

July 16, 2024



Victoria Houed, OUSEA
U.S. Department of Commerce
1401 Constitution Avenue NW
Room 4848
Washington, DC 20230

1345 Avenue of the Americas
27th Floor
New York, NY 10105
Phone: (202) 448-1985
Fax: (866) 516-6923

Dear Ms. Houed:

RE: AI-Ready Open Data Assets RFI, DOC– 2024–0007

We appreciate the opportunity to provide input to the Department of Commerce (DOC) Request for Information on AI-Ready Open Data Assets. We agree with the DOC view that AI systems should be powered by data that is not just machine-readable and accessible, but that is “machine-understandable.” Artificial intelligence has enormous potential, but AI algorithms must leverage high-quality, unambiguously understood data to generate reliable, useful results.

The path to providing data of high integrity and accuracy is through open, nonproprietary (free) data standards, an approach that has been successfully employed around the world. Data standards programs have been adopted by 80 global regulators in close to 220 programs¹ for data collected from public and private companies, banks, regulators, and utilities. Data standards programs, established by U.S. regulators, have been in place for data collections for 15 years.

Furthermore, data standards are required to be used in support of the Financial Data Transparency Act (FDTA), legislation passed in December 2022, which impacts data collected by eight regulatory agencies² that are members of the Financial Stability Oversight Council. Data collection called for in the FDTA is expected to begin by 2027 which should result in the creation of large pools of structured, standardized data.

XBRL US is a nonprofit data standards organization, with a mission to improve the efficiency and quality of reporting in the U.S. by promoting the adoption of business reporting standards. XBRL US is a jurisdiction of XBRL International, the nonprofit consortium responsible for developing and maintaining the technical specification for XBRL (a free and open data standard widely used around the world for reporting by public and private companies, banks, and government agencies). Our members include accounting firms, public companies, software, data, and service providers, as well as other nonprofits and standards organizations. We support the use of open,

¹ XBRL International Project Directory: <https://www.xbrl.org/the-standard/why/xbrl-project-directory/>

² Federal agencies included in the FDTA: Treasury, SEC, FDIC, Federal Reserve, Consumer Financial Protection Board (CFPB), National Credit Union Administration (NCUA), Federal Housing Finance Agency (FHFA), Office of the Comptroller of the Currency (OCC).

nonproprietary data standards as they have been proven to reduce cost, improve timeliness, and efficiency for all stakeholders. As noted in the SEC 2024 Annual Report to Congress³, “*Studies show that machine-readable disclosures benefit investors, markets, and issuers. With respect to investors and, more broadly, to markets, making corporate disclosures machine readable has decreased information asymmetry between firms and investors by reducing information processing costs, making stock prices more informative (i.e., more reflective of firm-specific information), and reducing market inefficiencies and risks.*”

Economic, population, and environmental data collected and prepared by the bureaus and offices of the DOC is used by regulators, governments, academic researchers, and businesses across the nation. The ability to access that data in the same structured, standardized format that is used for data reported to regulators by banks, public companies, investment management companies, and ultimately by many more reporting entities, including state and local governments through the FDTA, will enable interoperability of multiple datasets, making it faster, less expensive and more efficient to conduct robust analysis.

We encourage the DOC to review the vast amount of structured, standardized data collected by US regulators, including the Securities and Exchange Commission (SEC), the Federal Deposit Insurance Corporation (FDIC), and the Federal Energy Regulatory Commission (FERC); and to consider the future datasets that will be made available through the roll-out of the FDTA. The ability for data users to access this data in the same format as DOC provided data, will enable significant economies of scale and cost savings for regulators, researchers, and businesses alike.

This letter addresses many of the questions raised in the DOC Request for Information.

Data Dissemination Standards

1. What data dissemination standards should Commerce adopt to support human-readable and machine-understandable public data?

We support the use of the eXtensible Business Reporting Language (XBRL) standard as the most appropriate standard to be used to support the collection and dissemination of data. XBRL can manage multiple data types and unit types, for example, monetary, volume, energy, length, time, mass, string. It has a defined method to manage and concretely report units that are necessary for associated data types such as currencies for monetary types and the varied units of measure for length as examples. Other datasets collected by U.S. regulators including the SEC, the FDIC and the FERC are prepared in XBRL. DOC data prepared in the same structured, standardized format would thus be interoperable with these other regulatory data collections, reducing cost across all data users.

2. What formats, metadata, and documentation should be prioritized to facilitate AI applications?

³ SEC 2024 Semi-Annual Report to Congress: <https://www.sec.gov/files/fdta-report-6-2024.pdf>

Regulatory datasets published by US agencies today vary widely in how they are provided. Many are provided as reports in paper-based documents such as PDF or Word, some are posted in CSV files for download, some in structured XBRL format. An AI application may need to consume multiple datasets provided by multiple agencies. Ideally, all U.S. federal agencies would provide their data following the same structured data standard. This approach would provide AI algorithms with data that is interoperable, “machine-understandable,” accessible, and automatable.

Data that is prepared using the XBRL data standard today such as data from public companies, public utilities, investment management companies, credit rating agencies, and banks (each reporting to their respective regulatory agency), meet these requirements, and can be easily housed in the same database and extracted and used in the same manner. At XBRL US, we maintain a database that contains data from FERC filings, SEC filings, European Single Electronic Formatting (ESEF) filings, and even financial statement data from state and local government entities. All of this data can be extracted in highly granular form using the same extraction or analytical applications because the data, while quite different, is structured in the same way using a standardized model or ontology. Standardization enables economies of scale and makes it less expensive to report, collect and extract data because there is a plethora of tools available on the commercial market.

The XBRL standard is not a “format,” rather it is a semantic data model that can be used to generate data in multiple formats, including XHTML, JSON, CSV, and XML. Different data collection programs may be better suited to one format versus another; therefore, the DOC should be prepared to adopt formats that are the right “fit” for the data collected. The XBRL standard is a semantic data model, rather than a format like XML. It has the flexibility to facilitate an approach that allows for more than one format.

Documentation is important to ensure that all stakeholders have a shared understanding of the data. The XBRL standard requires the creation of a taxonomy (or ontology) which contains all the documentation needed for all stakeholders including the concepts that can be reported along with their properties, labels, and references; and the relationships between those concepts such as mathematical and parent/child relationships. A taxonomy is a digital dictionary of terms that contains all the information needed for anyone involved in reporting, collecting, or using the data that is expressed by the taxonomy. When all stakeholders can refer to a single source such as a taxonomy (ontology), they have a shared understanding of what is reported, collected, and ultimately used.

3. How does raw data, such as data from the sensor networks, differ from derived data, such as statistical data from the U.S. Census Bureau, when it comes to metadata standards?

XBRL classifies different data based on dimensions. This data can fit into distinct categories. Most of the data collected by Federal agencies is statistical or aggregated data that represents positions or balances at a point in time or flows or activity over a period of time such as GDP for a quarter. XBRL refers to this as fact-based data. Additional data sets are classified as follows:

- Time Series Data (Data that is measured on a regular basis)

- Event Based Data (Data that is measured when an event occurs such as transactional data)
- Reference Based Data (Securities Master File, Listing of Leases)
- Positional Data (Inventory listing, employee listing)
- Data sets recording the occurrence of an event (Journal entries)

Time series data is collected on a regular time interval either at a point in time such as ocean temperatures or data that is collected over a short duration of time such as peak load on an electric grid for a given hour. All of this data is captured using a period dimension in XBRL. This period dimension is consistent across all XBRL filings globally which means that this data is comparable across data sets.

Event based data occurs on an infrequent basis but is dimensionalized using an event-based dimension. It also must have an associated date representing the point at which the event occurred, that can be related to the period dimension. This typically captures transactional data such as sales and purchases. It can also be used for non-financial events such as planes landing at an airport.

Reference based data is time independent. This data does not change with the passage of time. This includes data sets like a listing of securities ever issued, or a listing of contracts that an entity has signed. With the passage of time the details of this data do not change. This data is often referenced by other data sets such as transactional data referencing a security that was sold.

Positional data is a breakdown of data by items that existed at a point in time. This may include all the securities an entity holds, or all the details of all transformers held by an electric utility at a point in time.

Journal entry data records how an event is actually recorded by an entity and how the data is classified in a system. This only applies to financial data.

All of the above data sets apply to the actual data captured. How these are defined is described in the XBRL taxonomy. Each of these data sets will have associated fields and dimensions. The taxonomy defines how these fields are related and what is the categorization of the data.

4. What data licensing practices, standards, and usage considerations should Commerce consider to support broad, equitable, and open access to its datasets and metadata?

All data collected by the DOC should be formatted using an open, nonproprietary standard to ensure that it can be freely used by consumers.

Data Accessibility and Retrievability

1. How can Commerce's data assets be made more accessible and valuable to the AI community (e.g., improved API access, web crawlability, etc.)?

When data is collected and made available in structured, standardized format, the reported files can be posted and accessible through notification feeds such as RSS to automatically update user databases and applications. Because standardized data in XBRL format is produced and used worldwide, numerous open-source and commercial tools are available that can immediately begin consuming this data. Data that is provided by regulators in custom XML format, or in static files like PDF, or even CSV or Excel, must be manipulated to fit the database or tool before the data can be used by consumers. That data preparation cost is duplicated across every user, resulting in unnecessary costs that can be eliminated or at least significantly reduced when data is in structured, standard format.

2. How can Commerce develop intuitive and accessible data portals that facilitate easy navigation and retrieval of data sets?

It is critically important to have a single data model that can be accessed and queried, that gives the user a complete understanding of what the different datasets represent and how they are related to each other. Providing discrete datasets with common data identifiers but no linkage between them forces users of the data to employ specialized background knowledge that is not available in the model. This makes it less useful for users and AI models.

Providing economic, population and environmental data that has been prepared in a standardized, structured, widely used format like XBRL will allow the commercial markets to extract data quickly and easily, using tools that are already able to access XBRL data. DOC data can then be easily commingled with other datasets maintained by users to produce robust, useful AI outcomes. The DOC could provide a simple portal for downloads of data and an RSS feed for market use; this is the approach followed by the SEC.

4. What measures can be taken to encourage user-friendly interfaces, including clear labeling and readable formats, for Commerce's online data resources?

Data prepared in structured, standardized format will ensure consistent labeling and unambiguous understanding of reported data.

Partnership Engagement

1. How can industry and academic stakeholders collaborate with the government to shape the design and dissemination of AI-ready open data?

XBRL taxonomies built to represent DOC data should be published for public exposure to solicit input from the market that uses the data.

2. What are the potential areas of partnership, and how can industry and academia contribute to enhancing data quality, integrity, and usefulness for AI purposes?

Industry and academia are eager to have access to highly granular, easily accessible, inexpensive data to support research studies, develop business strategies, and identify risk. Providing consistently prepared open-source structured data will be highly beneficial for all users. Their input during taxonomy public reviews and reviews of data quality can provide an important feedback loop to the DOC to make further enhancements to the taxonomy and to validation rules.

Data Integrity and Quality

1. What are best practices that industries have employed to enhance the integrity and accuracy of public data when used in AI applications? What are best practices for data verification and validation? What are best practices for conducting regular audits and quality checks of data used in AI applications?

Data integrity is achieved when the creators and users of the data have a clear, shared understanding of the data. Creating data in structured, standardized format imposes the discipline needed for data that is accurate and transparent. A single data point like the value 12,277 highlighted on the income statement below must have the properties of the fact shown in the green boxes on the right, embedded when it is transported from one machine to another. Even less complex data, such as labels or text blocks, must have comparable properties connected to the fact in order for it to be unambiguously “read” and understood by a machine.

	For the Fiscal Years Ended May		
	2024	2023	2022
Net sales	\$ 12,050.9	\$ 12,277.0	\$ 11,333.9
Costs and expenses:			
Cost of goods sold	8,717.5	9,012.2	8,497.1
Selling, general and administrative expenses	2,480.6	2,189.5	1,492.8
Pension and postretirement non-service income	10.3	24.2	67.3
Interest expense, net	430.5	409.6	379.9
Equity method investment earnings	177.6	212.0	145.3
Income before income taxes	\$ 610.2	\$ 901.9	\$ 1,178.7
Income tax expense	262.5	218.7	190.5
Net income	\$ 347.7	\$ 683.2	\$ 888.2
Less: Net income (loss) attributable to noncontrolling interests	0.5	(0.4)	-
Net income attributable to Conagra Brands, Inc.	\$ 347.2	\$ 683.6	\$ 888.2
Earnings per share — basic			
Net income attributable to Conagra Brands, Inc. common stockholders	\$ 0.73	\$ 1.43	\$ 1.85
Earnings per share — diluted			
Net income attributable to Conagra Brands, Inc. common stockholders	\$ 0.72	\$ 1.42	\$ 1.84

- Net Sales
- Monetary
- Definition
- Authoritative FASB References
- Millions
- 12 Mos Ending 5/28/2023
- Conagra Brands

Accuracy and the ability to verify data is achievable when the data model contains rules that explain the relationships between reported facts. In the example above, a fact reported for “Income before income taxes” has a positive calculation weight with “Net sales” and a negative calculation weight with the various components of “Costs and expenses.” By establishing these relationships in the data model (the taxonomy or ontology), preparers of the data can run clearly defined validation rules that alert them when a rule is not met; users of the data can run the same rules to identify when a fact is not in compliance with a required rule. Establishing rules that are used across all on the supply chain encourages quality and consistency in datasets.

4. How can Commerce promote transparency in data sourcing and processing methods to enhance trust and reliability? What is the expectation for reporting the quality of its data and how can we ensure that information will be carried through and presented to the end user?

Authoritative references and detailed metadata can be included in the taxonomy so that all supply chain users have the same “view” into what a reported fact represents. Metadata required to be connected to each concept must be clearly defined and unambiguous.

5. What validation processes can be established to maintain and verify data accuracy and consistency?

As noted in the response to question 1 in this section, sophisticated validation rules can be established that check for completeness, reasonableness, economic relationships, signage issues, and for many other situations that may require complex examination. Rule sets can be provided to the market and leveraged at the point of data creation, collection and extraction for a multi-pronged check. The granularity of XBRL lends itself to complex validation rules; and open-source processing languages are used today to write detailed rules for FERC and SEC filers so that they can check their filings before regulatory submissions. These rules have been highly effective at increasing the quality of reported data as shown by the Aggregated Real-time Filing Errors charts⁴ posted on the XBRL US Data Quality page.

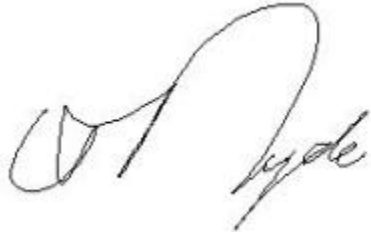
6. How can Commerce facilitate comprehensive and transparent data documentation for replication and analysis?

When data standards are used, supported by a taxonomy, the taxonomy provides comprehensive, detailed documentation and rules that provide guidance to data preparers and users. When a taxonomy is updated or revised, to reflect changes in concepts or guidance in what and how to report, that instruction is communicated simultaneously to all because the taxonomy is the primary source of information.

⁴ XBRL US Data Quality Committee, Aggregated Real-time Filing Errors: https://xbrl.us/data-quality/filing-results/dqc-results/**

Thank you again for the opportunity to provide input to this RFI. I am available to discuss this recommendation further or to answer any questions you may have. I can be reached at (917) 582-6159 or Campbell.Pryde@XBRL.US. I look forward to discussing this with you further.

Sincerely,

A handwritten signature in black ink, appearing to read "Campbell Pryde". The signature is fluid and cursive, with the first name "Campbell" written in a larger, more prominent script than the last name "Pryde".

Campbell Pryde
President and CEO, XBRL US