Note that some of these factors came from the draft readiness matrix developed by the Subcommittee on Open Science [1], and some have been added based on further research. Definitions for some concepts are listed at the end of this document. This checklist is developed through a collaboration of ESIP Data Readiness Cluster members including representatives from NOAA, NASA, USGS, and other organizations. The checklist will be updated periodically to reflect community feedback.

Before the assessment: Ideally for AI-ready assessment, a dataset should be defined as the minimum measurable bundle (i.e., a physical parameter/variable of observational datasets or model simulations). The assessment at this scale will enable better integration of data from different sources for research and development. However, it can be an intensive process for manual assessment without automation. Therefore, we recommend current assessments be done on the data file level. If the dataset has different versions, the checklist should be applied to each dataset type (e.g. raw, derived).

Version: 1.0; *Last updated*: 2023-06-16.

General Information:

- Link to the dataset landing page (questions below could be automatically filled from the landing page in the future)
    - Name of the dataset
    - Current version of the dataset
    - Point of contact for the dataset
    - When was the dataset originally published?
- Is this raw data or a derived/processed data product? *Raw / Derived*
- **Is this observational data, simulation/model output, or synthetic data?** *Observed / Modeled / Synthetic.*

  SEI NOTE: The dataset should include observational data and any features necessary for application of the AIR Tool. (The current version of the AIR Tool does not currently support processing image or NLP data.)
- Is the data single-source or aggregated from several sources? *Single-source / Aggregated*

Data Quality

- Questions on timeliness
    - Will the dataset be updated? *Yes, it will be updated. / No, it will not be updated*
        - If the data will be updated, how often will it be updated? *[choose from]*
            - *Updated when new data are added*
                - Choose the update frequency that best describes the dataset:
                    - *near-real-time (irregularly) / hourly or sub-hourly /*

daily or sub-daily / weekly / monthly / yearly / longer than a year

- Will there be different stages of the update (e.g., updated with preliminary data first and replaced by a later update of the full record)? *Yes/No/Not applicable.*
  - If yes, what is the delay between different stages? [short answer]
- *Updated when a new version of the dataset is available.*
  - Should the new version of the dataset supersede the current version? *Yes / No / Others*
    - Provide an explanation for "*Others*"

- Questions on data completeness
  - **Is there any documentation about the completeness of the dataset?** *[✓ Yes] / No*
    - If *yes*, link to report/document
  - How complete is the dataset compared to the expected spatial coverage? *Complete / Partial / Unknown / Not applicable*
  - How complete is the dataset compared to the expected temporal coverage? *Complete / Partial / Unknown / Not applicable*
    - SEI Note: Is this dataset subject to confounding? *The current version of the AIR Tool does not currently support significant confounding (i.e., any common causes of two or more variables in the dataset are themselves captured in the dataset).*
- Questions on data consistency
  - **Is this dataset self-consistent in that its units, data types, and parameter names do not change over time and space?** *[✓ Yes] / No / Not applicable*
  - Is this dataset's units, data types, and parameter names consistent with similar data collections? *Yes / No / Not applicable*
  - Are there processes to monitor for units, data types, and parameter consistency? *Yes / No / Not applicable*
    - If *yes*, what measures are taken? *Manual review / Automated review*
- Questions on data bias
  - Is there known bias in the dataset? *Yes / No*
    - SEI Note: If *yes*, provide more information. Ideally, no significant measurement error in the data
  - Have measures been taken to examine bias? *Yes / No*
    - If *yes*, what measures were used?
    - Is the bias metrological traceable?
  - Is there reported bias in the data? *No known bias / Bias found and reported / No information available*
    - (optional) Link to the report/document on the bias
    - (optional) Link to tools available to reduce bias
    - (optional) Link to a bias-corrected or bias-reduced version of the dataset
- Is there quantitative information about data resolution in space and time? *Yes / No / Not applicable*
- Are there published data quality procedures or reports? Yes / No
  - *If there is published quality information, please provide the link to the information.*
- Is the provenance of the dataset tracked and documented? *Yes / No / Not applicable*
- Are there checksums / other checks for data integrity? *Yes / No / Not applicable*

- **What is the size of the dataset?** Depending on the resource, this might be total data volume, dimensionality, number of images, data files, table rows, image size, etc. *Short Answer*
    - *SEI Note:* We'd strongly prefer if the dataset had fewer than about 1000 variables. The current version of the AIR Tool does not support images.

Data Documentation

- Does the dataset metadata follow a community/domain standard or convention? *Yes / No / Not applicable*
    - If the metadata follows a community/domain standard, which standard is it? Select from a list [CF, …, TBD, others]
    - Is the dataset metadata machine-readable? *Yes / No / Not applicable*
    - Does it include details on the spatial and temporal extent? *Yes / No / Not applicable*
- **Is there a comprehensive data dictionary/codebook that describes what each element of the dataset means? parameters?** *Yes / No / Not applicable*
    - Is the data dictionary standardized? *Yes / No / Not applicable*
    - Is the data dictionary machine-readable? *Yes / No / Not applicable*
    - Do the parameters follow a defined standard? *Yes / No / Not applicable*
        - *If the parameters follow a defined standard, which standard it is?*
    - Are parameters crosswalked in an ontology or common vocabulary (e.g. NIEM)? *Yes / No / Not applicable*
- Does the dataset have a unique persistent identifier, e.g. DOI? *Yes, [supply identifier] / No / Not applicable*
- **Is there contact information for subject-matter experts?** *Yes / No / Not applicable*
- Is there a mechanism for user feedback and suggestions? *Yes / No / Not applicable*
- Are there example codes/notebooks/toolkits available showing how the data can be used? *Yes / No / Not applicable*
- What is the license for the data?
    - *Pick from a list of data licenses + others + no official data license*
    - Is the license standardized and machine-readable (e.g. Creative Commons)? *Yes / No / Not applicable*
- **Has this dataset already been used in AI or ML activities?** *Yes / No Link to publications/reports*.
    - SEI Note: We'd strongly prefer if the answer is "Yes" because it adds to the business value for employing the AIR Tool—not for application of the AIR Tool to the data.
- **Are there recommendations on the intended use of the data, and uses that are not recommended?** *Yes / No/ Not applicable*
    - SEI Note: There needs to be one or more controllable (intervenable) variables in the dataset (to represent scenarios we might take action on).

Data Access

- **What is/are the major file formats?** *Pick from a list of common data formats/ "other" (choose all that apply)*
    - **Is this format machine-readable?** *Yes / No / Not applicable*
        - *SEI Note: CSV is preferred.*
    - Is the data available in at least one open, non-proprietary format? *Yes / No / Not*

applicable

  - *Are there tools/services to support data format conversion? Yes / No*
    - *If so, provide the link to the tools/services*
- Data delivery:
  - Does data access require authentication (e.g., a registered user account)? *Yes / No / Not applicable*
  - Can the file be accessed via direct file downloading or ordering? *Yes / No / Not applicable*
  - Is there an Application Programming Interface (API) or web service to access the data? *Yes / No / Not applicable*
    - If there is an API, does the API follow an open standard protocol (e.g., OGC)? *Yes / No*
    - If there is an API, is there documentation for the API? *Yes / No*
      - If answered "Yes", please provide a URL to the documentation.
  - Is the data available publicly via cloud services? *Yes / No / Not applicable*
- For restricted data, have measures been taken to provide some access while still applying appropriate protection for privacy and security? *Yes / No / Not Applicable*
  - Has the data been aggregated to reduce granularity? *Yes / No / Not applicable*
  - Has the data been anonymized / de-identified? *Yes / No / Not applicable*
  - Is there secure access to the full dataset for authorized users? *Yes / No / Not applicable*

Data Preparation

- **Have null values/gaps been filled?** *[✓ Yes] / No / Not applicable*
- SEI Note: Has any functional determinism (e.g., no variable is the sum of two other variables) and high-intercorrelations (i.e., two variables whose correlation is > 0.9 or < -0.9) among the variables been addressed (causal discovery generally assumes "yes")? *Yes / No*
- **Have outliers been identified?** *[✓ Yes, tagged] / Yes, removed / No / Not applicable*
- **Is the data gridded (regularly sampled in time and space)?**
  - Options to *choose from (choose all that apply):*
    - *Regularly gridded in space / Constant time-frequency / Regularly gridded in space and constant time-frequency / Not gridded / Not applicable*
  - If the data is gridded, was it transformed from a different original sampling? *Yes, from irregular sampling / Yes, from a different regular sampling / No, this is the original sampling*
  - If the data is resampled from the original sampling, is the data also available at the original sampling? *Yes / No / Only available at request / Not applicable*
- **Are there associated targets or labels for supervised learning techniques (i.e., can this be used as a training dataset for supervised learning techniques)?** *[✓ Yes] / No / Not applicable*
  - *If there are associated targets/labels, are community labeling standards implemented (e.g., STAC label extension, ESA AIREO specification, etc)?*
    - *A list of label standards to choose from + "others" option*

Definitions

Quality

- Completeness: the breadth of a dataset compared to an ideal 100% completion (spatial, temporal, demographic, etc.); important in avoiding sampling bias
- Consistency: uniformity within the entire dataset or compared with similar data collections; for example, no changes in units or data types over time; item measured against itself or its counterpart in another dataset or database
- Bias: a systematic tilt in the dataset when compared to a reference, caused for example by instrumentation, incorrect data processing, unrepresentative sampling, or human error; the exact nature of bias and how it is measured will vary depending on the type of data and the research domain.
- Uncertainty: parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.
- Timeliness: the speed of data release, compared to when an event occurred or measurements were made; requirements will vary depending on the timeframe of the phenomenon (e.g., severe thunderstorms vs. climate change, or disease outbreaks vs. life expectancy trends)
- Provenance: identification of the data sources, how it was processed, and who released it
- Integrity: verification that the data remains unchanged from the original; aka data fixity.

Documentation

- Dataset Metadata: complete information about the dataset: quality, provenance, location, time period, responsible parties, purpose, etc.
- Data Dictionary / Codebook: complete information about the individual variables/ measures/ parameters within a dataset: type, units, null value, etc.
- Identifier: a code or number that uniquely identifies a dataset
- Ontology: formalized definitions of concepts within a domain of knowledge, and the nature of the inter-relationships among those concepts

Access

- Formats: standards that govern how information is stored in a computer file (e.g., CSV, JSON, GeoTIFF, etc.); different AI user communities will have different requirements, so the best practice is to provide several format options to meet the needs of multiple high priority user communities.
- Delivery Options: mechanisms for publishing open data for public use (e.g., direct file download, Application Programming Interface (API), cloud services, etc.); different AI user communities will have different requirements, so the best practice is to provide several delivery options to meet the needs of multiple high priority user communities.
- License / Usage Rights: information on who is allowed to use the data and for what purposes, including data sharing agreements, fees, etc.; some federal data needs to have restrictions and some will be fully open, so rights should be documented in detail

- Security / Privacy: protection of data that is restricted in some way (privacy, proprietary/business information, national security, etc.)

References:

1. OSTP Subcommittee on Open Science (2019), Draft AI-ready data matrix. [*This draft document is not an official publication of the committee.*]