July 16, 2024

**Commerce Department RFI:** 89 FR 27411
**Organization:** The Center for Security and Emerging Technology (CSET)
**Respondent type:** Academic institution / Think tank
**Primary respondent:** Catherine Aiken, Director of Data Science and Research
**Additional respondents:** James Dunham, Jacob Feldgoise, Rebecca Gelles, Ronnie Kinoshita, Mina Narayanan, and Christian Schoeberl.

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following response to the Request for Information on AI and Open Government Data Assets. A policy research organization within Georgetown University, CSET provides decision-makers with data-driven analysis on the security implications of emerging technologies, focusing on artificial intelligence, cybersecurity, and biotechnology. We appreciate the opportunity to offer these comments.

CSET supports Commerce's efforts to advance public data accessibility, quality, and transparency. We encourage Commerce to consider existing standards, tools, and best practices for making data usable by humans, as they go hand-in-hand with making data AI-ready. To that end, we encourage Commerce to:

- Leverage existing platforms, forums, and dissemination practices (e.g. GitHub, Zenodo)
- Prioritize clear, understandable, comprehensive data documentation (e.g. data cards)
- Align data assets with existing tools and data sets, including incorporating open organization identifiers and existing occupational codes (e.g. ROR, SOC)

These priorities will help make data usable, ensure accuracy, foster responsible use, and mitigate bias. These priorities also enable consistency and data linkage, two critical data features for human use and AI applications.

CSET has published two Data Snapshots that offer relevant suggestions on the topic of using open Commerce data for analysis. Please see BIS Best Data Practices: Part 1 and BIS Best Data Practices: Part 2.

Our response is structured according to the topical questions outlined in the Request.

## Data Dissemination Standards

Machine-readable data, metadata, and documentation are critical features to facilitate AI applications with open data. The AI system must be able to understand the data set and metadata, and human users must be able to understand the system's features. For a maximum utility, a low cost but high payoff priority should be explicit, understandable descriptions of what each data point in the data set means (i.e., what each point is counting vs not counting, how counts are calculated), all of which is information not immediately obvious to users. It would be very helpful to clearly articulate the limitations of each dataset provided to the public. What is *not* captured by the scope of the data? What types of conclusions can *not* be derived from the data? Presumably, Commerce is familiar with how the public or researchers have misused/misunderstood the dataset in the past, and documenting those misunderstandings would help.

Public datasets and documentation need to be downloadable in bulk, machine-readable, and available as a csv or json file format. Ideally, they would include standard, open-source or non-proprietary entity identifiers and occupation codes to allow users to connect datasets, and maintain consistency over time to allow users to analyze trends over time.

As raw and derived data are described in the Request, one distinction is that derived data likely has had privacy protections applied to it. If that is not the case, it should be a requirement. Raw data may not have gone through the same anonymizing or privacy preserving processes, since it may be assumed that raw data cannot be directly linked to users or used to extract information about individuals. But considering what kind of personally identifiable information (PII) could be leaked through raw data is important and should be prioritized. We emphasize that many of the recommendations we provide apply to raw and derived data, and that metadata standards should apply consistently.

Commerce should release data under open licenses to support broad, equitable, and open access to Commerce data sets and metadata to signal to users that they provide public data. If the intent is to more clearly signal that data is available for use by AI systems, Commerce could consider releasing data under licenses with specific allowances for AI use, possibly specifying acceptable, or unacceptable, uses (see Responsible AI Licenses). To further signal data is available for AI development, Commerce could make data available not just on .gov websites, but also via forums widely used by the AI development community. For example, increasing activity and availability through Commerce's GitHub or by sharing data assets on Zenodo.

Encouraging use of data assets in projects and outreach under the National Artificial Intelligence Research Resource (NAIRR) Pilot would likely help as well.

While open is generally good, it is important to consider the potential for PII leakage, and balance that against the benefit of useful open data. Commerce should ensure that individuals cannot be identified from the data. Other potential harms and biases should also be considered, and if the primary or only use case for a data set is likely to be one that is harmful rather than beneficial, that should shape decisions about releasing the data openly, or with specific access and use restrictions.

## Data Accessibility and Retrieval

Data can be more accessible and valuable to the community if it were provided as downloadable data sets (json or csv) or via an API to pull bulk data. A priority should be providing the data in a format that is not PDFs. Existing Commerce data sets our team has worked with were available as PDFs of collected information, with inconsistencies both by agency and by year. Specific priorities could be to provide downloadable CSV data sets for aggregated BIS licensing data and the Commerce Control List. In addition, Commerce could make its existing CSV data set for the BIS Entity List more usable by addressing the following data quality issues:

1. Ensure that the effective date and country fields are filled out for each listed entity, and
2. Standardize how Entity List modifications are reflected in the data set (i.e. adding a new row or modifying an existing row).

A site that centralizes the data to make scraping easier would improve web crawlability, though reliance on web scraping does reduce accessibility to many researchers and teams. A centralized location would also reduce effort required to track data across time and agencies, and result in more consistent and higher-quality data. Existing data assets would also be more valuable if they included standard, open-source/non-proprietary entity identifiers, occupation codes, and similar metadata.

While there are several things Commerce should consider to develop intuitive, user-friendly, and accessible data portals and interfaces, we also encourage leveraging existing open source options. These come with the benefit of widely used and well-documented methods and tools for navigation and retrieval. Commerce could have a presence on existing public data portals (e.g., Zenodo) and rely on existing knowledge and documentation.

Assuming Commerce will additionally, or alternatively, maintain its own data portal, the important things are to centralize data in one place and make download and API access straightforward. But Commerce should invest in understanding the needs of data users, and the impact of investing in data availability. There is existing literature on user testing methods and best practices that Commerce could work from, as well as research on early impact of other open government data initiatives to learn from (like here).

## Partnership Engagement

Areas of partnership or collaboration to enhance data quality, integrity, and usefulness could focus on making historical data more usable, improving data quality and usefulness in the process. For example, Commerce could sponsor a challenge on extracting data from old PDFs, or updating and translating existing documentation into new, machine-readable formats. Other forms of prizes or hackathons would also increase visibility, use, and data quality. A partner organization could build on work done in the challenge to produce publishable research, increasing awareness of the usefulness of Commerce data.

## Data Integrity and Quality

There is an expectation for reporting data quality in clear, understandable documentation. That is a critical step in providing data that will be used in AI applications, as well as for human analysis. The same applies to transparent data sourcing, processing, and updating; it boils down to documentation. Public data documentation should include a detailed description of data sourcing and processing methods, and if these differ across the data set, each different method or source should be enumerated. Any known quality issues, and any evaluation done to assess quality, should also be documented as part of the data description.

We know of no universal solution to ensure documented quality and processing information is carried through to an AI end user. The ability to ensure that depends on the AI application. For some AI applications, it is more feasible to encourage information from data documentation to be included in system outputs. But if, for example, the data is used for training a large language model, and is one of many original data sources, it is less likely documented limitations or sourcing will make its way into a final product. To encourage such behavior though, Commerce could adopt a known, standard data documentation format, like releasing a data card with each data set (see here and here) or require it in data use specifications or licenses.

Adopting a familiar format for data documentation will also facilitate transparency. Another recommendation, specific to replication and analysis, would be to host assets and documentation in GitHub repositories (see this USASpending example). Storing data documentation (and data) within a GitHub repository makes it easy to post and follow issues, track changes in documentation**,** and find example cases that can be tailored for a researcher's specific needs. This would have the added benefit of facilitating communication with researchers and stakeholders interested in using the data. If there were questions about the data documentation, or issues found with the data itself, these could be communicated easily by users to Commerce using Github issues, and Commerce can clarify and provide updates in turn.

## Data Ethics

AI systems are only as good as the datasets they are trained on; one effective method to promote equitable outcomes is to ensure Commerce's released data sets are as representative and complete as possible. A biased data set might have an overrepresentation from majority or advantaged groups and very limited representation from minority or disadvantaged groups. These data sets are much more likely to result in biased algorithms. In addition to avoiding the release of highly imbalanced data sets, Commerce should clearly document moderate imbalances between groups within a data set and point users towards existing tools and literature for mitigating bias.

We recommend tracking the sources of and modifications to data to identify and protect stakeholders' data. GitHub repositories can offer a useful platform for Commerce to record the provenance of the data sets: data sources, how data has been changed, who is responsible for maintaining the data, obligations of data providers and users. For data sets that might contain PII, access restrictions licenses can protect the privacy and property rights of data subjects.