Note that some of these factors came from the draft readiness matrix developed by the Subcommittee on Open Science, 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.

Version: 0.2; *Last updated*: April 1, 2022.

Data Preparation

- Have null values/gaps been filled? *Yes / No / Not applicable*
- Have outliers been identified? *Yes, tagged / Yes, removed / No / Not applicable*
- Is the data single-source or aggregated from several sources? *Single / Aggregated*
- Has the data been gridded (regularized in space and time) or is it in the originally sampled resolution? *Gridded / Not gridded / Not applicable*
- Have targets been identified and labeled (i.e. can this be used as a training dataset for supervised learning techniques)? *Yes / No / Not applicable*

Data Quality

- Have measures been taken to ensure completeness? *Yes / No / Not applicable*
- Are there automated processes to monitor consistency? *Yes / No / Not applicable*
- Have measures been taken to reduce bias? *Yes / No / Not applicable*
- What is the timeliness of the data? *Near real-time, 1 week, 1 month, 1 year, more than 1 year*
    - Is there a difference between raw near real-time access vs fully quality-controlled data that has an additional delay? *Yes / No / Not applicable*
- Are there quantitative measures of uncertainty? *Yes / No / Not applicable*
- Is there quantitative information about data resolution in space and time? *Yes / No / Not applicable*
- Are there published data quality procedures or reports? *Link to reports.*
- Is the provenance tracked and documented? *Yes / No / Not applicable*
- Are there checksums / other checks for data integrity? *Yes / No / Not applicable*
- How big is 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*
- Is this essentially raw data or a derived/processed data product? *Raw / Derived*
- Is this observational data or simulation/model output? *Observed / modeled*
- Has the data been peer-reviewed? *Yes / No / Not applicable*
- Has it been down-sampled to reduce resolution, or is it raw? If so, are the raw data available?

Data documentation

- Does the dataset have metadata? *Yes / No / Not applicable*
    - Is the dataset metadata standardized? *Yes / No / Not applicable*
    - 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 to describe 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*
    - 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 / 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*
- Is there a clear data license? *Yes / No/ Not applicable*
    - 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? *Link to publications / reports*.
- Are there recommendations on the intended use of the data, and uses that are not recommended? *Yes / No/ Not applicable*

Data access

- What is the file format? *Pick from list / "other"*
    - Is it machine-readable? *Yes / No / Not applicable*
    - Is it available in at least one open, non-proprietary format? *Yes / No / Not applicable*
    - Is it available in several different file formats? *Yes / No / Not applicable*
- Data delivery:
    - Direct file download or ordering? *Yes / No/ Not applicable*
    - Is there an API? *Yes / No/ Not applicable*
        - Custom-developed or open, standard protocol? *Custom / Standard*
    - On the cloud? *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*

Definitions

(outlined in the draft OSTP Open Data Sub-committee 2019 data AI maturity matrix)

Quality

- Completeness: the breadth of a dataset compared to an ideal 100% completion (spatial, temporal, demographic, etc.); important in avoiding 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, 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
- 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.)