What is AI-Ready Open Data?

Tyler Christensen (NOAA / NOS), Cassandra Ladino (USGS),
Dee Clarkin (NOAA Central Library), Bob Williams (HHS)

NOAA Workshop on Leveraging AI in Environmental Sciences
October 2020

Background and context

2019 exec order on AI Research & Development

"... agencies shall improve data and model inventory documentation to enable discovery and usability, and shall prioritize improvements to access and quality of AI data and models based on the AI research community's user feedback."

OSTP Subcommittee on Open Science (SOS) asked...

What specific improvements would make data more useful?

What is AI-ready data?

AI research community's user feedback

- Two federal register RFIs
  - Updating the National AI R&D Strategic Plan
  - Priority AI data & model improvements
- OMB “AI Inventory Guidance”
- DoE “Data for AI” roundtable
- OSTP SOS agency survey
  - 16 federal science agencies
  - AI researchers and data stewards

What factors are most important?

Data quality

Access

Documentation

What factors are most important?

Data quality

- Completeness (spatial / temporal / demographic)
- Consistency (uniformity within the dataset)
- Lack of bias (no systematic “tilt”)
- Timeliness (speed of data release)
- Provenance & Integrity (unchanged from a trusted source)

Access

Documentation

What factors are most important?

Data quality

Access

- Formats (variety of formats is preferred)
- Delivery Options (again, variety is preferred)
- Usage Rights (clear, machine-readable license)
- Security / Privacy (protecting restricted data)

Documentation

What factors are most important?

Data quality

Access

- Formats (variety of formats is preferred)
- Delivery Options (again, variety is preferred)
- Usage Rights (clear, machine-readable license)
- Security / Privacy (protecting restricted data)

Documentation

PLEASE NOTE

“AI-Ready”
≠
“Dump any data onto the Cloud”

What factors are most important?

Data quality

Access

Documentation

- Dataset Metadata (info about the data)
- Data Dictionary (info about each parameter)
- Identifier (number / code that uniquely identifies the dataset)

Creating a first-draft readiness matrix

Level 0 = not AI-ready

- The dataset meets basic requirements for Open Data, but does not specifically facilitate AI/ML.

Level 3 = Optimal

- Data pipeline ensures versioning, provenance, data integrity, and protection of sensitive information. Data has robust machine-readable metadata, license, and data dictionary.

Readiness example: data quality

Consistency (*uniformity within the dataset*)

Level 0 / Not AI-Ready	no formal effort to ensure internal consistency before data are published
Level 1 / Minimal	manual checks for consistency
Level 2 / Intermediate	some consistency checks are automated, some documentation of results
Level 3 / Optimal	fully-automated internal consistency checks and reporting; some consideration for external consistency among community datasets

Readiness example: data access

Delivery Options

Level 0 / Not AI-Ready	open for public use only by request or via an ordering system
Level 1 / Minimal	one non-programmatic access option only, such as file download
Level 2 / Intermediate	multiple delivery options including at least one programmatically accessible method, such as bulk file download plus API or cloud
Level 3 / Optimal	multiple delivery options (download, API, cloud, HPC, data-as-a-service, etc.)

Readiness example: documentation

Data Dictionary (*info about each parameter*)

Level 0 / Not AI-Ready	no data dictionary available, or in non machine-readable format (e.g. pdf)
Level 1 / Minimal	data dictionary in machine-readable format (e.g. csv, xml, json)
Level 2 / Intermediate	data dictionary uses a machine-readable metadata standard
Level 3 / Optimal	machine-readable metadata standard; parameters are harmonized with other agency datasets, across Federal agencies, or domain standards

Next steps

~~OSTP Subcommittee on Open Science~~

OSTP Committee on AI - data working group

OSTP request for pilot projects

NOAA AI Center: Test readiness matrix for pilot NOAA datasets

More input from external AI researchers?

Other ideas on how to proceed?

What would make NOAA's open data easier to use in your own AI projects?

Thank you!

Full document here (outside NOAA? Just ask!)

Questions or suggestions: tyler.christensen@noaa.gov