BrightQuery®
The Factual AI Company™

AI Readiness Case Study: National Center for Science and Engineering Statistics (NCSES)

AI-Ready Data Products to Facilitate Discovery and Use

ADC-AIRD-24-N22
Feb 7, 2025

Report Authors
Jose M. Plehn-Dujowich, Ph.D., CEO and Founder
jplehn@brightquery.com

Chris Shaffer, VP of Technology
chris@brightquery.com

Shyam Peri, Chief Data & AI Officer
shyam@brightquery.com

Peyton Marsh, VP of Communications
peyton@brightquery.com

Table of Contents

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. NCSES has reviewed this product for unauthorized disclosure of confidential information and approved its release (NCSES-DRN26-054).

Glossary

- **AI (Artificial Intelligence):** A field of computer science that enables machines to perform tasks that typically require human intelligence, such as understanding natural language, recognizing patterns, or making decisions.
- **AI Readiness:** The process of preparing datasets to be machine-readable and machine-understandable, enriched with metadata, and organized in a standardized format to optimize their usability by AI technologies.
- **API (Application Programming Interface):** A set of rules and tools that allow different software applications to communicate and share data. The NCSES API enables users to retrieve NCSES datasets programmatically.
- **ANSI:** American National Standards Institute; (among other things) maintains a list of codes to classify various reference data.
- **CIP:** Classification of Instructional Programs; maintained by the National Center for Education Statistics (NCES), categorize academic programs in the U.S.
- **Croissant Tagging:** An advanced metadata framework designed for AI and machine learning applications. It helps datasets become more discoverable and usable by enabling context-aware data interpretation.
- **Data Dictionary:** A centralized document or resource containing definitions of data items, descriptions of parameters, and metadata to help users understand the structure and meaning of a dataset.
- **Digitization:** The process of converting information from physical or analog formats into digital formats that can be accessed and analyzed using computers.
- **Granularity:** The level of detail in a dataset. High granularity means data is broken down into finer details, such as by geography (e.g., county level) or time (e.g., monthly).
- **Interactive Data Tools:** Web-based tools that allow users to explore datasets dynamically by applying filters or generating visualizations, such as tables, charts, or graphs.
- **Interoperability:** The ability of different systems, datasets, or tools to work together seamlessly, often by adhering to standardized formats or metadata structures.
- **ISO:** International Standards Organization; (among other things) maintains a list of codes to classify various reference data.
- **Metadata:** Data that provides information about other data, such as the title, description, units, and date of publication. Metadata helps users and AI systems understand the content and context of datasets.
- **NAICS (North American Industry Classification System):** A standard used by federal agencies to classify businesses and industries. It organizes sectors hierarchically for consistent reporting and analysis.
- **PII (Personally Identifiable Information):** Information that can identify an individual, such as a name, address, or Social Security number.
- **Schema.org Tagging:** A semantic vocabulary that provides metadata for web content. It enhances search engines' ability to understand the context of datasets and improve their discoverability.
- **SDMX:** A global initiative to improve Statistical Data and Metadata eXchange.
- **SOC:** Standard Occupational Classification; maintained by the Bureau of Labor Statistics (BLS) to classify and categorize occupations in the labor market.
- **Semantic Interoperability:** The ability of different datasets to be linked and understood in terms of their meaning and context, often facilitated by metadata standards like Schema.org.
- **Structured Data:** Data organized into a defined format, such as rows and columns in a table or fields in a database, making it easily machine-readable.
- **Unstructured Data:** Data that does not have a predefined structure, such as text documents or reports, requiring additional processing to extract information.
- **Version Control:** The process of tracking and managing changes to datasets over time, including initial releases, revisions, and final versions.
- **Visualization:** The graphical representation of data, such as charts, graphs, or maps, to make patterns and trends easier to interpret.
- **Web Scraping:** The automated process of extracting data from websites. It can be challenging for tools to process data not explicitly structured or tagged.
- **XML (Extensible Markup Language):** A markup language that defines rules for encoding data in a structured format, often used for sharing information between systems.

Executive Summary

In today's world, pursuit of the goal of creating FAIR/O – Findable, Accessible, Interoperable, Reusable, and Open-License – data for use by policymakers, researchers, journalists, industry, and the general public requires AI readiness. In the words of the Department of Commerce, which are broadly applicable across government, academia, and industry, this means "data that is not just machine-readable, but machine-understandable; data that is enriched with contextual metadata and organized in interpretable standard formats. AI models can then better interpret Commerce data, link them to similar data, and return accurate results from authoritative sources."

This report provides an in-depth landscape analysis of the National Center for Science and Engineering Statistics (NCSES) functional content—APIs, Unstructured Reports, Tables, Interactive Charts, and Website & Metadata, evaluating their readiness for integration into Artificial Intelligence (AI) and Machine Learning (ML) frameworks. As the role of AI continues to expand in transforming data into actionable insights, ensuring NCSES content meets the stringent requirements of AI systems is critical to unlocking their full potential.

The objective of this evaluation is twofold: first, to analyze the current state of NCSES's information published in terms of structure, metadata, interoperability, and other critical AI readiness criteria; and second, to identify actionable steps that can enhance their utility in AI applications. By doing so, the report will demonstrate how we intend to bridge the gap between traditional data practices and the advanced capabilities of AI technologies, **enabling search engines such as Google, and Large Language Models like ChatGPT to cite NCSES's statistical products more accurately.** Title II of the Evidence Act (or the "OPEN Government Data Act") "requires public government data assets to be published as machine-readable data and also machine understandable. In order to accomplish this, BrightQuery will help the NCSES place statistical data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost.

BrightQuery's efforts are supported by a team of esteemed consultants and strategic partnerships with leaders in AI and data technologies, and this has allowed BrightQuery to implement Schema.org tagging solutions that enrich metadata and improve discoverability for NCSES datasets, as well as Croissant tagging to optimize NCSES's data for machine learning applications. BrightQuery also partners with Google Data Commons to enhance data linkage capabilities. Through the AI Alliance, BrightQuery works with IBM and Meta as part of the Open Trusted Data Initiative to establish standards for secure and authoritative AI-ready data. These collaborations underscore BrightQuery's ability to deliver innovative and robust solutions for advancing NCSES's AI readiness.

Throughout this assessment, we outline the strengths of NCSES published information while identifying opportunities for improvement to align them with the standards necessary for advanced AI tools. This work also introduces key strategies and methodologies that can aid NCSES in transitioning its published content toward a more AI-ready framework, ensuring seamless integration and optimal performance in AI-driven environments.

Key Findings. See **Conclusion** for more; a few areas stand out above the others:

- Public Use Files: NCSES publishes individuals' responses to surveys (with personally identifiable information removed) that allows for more advanced data analysis
- Restricted Use Files: NCSES provides a clear procedure to request access to query that withheld information, and to do so in a way that minimizes risks to privacy
- Functional: No API Present
- Criteria: History
- Criteria: Time Series Consistency
- Theme: Discoverability and Navigation
- Criteria: AI / ML Specific Features

The Importance of AI Readiness. Advancements in artificial intelligence (AI) are transforming how data is accessed and utilized. For the National Center for Science and Engineering Statistics (NCSES), ensuring its datasets are AI-ready enhances their usability, accessibility, and relevance in today's data-driven world.
AI readiness involves two key elements:

- **Machine-Readability:** Structuring data in formats like APIs and tables that can be processed automatically.
- **Machine-Understandability:** Enriching data with metadata, such as definitions and units, to enable accurate interpretation by AI systems.

NCSES's datasets underpin critical economic analyses, from GDP to trade statistics. By improving metadata and adopting frameworks like Schema.org and Croissant tagging, NCSES can make its data more discoverable and interoperable with other systems, increasing its value for policymakers, researchers, and the public.

This report outlines a roadmap for NCSES to enhance AI readiness by addressing gaps in metadata, structure, and accessibility. By adopting these changes, NCSES can set a standard for transparent, AI-optimized federal data systems that empower economic decision-making.

Restatement of Objectives. The primary objective of this project is to explore how a future National Secure Data Service (NSDS) could provide shared information and tools for making statistical data products more readily ingestible by AI technologies. Generative AI (GenAI) is transforming industries and influencing the daily decisions of Americans. Typically, GenAI models are trained on a limited set of data, often excluding the most recent information. To make GenAI truly resourceful with the latest data, it is necessary to enhance these models with additional datasets using technologies such as Retrieval-Augmented Generation (RAG), Text2SQL, and model fine-tuning. These technologies significantly expand the knowledge base of GenAI models. However, these processes impose stringent requirements on the datasets used for augmentation, including data format, schema, context, provenance, and metadata. The same challenges apply to statistical datasets produced by government agencies, including NCSES. The BrightQuery (BQ) team proposes developing an open-source platform called **Government Data Agent (GDA)** with two components: **GDA Evaluator** to assess the AI readiness of a dataset; and **GDA Transformer** to perform the necessary transformations recommended by the evaluator. This multi-agent system powered by a Large Language Model (LLM) will intelligently assess if a dataset is AI-ready and then transform problematic datasets to meet AI standards. GDA will primarily serve federal agencies by evaluating, transforming, and upgrading their datasets to meet the demands of the rapidly growing GenAI applications. This ensures that the information retrieved by GenAI applications is accurate, trustworthy, and traceable to its original source.

Background on AI Readiness. To be AI-ready, datasets and tools must meet several critical criteria that ensure compatibility with advanced AI and ML frameworks. AI readiness encompasses the following key attributes:

- **Structured Data Formats**: Data should be structured in formats that enable seamless integration with AI systems, such as relational databases, JSON-LD, or RDF formats for knowledge graphs.
- **Comprehensive Metadata**: Metadata provides essential context, including provenance, schema details, and updating logs, ensuring datasets can be accurately interpreted and processed by AI models.
- **Interoperability**: Datasets must integrate easily with other systems or datasets, often achieved through standardization and adherence to widely accepted data exchange formats.
- **Granularity and Scalability**: AI systems require data at granular levels to support diverse use cases, along with scalability to accommodate increasing data volume and complexity.
- **Temporal and Contextual Consistency**: For time-series data, such as economic indicators, datasets must represent both the time of data capture and the time to which the data applies. This ensures accuracy in temporal analyses.
- **Semantic Enrichment**: Techniques like triplification enable datasets to include explicit relationships between entities, enhancing machine understanding.

Examples of Successful Implementations. Outside of government applications, several organizations have successfully implemented AI-ready datasets and tools, including:

- **Google Knowledge Graph**: Utilizes structured triples to connect and enhance search results, demonstrating the power of semantic data organization.
- **FactSet**: A leading provider of extensive financial data and analytics, **FactSet relies on BrightQuery's structured and real-time data feeds** to power its platforms. By integrating BrightQuery's comprehensive datasets, FactSet enhances its ability to deliver actionable insights and enables decision-making in capital markets.
- **LinkedIn Economic Graph**: Maps global workforce dynamics through structured and semi-structured datasets, supporting AI-driven recommendations and insights.

For further insights, recommended resources include:

"The Role of Knowledge Graphs in AI" (https://www.croissant.ai/knowledge-graphs)
"Entity-Attribute-Value Model for Semantic AI" (https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model)
Google's Data Commons Framework (https://datacommons.org/).

How these criteria were determined. BrightQuery's team includes a number of data and AI/ML practitioners. The criteria were derived largely from our firsthand experience, asking the hypothetical question "what would I need and/or want if I were building against this data set?". BrightQuery is also building a chatbot for government data as part of an NSF grant (see: https://www.americasdatahub.org/data-access-alternatives-artificial-intelligence-supported-interfaces-daa/), so the question is not entirely hypothetical.

See also: Dept. of Commerce - Generative Artificial Intelligence and Open Data: Guidelines and Best Practices https://www.commerce.gov/sites/default/files/2025-01/GenerativeAI-Open-Data.pdf

About This Report. This report covers a human-prepared evaluation of the readiness of NCSES data sets for ingestion by AI and LLM tools. In addition to offering this assessment and providing a starting point for estimation of the effort required to transform and prepare NCSES data sets, this will serve to inform the design of the automated tool that we'll be developing in Q2 of 2025.

Request for Feedback. We are actively seeking validation of the approach we've taken by hand before we automate it and roll that methodology out across additional agencies. We expect the scoring methodology described below to change slightly over the next 3 months. We hope, *primarily*, to achieve completeness – covering all aspects that are necessary or useful for machine ingestion and model training – and, *secondarily*, to rank and score those aspects in order of importance.

Next Up. We're currently planning to evaluate the **Bureau of the Census** next. The list of agencies we've identified as high priority, following those (in no particular order):

- Bureau of Transportation Statistics
- Bureau of Labor Statistics
- Bureau of Economic Analysis
- Economic Research Service
- National Agricultural Statistics Service
- Bureau of Justice Statistics
- Center for Behavioral Health Statistics
- Statistics of Income
- Microeconomic Surveys Unit
- National Center for Education Statistics
- National Center for Health Statistics
- Office of Research, Evaluation, and Statistics
- Energy Information Administration
- National Animal Health Monitoring System

The BQ AI Readiness Scorecard

The BQ AI Readiness Score Table below provides a comprehensive evaluation of NCSES's key functional content areas—APIs, Unstructured Reports, Tables, Interactive Charts, and Website & Metadata. Each area is assessed against critical dimensions, including Scope, Data Structure & Format, Metadata & Context, Interoperability, Granularity, & Time Series Continuity, Technical Implementation, AI/ML-Specific Features, and Governance & Compliance.

These functional areas form NCSES's data ecosystem, supporting a wide range of users, from policymakers and academics to developers and AI systems. Evaluating their readiness for AI ensures that NCSES datasets and tools can meet the growing demands of data-driven innovation. By addressing gaps and leveraging opportunities for improvement, NCSES can enhance the usability, accessibility, and impact of its data assets.

Here we delve into each functional area, offering insights into their current capabilities, challenges, and potential for transformation. This structured analysis provides actionable recommendations to help search engines such as Google, and Large Language Models like ChatGPT cite NCSES's data more accurately.

Functional Content Areas

1. APIs

APIs are the backbone of modern data interoperability, offering programmatic access to datasets for seamless integration into external applications. They enable users to query, retrieve, and manipulate data dynamically, making them indispensable for building scalable, AI-powered systems. NCSES does not provide an API, though the interactive table builders do provide much of the powerful data filtering and aggregation functionality typically expected from an API.

Learn more: Understanding RESTful APIs for foundational knowledge. Google Data Commons API for an example of effective public APIs.

2. Unstructured Reports

NCSES hosts a broad range of reports, documents, and unstructured data. These documents not only capture valuable qualitative insights which complement quantitative data – including contextual explanations, detailed interpretations, and domain-specific knowledge critical for a comprehensive understanding of the data tables. In addition, they frequently capture quantitative data that isn't part of the NCSES data sets, qualitative insights regarding non-NCSES data, and wholly qualitative discussions. However, their unstructured nature poses challenges for AI systems, which thrive on structured data. Transforming these reports into AI-ready assets, through methods like triplification and semantic tagging, can unlock their potential for machine-driven insights and enhanced decision-making.

Learn more: Introduction to Semantic Annotation for enriching text documents. The Role of Knowledge Graphs in AI.

3. Tables

Tables represent one of the most fundamental and widely used data formats in NCSES's ecosystem. They organize quantitative and categorical data into structured formats, making them easily accessible for traditional analyses and machine learning models. Tables provide a foundation for AI applications, as their structure inherently supports efficient data processing. Enhancing their metadata and granularity can elevate their utility, allowing richer, more nuanced insights to be derived from AI systems.

3a. Aggregated Tables

These tables are typically the primary way of interacting with data in a structured manner. These are presented on the website and in PDF format, but more importantly as Excel or CSV downloads.

3b. Public Use Files

NCSES publishes individual survey responses (albeit anonymized, with PII stripped) in addition to aggregated data tables. This has enormous potential, though it's unable to be fully realized as it's in a semi-proprietary SAS file format.

Learn more: Data Cleaning Techniques for improving structured data. Schema.org for metadata standards.

4. Interactive Charts

Interactive charts offer dynamic visualization tools which bring NCSES's datasets to life. By enabling users to explore trends, patterns, and relationships in real time, these charts transform raw data into an engaging, user-friendly experience. For AI applications, however, these visualizations often lack the structure and accessibility required for direct integration. Improving the underlying data connections and embedding AI-ready features into these charts can enhance their role as a bridge between data exploration and actionable insights and improve public engagement through AI-driven predictive visualizations.

Learn more: D3.js Documentation for charting interactivity. Charting and AI Integration for bridging visualization and machine learning.

5. Website & Metadata

The NCSES website serves as the central hub for disseminating data to the public and stakeholders. It hosts datasets, tools, and metadata essential for both human users and AI systems. High-quality metadata is the linchpin of discoverability and usability, providing the context necessary for effective data integration and application. By enhancing the website's metadata and structural elements, NCSES can ensure its digital assets are both accessible and actionable, driving innovation and adoption across AI-powered systems.

Learn more: Schema.org Markup Guide for semantic web best practices. Open Government Data Principles for aligning with accessibility standards.

Survey Scope – Mapping NCSES Data to Scorecard Categories

Data are organized by NCSES into categories: Data Interactive, Data Tables, Data Directorate Profiles, Info Briefs, Info Bytes, Info Charts, References, Reports, and Working Papers. There are also some things that don't fit neatly into any of these categories. For our own AI/ML focus we'd suggest thinking of them in terms of a few broader categories.

- **Survey Releases**: Results of surveys released to the public, as-is. This would include data tables, whether interactive or static, and the PDF documents they originally appear in, containing an introduction and summary of the results.
- **Analysis of Survey Releases**: NCSES also publishes research and analysis pertaining to its survey results. These include discussions of survey results, interpretation, potential causal relationships. They may also include data, tables, and charts derived directly from survey results and comparisons across years.
- **Analysis of non-NCSES Data**: NCSES links to research papers which don't rely (or only rely in part) upon NCSES's own data. This may include data from other government agencies, academic institutions, or private surveys. Some documents straddle both this and the previous category – the published conclusions of research involving both NCSES and non-NCSES data.
- **Methodology Discussion**: This covers discussions of survey methodology, questions, response rates, and similar topics. These discussions frequently cover how to improve the quality and explanatory value of data being received, how to deal with methodological differences in the best way possible, and how to remediate potential bias and weaknesses in the data.

Locations Included in Our Survey

Data Explorer.
Link: https://ncsesdata.nsf.gov/explorer
This is the gateway to browse and discover data sets – surveys, files, and variables.

Browse / Search Data.
Link: https://ncses.nsf.gov/browse-library
This is the gateway to browse and discover individual documents. It links out to the below:

- Full Reports, Modern Format
- Data Tables
- Full Reports, Web Archive
- Analysis Reports

Full Reports, Modern Format.
Example: https://ncses.nsf.gov/surveys/higher-education-research-development/2023
These each include both **Survey Releases** and the **Methodology Discussions** that go along with them. The former is in a natively structured format (downloadable tables), but also available unstructured (PDF). The latter is inherently unstructured. These are consistently formatted from 2021 onward.

Data Tables (via search function).
Example: https://ncses.nsf.gov/pubs/nsf22310
These are labeled as Data Tables but typically contain the same information (both **Survey Releases** and the **Methodology Discussions)** as the Modern Format. The layout is less consistent (e.g., methodology isn't *always* present, don't have direct links to related **Analysis Reports** and **Public Use Files**), but they're near-parity. Typically, available from the late 2010s to 2021.

Full Reports, Web Archive.
Example: https://wayback.archive-it.org/5902/20150627202709/http://www.nsf.gov/statistics/rdexpenditures/datbrief/datbrief.htm
These include both **Survey Releases** and the **Methodology Discussions** that go along with them. In contrast to the Modern Format, they don't necessarily include *both* of those (sometimes both are available together, sometimes they're separate documents, sometimes only one is available, and other times neither). **Survey Releases** are almost always structured when present, though not always the same between years. This is often the only source of data prior to 2021, although it's not available for the full history (usually beginning sometime in the 1990s for most surveys)

Analysis Reports, Modern format and Web Archive.
Example: https://ncses.nsf.gov/pubs/nsb20214
These include both **Analysis of Survey Releases** and **Analysis of non-NCSES Data**, sometimes in the same document. These are broadly construed as unstructured data. While these sometimes do include structured data, it's never "primary" data, and rarely in a consistent format.

Public Use Files.
Link: https://ncses.nsf.gov/explore-data/microdata
These are **Survey Releases**, with de-identified individual record level data (as opposed to aggregated tables), available for download – there are SAS files and supplemental files.

Restricted Use Files.
Link: https://ncses.nsf.gov/licensing

These are **Survey Releases**, with individual record level data (as opposed to aggregated tables), researchers can apply for access through the Standard Application Process (SAP). We did not access this data, though we did review the policies for doing so.

SESTAT Interactive Tool.
Link: https://ncsesdata.nsf.gov/sestat/
This was included in our survey under the Interactive Tools category.

Chart Builder and Table Builder.
Links: https://ncsesdata.nsf.gov/builder/welcome?type=chart | https://ncsesdata.nsf.gov/builder/welcome?type=table
These are included in our survey under the Interactive Tools category.

Table 1. BQ AI Readiness Scorecard

✓	Complete / Good	✗	Missing / Not usable
◐ ◐ ◐	Partial (slightly, somewhat, mostly)		
NA	Not applicable	?	Unable to determine

	Functional Area					
	APIs	Unstructured Reports	Aggregated Tables	Public Use Files	Interactive Charts	Website & Metadata
Scope						
Breadth	✗	✓+ ≥ 2021	✓ ≥ 2021	◐ ≥ 1993	◐ ≥ 1993	✓ ≥ 2021
History	✗	◐ ≥ various	◐ ≥ various	◐ ≥ 1993	◐ ≥ various	◐ ≥ various
Time Series Revisions Dimension	✗	NA	NA	NA	NA	NA
AI / ML Specific Features						
Schema.org Tagging	NA	✗	✗	✗	✗	✗
Croissant Tagging	NA	✗	✗	✗	✗	✗
Triplification	NA	✗	✗	✗	✗	✗
Training Data Labels	NA	?	?	?	?	?
Bias Documentation	NA	✓	◐	◐	◐	✓
Data Structure & Format						
Digitization	NA	✓ ≥ 2021 / ◐	✓	✓	✓	✓
Standardized Schema	NA	NA	◐	◐	NA	◐
Standardized Layout	NA	✓ ≥ 2021 / ◐	NA	NA	✓	◐
Tabular Layout	NA	✓	✓	✓	✓	NA
Field/Column Labels	NA	◐	◐	✓	✓	NA
Metadata & Context						
Units Discernable	NA	◐	◐	✓	✓	✓
Data Dictionary	NA	✓	✓ (?)	✓ (?)	✓	✓
Methodology Described	NA	✓ ≥ 2021 / ✗	✓ ≥ 2021 / ✗	✓ ≥ 2021 / ✗	✓ ≥ 2021 / ✗	✓ ≥ 2021 / ✗
Release Date/Time	NA	✓ ≥ 2021 / ?	✓ ≥ 2021 / ✗	✓ ≥ 2021 / ✗	✗	✓ ≥ 2021 / ?
Interoperability, Granularity, & Time Series Continuity						
Geographical	NA	◐	◐	◐	◐	◐
Vertical (Field of Study & Sector)	NA	◐	◐	◐	◐	◐
Reporting Period	NA	◐	◐	◐	◐	◐
Methodology	NA	◐	◐	◐	◐	◐
Technical Implementation						
Direct Access Links	NA	✓	✓	✓	✗	✓
Documentation	NA	NA	NA	◐	✓	NA
Rate Limiting	NA	?	?	?	?	?
Authentication Methods	NA	✓	✓	✓	✓	✓
Open File Format	NA	✓	◐	◐	✓	✓
Governance & Compliance						
Version Control	NA	✗	✗	✗	✗	◐
Usage Rights	NA	✓	✓	✓	✓	✓
Privacy Annotations	NA	✓	✓	✓	✓	✓
Individual Linkages Possible	NA	✓	✓	✓	✓	✓
PII Action Required	NA	✓	✓	✓	✓	✓

Table 2. NCSES's AI Readiness - Summary KPIs

By Functional Area		By Category	
Overall	**59%**	**Scope**	**49%**
APIs	**0%**	**AI / ML Specific Features**	**22%**
Unstructured Reports	Recent Data: **70%** Historical: **60%**	**Data Structure & Format**	Recent Data: **83%** Historical: **81%**
Aggregated Tables	Recent Data: **67%** Historical: **56%**	**Metadata & Context**	Recent Data: **93%** Historical: **38%**
Public Use Files	Recent Data: **63%** Historical: **52%**	**Interoperability, Granularity, & Time Series Continuity**	**25%**
Interactive Charts	Recent Data: **61%** Historical: **58%**	**Technical Implementation**	**85%**
Website & Metadata	Recent Data: **62%** Historical: **54%**	**Governance & Compliance**	**81%**

By Objective	
Search Engines *Website & Metadata, Unstructured Reports* *Scope (All), Schema.org Tagging, Digitization, Field/Column Labels, Data Dictionary, Direct Access Links, Rate Limiting, Authentication Methods, Open File Format, Privacy, PII*	Recent Data: **83%** Historical: **82%**
LLMs *Website & Metadata, Unstructured Reports* *Scope (All), AI/ML Specific Features (All), Data Structure & Format (All), Metadata & Context (All), Direct Access Links, Rate Limiting, Authentication Methods, Open File Format, Usage Rights, Privacy, PII*	Recent Data: **76%** Historical: **64%**
BrightQuery Chatbot Project (DAA-24) *All functional Areas* *Scope (All), Triplification, Training Data Labels, Bias Documentation, Data Structure & Format (All), Metadata & Context (All), Interoperability, Granularity, & Time Series Continuity (All), Technical Implementation (All), Governance & Compliance (All)*	Recent Data: **68%** Historical: **59%**
ML & Predictive Analytics *APIs, Aggregated Tables, Public Use Files* *Scope (All), Triplification, Bias Documentation, Data Structure & Format (All), Metadata & Context (All), Interoperability, Granularity, & Time Series Continuity (All), Open File Format, Governance & Compliance (All)*	Recent Data: **64%** Historical: **51%**

Description and Importance of Criteria

Scope

Breadth. Does the functional area cover all data sets made available? This evaluates whether all datasets in one format are also available in others. Being able to gather all necessary data from one functional area makes the ingestion process easier; having to collate data from multiple increases effort required. In addition, having all of the same data specified in multiple formats increases the ability of machine-learning systems to check their own work. Of course, if some data isn't available online at all, it can't be ingested.

History. This determines if datasets include their complete historical range in each format.

Time Series Revision Dimension. In certain data sets, time is inherently two-dimensional: the time **at which** the data was produced and the time **to which** the data refers. This is typically expressed as estimates, revisions, and finalized values – in which there are multiple reported values for the same data point and time period (i.e. US GDP for Q4 2023). This does not currently apply to the main NCSES data sets.
See **Appendix B: Modeling Time Series data with Revisions.**

AI / ML Specific Features

Schema.org Tagging. Schema.org is a widely adopted semantic vocabulary used to add structured metadata to web content. By embedding Schema.org tags into datasets and web pages, this metadata enhances search engines' understanding of the data's context, structure, and relevance. Is Schema.org metadata present? Is it complete? This is a standard set of metadata which various commercial web crawlers look for and use to understand documents. While it's not sufficient for our purposes, it will mean more relevant search results for people coming from the outside internet.

BrightQuery is uniquely positioned to implement Schema.org tagging through its collaboration with Schema.org. This partnership allows NCSES to benefit from cutting-edge expertise in metadata enrichment, enabling its datasets to achieve maximum visibility and utility across both search engines and AI applications.
Learn more: Schema.org Docs

Croissant Tagging. Croissant is an advanced metadata framework tailored to AI and machine learning applications. Unlike Schema.org, which focuses on search engine optimization, Croissant tagging is designed to make datasets more discoverable and usable by AI systems. Is Croissant Metadata present? Is it complete? This is a standard set of metadata which various commercial web crawlers look for and use to understand tabular data. While it's not sufficient for our purposes, it may improve the performance of commercial LLMs over the long term.

BrightQuery's partnership with MLCommons, brings unparalleled expertise in implementing Croissant tagging to optimize datasets for AI and machine learning applications. By actively participating in MLCommons, BrightQuery can help to ensure statistical products are aligned with the highest standards for discoverability and usability by advanced AI systems. This collaboration also leverages MLCommons' network, including Microsoft, Google, Meta, and OpenML, to ensure the tagging framework is robust, interoperable, and future-proof, empowering NCSES to deliver cutting-edge data solutions across a wide array of analytical domains.
Learn more: MLCommons Croissant

Triplification. In the context of AI/ML, "triplified data" refers to data that is represented in the form of triples, which are structured as subject-predicate-object. This format is commonly used in knowledge graphs and semantic web technologies to express relationships between entities in a way that is both machine-readable and semantically rich. This is described in more detail in **Appendix A.**

Training Data Labels. These are question/answer pairs designed to aid in developing and validating lexical search results and natural-language interfaces, for example, "how has the employment landscape for computer engineers changed over the past decade?" These question-and-answer pairs are crucial for the translation of complex economic data into meaningful answers, either to retrieve the proper data set in a lexical search or carry out a conversation in a chat-style interface.

Bias Documentation. Is an explanation of likely statistical bias provided? For example, a land-line phone survey in the early 2000s would have under-sampled young adults. This will be useful for predictive modeling or for more advanced use cases, such as those in which the system draws conclusions. More prosaically, it should be presented to humans who will make those determinations, in order to raise awareness of potential blind spots.

Data Structure & Format

Digitization. Data not digitized must be processed using OCR (optical character recognition) software. While this technology is relatively mature, it does require additional effort and processing power, while introducing a small chance of transcription errors. Verifies if data is alphanumeric and not an image.

Standardized Schema. Checks if data adheres to a consistent schema. A consistent data schema which rarely or never varies allows for processing code to utilize a single, simpler, algorithm. This greatly increases testability, accuracy, and reliability.

Standardized Layout. Evaluates whether datasets are consistently organized. A consistent layout and document/HTML structure which rarely or never varies allows for processing code to utilize a single, simpler, algorithm. While not quite as useful as a structured schema, this still greatly increases testability, accuracy, and reliability compared with no standardization at all.

Tabular Layout. Assesses whether data is presented in table format. Tabular layouts are simpler to process and verify, particularly when the layout is not otherwise standardized.

Field/Column Labels. Ensures clear naming conventions for columns. Clear labels can be read unambiguously by a machine, as opposed to labeling schemes requiring context and cross-referencing.

Metadata & Context

Units Discernable. Verifies that units (e.g., billions of dollars, percentages) are clearly specified. This is particularly important when making comparisons or calculations across data sets, or when a natural language question requires conversion.

Data Dictionary. Determines if comprehensive descriptions of terms are available. A data dictionary will greatly improve the performance and accuracy of lexical searches when a user doesn't know the exact terminology, or would like a list of choices, by providing descriptions (which can be further enriched with synonyms).

Methodology Described. Checks for detailed explanations of data collection methods. This is useful for simple presentation and explanation to users; it can also be used by more advanced LLM-based systems to speculate as to possible biases and the relevance of certain comparisons between data sets or across time.

Release Date/Time. Assesses if the publication date and revision history are clear.

Interoperability, Granularity, & Time Series Continuity

Interoperability allows for making comparisons and drawing conclusions across multiple data sets. All other things considered, an interoperable data set along more dimensions at a higher granularity will be more useful in conjunction with others. It's unlikely there will be much to be done about non-interoperable data in the near term.

While a lack of interoperability doesn't affect the ability to ingest and analyze data in isolation, it limits the ability to ask questions like "how does [national trend] affect the economy in [city]" and get meaningful and precise answers.

Geographical. Is the geographic breakdown compatible across all data points? Between this data set and those from other agencies? For example, county-level data can be aligned with state-level data, but metropolitan statistical area -level data cannot (as MSAs may include parts of multiple states).

Vertical: Field of Study & Industry/Sector. Is the industry/sector breakdown compatible across all data points? What about fields of study? Between this data set and those from other agencies? For example, NAICS is compatible with SIC, but not with GICS. A free-text industry description would not be compatible with either (unless it corresponds directly to codes specified in a data dictionary).

Reporting Period. Is the reporting period breakdown compatible across all data points? Between this data set and those from other agencies? For example, monthly or quarterly data can be aligned with the Federal Government's fiscal year; annual data aligned to the calendar year (covering Jan-Dec) cannot.

Methodology. Is the methodology compatible across all data points? Between this data set and those from other agencies? Across time? For example, the Census has changed or added race/ethnicity choices on multiple occasions; this creates fissures across which data cannot be compared on a like-for-like basis.

Technical Implementation

Direct Access Links. Is it possible to link directly to a data item or table, or is source information buried in a page with numerous other tables and/or tests? What about a single number? Direct links will be necessary for a system to cite its sources. It also greatly helps the QA process.

Documentation. Specifically, technical documentation. Is it possible to use an API or interactive tool effectively without technical support from agency personnel, extensive guessing, or trial and error?

Rate Limiting. Are rate limits clearly described and high enough to allow full ingestion of the data set? Rate limits enable researchers to download comprehensive datasets without overloading the system.

Authentication. Are there access controls? If so, is the sign-up free and quick? Is there a waiting period? These all relate to the ease of data ingestion and processing, and the amount of engineering time and effort required.

Open File Format. Do the files use a format that allows it to be read, by humans or machines, using free and readily available software?

Governance & Compliance

Version Control. Are updates and revisions to published data and their contents tagged and auditable? *Not typically relevant for data sets (e.g., economic indicators) in which the published date is a more fundamental property of the data.* This will be important for testing and ensuring the accuracy of our data retrieval or simply keeping our system up to date. Records of when what changed, how, why, and by whom can also build public trust.

Usage Rights. Are there restrictions on use or publication of the data? *Not typically relevant for government data sets; this is more commonly applied to data compiled by for-profit companies.*

Privacy Annotations. Is private or personal data clearly marked? It's crucial to avoid publishing or using (even for training) private data inappropriately. More prosaically, data without open usage rights can't be published in an openly accessible tool.

Possible Individual Linkages Is non-anonymized information present at the individual level, in such a way as to enable individual linkages across data sets? When possible, it opens entirely new categories of question for study or analysis which previously would have required dedicated one-off research efforts. How does an *individual's* response to a consumer sentiment survey relate to their spending patterns? Their educational outcomes? Their health? What effect does losing a driver's license have on an individual? Does this differ across geographies? Age groups?

PII Action Required. Must private or personal data be redacted before publication? May it be shared publicly or between agencies? May data be shared if it undergoes an anonymization process? An aggregation process? *Action will almost always be required if PII is present; exceptions would be made for data which is already public (for example, FEC donations).* This greatly increases the amount of effort needed for security, testing, and responsible publishing of data. Data with PII (which isn't otherwise public information) can't be fed into machine-learning systems at all, in many cases.

Analysis Details

Scope

Breadth. All main data sets appear available as data tables, unstructured reports, and on websites like HTML, for at least some (usually the most recent) time frames. *This comes with the standard caveat that, were the NCSES to publish results of a survey, and not make it available in any format or any location on their website, it would not be included in this analysis.*
Some unstructured reports rely, in part, on data that's not available in structured tables. *We decided to score this as "extra credit" for the reports, rather than penalize the structured tables, as those cases aren't using data, we'd expect to find on the NCSES website. (For example, international comparisons rely, in part, on data from the World Bank. Some reports also incorporate private data sets which are not available to the general public)*
The interactive data tools do not, however, cover all areas. For example, we found the following to be missing:

- Survey of State Government R&D
- Annual Business Survey (ABS)
- Business Enterprise Research and Development (BERD) Survey
- Nonprofit Research Activities (NPRA) Survey

History. Some key surveys, including the SDR and NSCG, began in the 1970s. However, data available from the NCSES varies on the dates it becomes available, depending on format:

- **Interactive Data Tools:** Data is not available prior to 1993
- **Aggregated Tables:** The date at which aggregated tables become available varies widely. There was no clear cutoff date we could discern; for instance, this is 1993 for the SDR and 2010 for the NSCG. *
- **Public Use Files:** Data is not available prior to 1993
- **Unstructured Reports:** The date at which unstructured reports become available varies widely; while the original **Survey Releases** containing the tables are not always accessible as far back as the raw data, *other* publications including **Analysis** of both **NCSES Survey Releases** and **non-NCESES Data** may be available prior to the earliest corresponding tables * **
- **Website & Metadata:** The date at which data becomes available or searchable on the website varies widely; a number of the aforementioned reports and tables are archived, linking to legacy formats stored on https://wayback.archive-it.org

** Some reports and data tables also contain (restated) figures from earlier reports, though these are not comprehensive – for example, the Survey of Federal Funds for Research and Development shows "Federal obligations for research and development, by type of R&D, and for R&D plant" all the way back to 1951, while the other tables only show the current year.*

*** While strictly speaking, the earliest available unstructured documents sometimes predate the earliest available corresponding data tables, we've given it a lower score because the specific reports in which the tables were originally published are not always available in their entirety.*

As an overall finding, there are few clear cutoff dates for which a type of data becomes available or discoverable via search functions. Data is well organized and consistently available from 2021 onward. Interactive data tools and Public Use Files *typically* go back to 1993. Before 2021, however, there is little discernable pattern – "what data is available in what formats and where do I find it?" seems answerable only by guess-and-check.

Time Series Revision Dimension. This is not applicable to NCSES data, as it's not typically revised in the same manner as other government data sets (e.g., GDP and other economic indicators). This topic is similar to but distinct from corrections. NCSES does issue corrections to specific errors discovered in data releases; these corrections are discussed under **Governance & Compliance > Version Control**.

AI/ML Specific Features

Schema.org Tagging. We did not find any Schema.org metadata present on NCSES's website.

Croissant Tagging. We did not find any Croissant metadata present on NCSES's website. Note that applying Croissant metadata would require some minor restructuring of Aggregated Data Table downloads; **See Appendix C: Primer on Schema.org and Croissant Standards** for more detail and an example. This would also dovetail with our recommendation to include CSV in addition to Excel file downloads.

Triplification. We did not find any Triplified data present on NCSES's website. While the vast majority of triplification, at least within the context of each table, can be done with software tools (given the tabular data or API), there is still a non-trivial amount of engineering effort required to ingest this data into a graph. There is also a considerable amount of effort involved in determining and building graph relationships out of the relationships that exist *between* data tables and their line items. This is increased or reduced by the level of standardization of schemas or formats.

Training Data Labels. Pre-built training data labels are not available for NCSES datasets, which is consistent with the broader context of government data. BrightQuery is working with NCSES personnel to develop these labels, which are essential for training AI systems effectively.

Bias Documentation. Survey methodology is discussed, including topics such as potential sampling errors, impacts of this bias, and remediation steps taken to control for deficiencies. This discussion is robust, tied directly to each survey, and specific to each year (in recent years; for older data tables, the corresponding discussion may be hard to find or not present).

Data Structure & Format

Digitization. The vast majority of data discovered was fully digitized. There were, however, a handful of historical reports and documents in the archive that were only available as non-text PDF, or which included figures as scanned images (with no corresponding digitized data tables).

Standardized Schema. Tables are standardized for the latest years; however, the formats vary wildly in previous years. Much historical data is available only in SAS files, with a slightly different schema for each year of each survey.

Standardized Layout. The practice in more recent years of giving tables a unique identifier that's comparable across years is helpful (e.g., "Table 1-1: College graduates, by level of highest degree, minor field of highest degree, and labor force status: 2023"). Finding things in the archive can be tricky due to inconsistent naming and labeling conventions, however (e.g., "I want to see Table 1-1 from the NCSG across each year available")

Tabular Layout. Across all formats, including unstructured data, key information is consistently presented in a tabular format rather than buried within prose. Even the oldest plain-text documents display data with a clear tabular layout, making it easy to interpret.

Field/Column Labels. Field and column labels are generally present across data formats. Historical data in SAS files is extremely difficult to interpret and determine labels for, however, without proprietary third-party software.

Metadata & Context

Units Discernable. Units are marked in a way that is clear to a human reader, though with some inconsistency in the format (sometimes units are mentioned in a header, or outside of the table altogether in the header). Public Use Files may require cross-referencing across multiple items in a zip package in order to determine units.

Data Dictionary.
A full data dictionary is available here, in HTML format: https://ncsesdata.nsf.gov/explorer/variables
These definitions are also helpfully linked from the Chart Builder tool.

We were able to obtain a structured download of the entire data dictionary. *This link was constructed by inspecting XHR requests and modifying the page size by hand:*
https://ncsesdata.nsf.gov/api/v1/metadata/variables?profile=summary&page=0&size=1000&key=header.payload.signature

Methodology Described.
Recent surveys have their methodology described in-depth and directly linked to the data itself. However, we were unable to discover similar descriptions and discussions for historical or archived data releases.

Release Date/Time.
These are provided and appear accurate for very recent data.
For unstructured reports and web pages prior to the early 2020s, however, we're unsure whether the publication times are accurate and/or meaningful. *One common pattern is reports with headers such as "This report provides data from the 2019 Annual Business Survey (ABS) (data year 2018)" with a publication date in 2022. We surmise that this report was, in fact, published in 2019 and the 2022 date is when the file happened to be uploaded (or re-uploaded) to the current CMS.*
Bulk data contained in the Public Use Files is not timestamped at all.

Interoperability, Granularity, & Time Series Continuity

Geographical.
Geographical classifications are difficult to align with other data sets. Most formats only show broad regions (Middle Atlantic, Mountain, etc.), while other data sets (NCES, for example) will often provide data at the state level. Aligning institutions with locations (e.g., in the HERD survey) requires a full download of the Public Use Files.
Data does appear to be *collected* at a much more granular level, including zip codes and even street addresses, though this is not presented publicly for obvious privacy reasons.

Vertical: Field of Study & Industry/Sector.
Categories such as field of study, employer industry/sector, and job code change frequently from year to year. These fields are sometimes missing entirely, making meaningful comparisons across years difficult.

Reporting Period.
Most data sets are annual; some are biennial. Seasonal differences because of changes in reference month from year to year create an extra complication, as does the use of fiscal years.

Methodology.
Most NCSES data sets have major methodological breaks. Some are quite frequent, with significant changes in every survey iteration.

Data that is not comparable year-to-year presents a problem for machine-learning systems that's similar to that which human analysts face; comparing or charting across multiple years leads to analyses that aren't necessarily valid, an knowing that requires an extensive read of footnotes. An AI system, however, will rarely be able to do more with those footnotes aside from present them to a human (in a good-case scenario).

An example with a small number of "breaks": https://ncses.nsf.gov/surveys/doctorate-recipients/2023#technical-notes_data-comparability
An example with a larger number: https://ncses.nsf.gov/surveys/higher-education-research-development/2023#technical-notes_data-comparability

Technical Implementation
Direct Access Links.
Direct links are available for all reports, downloads, tables, and files. The interactive data tools, however, do not generate unique links with the embedded parameters (though the parameters can be exported to a file).

Documentation.
The chart builder offers a "help" sidebar and links to a knowledge base.

Rate Limiting.
Rate limits are undocumented, making it difficult to evaluate their impact. Testing limits through responsible interaction is critical, as aggressive testing methods, like simulating a DDoS attack, are neither feasible nor appropriate.

Authentication.
No authentication is required, enabling easy public access.

Open File Format.
Public Use Files are in SAS format. While this format isn't *entirely* opaque to those without proprietary software (they're stored as delimited plain text and could be read with custom software or by hand), it's impractical to use without a SAS license. Aggregated Tables are downloadable in Excel and PDF – PDF isn't well-suited to analytical workflows, and Excel is technically proprietary, though widely available free tools will open Excel files.

Governance & Compliance
Version Control.
NCSES publishes a list of corrections since 2010, but this is not in a version control system, nor can a revision history be accessed. For older and unstructured data, we are advised: "In some instances, prior year data have been modified based on discrepancies noted during the consistency reviews of the data across years. To obtain accurate historical data, data users should use only the most recent publication, which incorporates corrections agencies have made in prior year data. Do not use previously published data."
https://ncses.nsf.gov/corrections
https://ncses.nsf.gov/surveys/federal-support-survey/2022#methodology

Usage Rights.
NCSES publishes aggregated statistics and reports that are entirely unrestricted in terms of usage rights, ensuring open access for researchers, policymakers, businesses, and the public. This open-use policy supports a wide range

of applications, from academic studies to AI-driven analyses, without the need for licensing or additional permissions.
https://catalog.data.gov/dataset/?publisher=National+Center+for+Science+and+Engineering+Statistics

Privacy Annotation.
PII Action Required.
Some interactive tools allow the user to run queries against individual responses. These responses are aggregated and blocked if the number of responses in a given grouping is low enough to compromise privacy.
Running individual linkages against restricted data may reveal private information to researchers; this is covered by a review and approval process that will minimize the risk of data leaks and ensure that data and analyses are only published in aggregate (or otherwise not compromising privacy).

Possible Individual Linkages.
There is a clear policy and human review process to access restricted data, which contains individual's responses and may compromise privacy.
https://ncses.nsf.gov/licensing

AI Readiness Action Items

The **AI Readiness Action Table** outlines tailored solutions to address the gaps identified in the BQ AI Readiness Scorecard, focusing on enhancing NCSES's data ecosystem to meet the demands of AI-driven insights. This table serves as a roadmap, detailing actionable steps designed to improve each functional area's performance across key dimensions: *Scope, Data Structure & Format, Metadata & Context, Interoperability, Granularity, & Time Series Continuity, Technical Implementation, AI/ML-Specific Features, and Governance & Compliance.* Our suggestions prioritize based on the perceived importance and prominence of each data set, the likely value of a change for data consumers, along with an educated guess as to the difficulty and cost of implementation.

Framework for the Solutions

The solutions in the **AI Readiness Action Items Table** are designed to:

- **Close Gaps**: Address specific shortcomings identified in the AI Readiness Headline Results.
- **Enhance Usability**: Improve accessibility, interoperability, and contextual clarity for end-users and AI systems.
- **Future-Proof Datasets**: Equip datasets with AI-specific features and compliance frameworks to meet evolving demands.
- **Maximize Impact**: Prioritize efforts on high-value datasets to demonstrate measurable improvements in NCSES's AI readiness.

Each functional area within the **AI Readiness Action Items Table** is mapped to these datasets, ensuring the proposed solutions are both actionable and strategically aligned with NCSES's goals.

Guiding Principles

The solutions proposed adhere to the following principles:

- **Scalability**: Solutions must scale across datasets and functional areas to maximize efficiency and reduce redundancy.
- **Interoperability**: Enhancements should align with federal and international data standards, such as schema.org and W3C.
- **AI Optimization**: Proposed actions emphasize features such as semantic tagging, metadata enrichment, and advanced querying capabilities.
- **Compliance and Governance**: Solutions address data auditability, bias mitigation, and adherence to privacy standards, ensuring ethical AI use.

Expected Outcomes

Implementing these action items will result in:

- Enhanced AI readiness for NCSES's datasets, enabling more accurate and timely insights.
- Improved discoverability and usability for a broader audience, including policymakers, researchers, and AI systems.
- Demonstrated leadership in adopting AI-ready standards for federal economic datasets.

Table 3. BQ AI Readiness Action Items

Low Effort, High Reward
Low Effort, Low Reward
Moderate Effort, Worthwhile Reward
High Effort, High Reward
High Effort, Low Reward
NA No Action Recommended

	Functional Area					
	APIs	**Unstructured Reports**	**Aggregated Tables**	**Public Use Files**	**Interactive Charts**	**Website & Metadata**
Scope						
Breadth	⚙	NA	NA	💎 / ?	⚙	NA
History	⚙	💎 💎 💎	💎 💎 💎	💎 💎 💎	💎	💎
Time Series Revisions Dimension	NA	NA	NA	NA	NA	NA
AI / ML Specific Features						
Schema.org Tagging	NA	⚙	⚙	⚙	⚙	⚙
Croissant Tagging	NA	⚙	⚙	⚙	⚙	⚙
Triplification	💎 💎	🛑	💎 💎	💎 💎	🛑	🛑
Training Data Labels	NA	💎 💎 💎	💎 💎 💎	💎 💎 💎	💎 💎 💎	💎
Bias Documentation	🛑 / ⚙	NA	🛑	🛑	🛑	NA
Data Structure & Format						
Digitization	NA	🛑	NA	NA	NA	NA
Standardized Schema	💎 💎 💎	NA	💎	💎	NA	🌟
Standardized Layout	NA	🛑	NA	NA	NA	🌟
Tabular Layout	NA	NA	NA	NA	NA	NA
Field/Column Labels	🌟	🛑	🌟	🌟	NA	NA
Metadata & Context						
Units Discernable	🌟	🛑	🫧	NA	NA	NA
Data Dictionary	🌟	NA	🌟	🌟	NA	NA
Methodology Described	?	?	?	?	?	?
Release Date/Time	?	?	?	?	?	?
Interoperability, Granularity, & Time Series Continuity						
Geographical	💎	💎	💎	💎	💎	💎
Vertical (Field of Study & Sector)	💎 / ?	💎 / ?	💎 / ?	💎 / ?	💎 / ?	💎 / ?
Reporting Period	💎	💎	💎	💎	💎	💎
Methodology	💎 / ?	💎 / ?	💎 / ?	💎 / ?	💎 / ?	💎 / ?
Technical Implementation						
Direct Access Links	🌟	NA	NA	NA	🫧	NA
Documentation	🌟	NA	NA	NA	NA	NA
Rate Limiting	NA	🫧	🫧	🫧	🫧	🫧
Authentication Methods	NA	NA	NA	NA	NA	NA
Open File Format	NA	NA	🫧	💎 💎 💎	NA	NA
Governance & Compliance						
Version Control	💎	💎	💎	💎	💎	💎
Usage Rights	NA	NA	NA	NA	NA	NA
Privacy Annotations	NA	NA	NA	NA	NA	NA
Individual Linkages Possible	NA	NA	NA	NA	NA	NA
PII Action Required	NA	NA	NA	NA	NA	NA

Action Items Details

Scope

Breadth.

Adding an API could be enormously beneficial – but not as a box-ticking exercise – we would only recommend doing so if it were to provide some data or functionality not currently available through other means (for example, full history, a standardized schema, version control & revisions).
Expanding the interactive chart builder tool to include all data sets would be beneficial for some lay users.

History.

To improve historical data accessibility, one key recommendation stands out:
Archive Expansion: All reports should be included in either the Public Use Files, data archive, or on each survey's page.
Ideally, these would be available as structured data – either in tables, downloadable files, or an API. Even if this is limited to scanned PDFs, having all NCSES survey results available in one place would still be beneficial. (some of these are available in disparate sources across the web, for example https://eric.ed.gov has SED results not available via the NCSES directly)

Time Series Revision Dimension.

While this isn't relevant for NCSES data in the same sense as it is for economic indicators – the NCSES is not entirely immune to corrections being required of previously published data. See the closely related concept of **Governance & Compliance > Version Control.**

Data Structure & Format

Schema.org Tagging.

Currently, Schema.org tagging is not implemented within NCSES's data infrastructure. The addition of this metadata framework would significantly enhance discoverability and semantic integration, particularly for search engines and AI applications.

Croissant Tagging.

Croissant tagging is also absent from NCSES datasets. Implementing this advanced metadata framework would improve usability for machine learning systems by enabling richer, context-aware data interpretation.

Triplification

The creation of a data set that was formatted as triples, for use in a knowledge graph, would greatly increase the ease of working with NCSES data for any consumer going down that path – as it currently stands, consumers will have to triplify data themselves to load it into a graph database, which puts it out of the reach of individuals or small teams for more than a handful of data points.
This would apply to both the relationships between data tables and the data points within them, and the data cells themselves. The former would be a one-time, largely manual (perhaps aided by an LLM) effort; the latter could be derived from the data tables as they were published (with a single piece of code for all data releases going forward, given a consistent format).
We would recommend focusing primarily on the effort of triplifying data that is already structured/tabular at this point.

Training Data Labels.

We're asking every agency we work with to provide a list of question/answer pairs for our natural language search tool and chatbot. This could also be made publicly available for other teams, including commercial AI chat tools.

Bias Documentation.

The discussion of methodology and potential bias in NCSES data sets is already strong. The improvements that could be made are:

- Adding this to Croissant metadata
- Providing this guidance for historical data, if possible

Data Structure & Format

Digitization.

For older research documents, preprocessing with optical character recognition (OCR) can enhance accessibility if native digital versions are unavailable. However, this is considered a lower priority compared to other action items.

Standardized Schema.

Field/Column Labels.

We would recommend a master list of columns, consistently applied across all downloads (or an API, were that to be developed), which would allow for easy comparisons across files. The bulk of the work here appears to have been done (Qlik Name, Public Use File SAS Name), they're just not labeled in non-SAS table downloads.

Standardized Layout.

Some standardization of layout, navigation, search, and nomenclature on the website and search functionalities would go a long way to both human and machine navigation and comprehension. For instance, a single listing of all Public Use Files for all surveys, a single listing of all data tables filterable by survey, a single listing of all original survey reports filterable by survey, consistent usage of survey names and abbreviations.

Tabular Layout.

No action is recommended.

Metadata & Context

Units Discernable.

Data Dictionary.

Adding units and data item Qlik/SAS codes to the non-SAS downloads would improve interpretability without required cross-referencing.

Methodology Described.

If methodology documentation regarding historical surveys' results is as readily available (internally) as for the most recent ones, making it accessible on the website is recommended. If not, this may be impractical or impossible to reproduce.

Release Date/Time.

To improve Metadata on release dates, we recommend embedding structured HTML metadata, such as:

```
<meta property="article:published_time" content="YYYY-MM-DDTHH:MM:SSZ">
```

If the original publication date of historical surveys' results is readily available (internally), exposing it visibly and within metadata is recommended. If not, this may be impractical or impossible to recover.

Interoperability, Granularity, & Time Series Continuity

Geographical.

NCSES data is available down to the zip code level. While it's unlikely that data could be provided publicly at that level of granularity without compromising privacy (or scrubbing enough cells as to render the output useless), an aggregation at the state or Metropolitan Statistical Area level would greatly improve the ability to related NCSES data to other data sets, if possible to produce responsibly.

Reporting Period.

While annual data is always preferable to biennial, we can safely assume this is known and the relevant cost/benefit considerations have been made. More granular reporting periods (quarterly, monthly) would be of limited (if any) use for most NCSES data sets. Alignment with fiscal years makes some alignments with other data sets difficult, though this is unavoidable. Consistency is reference months, however, may improve comparability year-to-year and with other data sets.

Vertical: Field of Study & Industry/Sector.

Methodology.

While many changes from year to year are expected, and are inherent to a changing education landscape, NCSES data has an abnormally high number of them, compared with other data sets. Data series are frequently not comparable across even a few years (let alone other, non-NCSES, data sets) without large caveats and footnotes.

While we don't presume having the full picture of the tradeoffs considered in making these changes, further consideration may be warranted. There are a few common techniques, however, some of which are already in use in some places in NCSES data:

- Changed methodologies could be presented as new times series / variable names / identifiers. For example, REGION_CODE, REGION_CODE_2010, REGION_CODE_2015. The Data Dictionary could describe the differences in methodology. This would allow for charting a "broken" chart with three time series, and make it difficult to *not* notice the lack of comparability and misinterpret an apparent trend.
- (not mutually exclusive) In the survey(s) immediately proceeding a methodology change, present both old and new numbers. This would allow for a period of comparability, and still allow for switching to an improved methodology. For example, the 2015 survey publication could include *both* a REGION_CODE_2015, which improves data collection, and a REGION_CODE_2010 which deliberately reproduces the flaws of the older methodology, so as to allow a comparison with the prior years.
- When it's possible to avoid methodological changes, this is of course the simplest solution. Ideally, old responses could be re-analyzed and updated to use current methodology, though that is of course often impossible. There may also be situations in which a small methodological improvement could be postponed to coincide with more in the future to reduce the number of breaks, though this of course carries a cost to accuracy.
- Adopting widely used standards for categorization, such as CIP (education), SOC (occupation), NAICS (industry), ISO (region), etc. will ease alignment with other data sets in the short term, and may reduce the need for changes in the long term.

SDMX provides further guidance around handling methodological changes and breaks in times series.

- Guidance: https://sdmx.org/wp-content/uploads/GUIDELINES_FOR_REPRESENTING_METHODOLOGICAL_CHANGES_IN_DSDs_version_1_0.docx
- An example: https://sdmx.org/wp-content/uploads/CL_AREA_2_0_March_2019.docx

Technical Implementation

Direct Access Links.

Both the SESTAT tool and the Chart Builder expose concise expressions of the query that's been run (SQL in the former, and exportable JSON in the latter). These are most likely small enough to be serialized into a URL parameter that would enable sharing a link.

Documentation.

No action is recommended for documentation, though if an API were to be developed, documentation would be strongly recommended.

Rate Limiting.
If limits exist or are desired, a "Crawl-delay" directive should be added to the robots.txt file at https://ncses.nsf.gov/robots.txt to guide automated access appropriately and prevent server overload.

Authentication.
No action recommended.

Open File Format.
The full survey results available in the Public Use Files would be more accessible if they were provided in another format (in addition to SAS). We would also recommend CSV downloads in addition to Excel.

Governance & Compliance
Version Control.
We would recommend, at minimum, that structured data files (Excel downloads, Public Use Files, and an API if developed) be stamped with their initial creation date, last revision date, version number, and a checksum. This will allow crawlers and other machine consumers to ensure they have the latest data without re-downloading entire files. See the Croissant specification, here: https://docs.mlcommons.org/croissant/docs/croissant-spec.html#dataset-versioningcheckpoints

Further improvements could be made with additional effort – updating the unstructured reports with corrections and maintaining a version history with all prior (overridden) versions of each file.

Usage Rights.
Privacy Annotation.
No action recommended.

Possible Individual Linkages
No action recommended.

PII Action Required.
No action recommended at this time. As a longer-term objective, options could be explored that would allow for the publication of synthetic data sets, which would contain more detail at an individual record level without compromising individuals' privacy.

Conclusion

This report serves as a step forward in the National Center for Science and Engineering Statistics (NCSES) journey toward AI readiness, reflecting BrightQuery's commitment to enabling seamless integration of NCSES's high-priority datasets into generative AI frameworks. By transforming the NCSES's statistical products into machine-readable and machine-understandable formats, NCSES can unlock new opportunities for enhancing data discovery, usability, and trustworthiness in AI-driven applications. This transformation will equip generative AI technologies to return accurate, authoritative insights, benefiting policymakers, researchers, and the public alike.

Key Findings
Our assessment identified critical gaps and opportunities across functional areas such as APIs, Unstructured Reports, Tables, Interactive Charts, and Website & Metadata. These areas form the foundation of NCSES's data ecosystem but require enhancements in metadata quality, interoperability, and technical implementation to fully support AI applications.

A few key areas stand out:

- **Public Use Files:** NCSES publishes individuals' responses to surveys (with personally identifiable information removed) that allows for more advanced data analysis than would be possible with only the aggregated information.
- **Restricted Use Files:** NCSES provides a clear procedure to request access to query that withheld information, and to do so in a way that minimizes risks to privacy.
- **Functional: No API Present.** While not necessary, an API could present a data consumer with a unified way to discover and download structured data for all survey sets across all history. It is a very particular class of user that uses an API – the *exercise* of creating an API, along with surfacing all of the data that would go into it, may be more valuable that the API itself.
- **Criteria: History.** Data from 2021 onward is very strong across most dimensions. As one travels back in time, however, data availability and formatting deteriorates rapidly – this effect is universal across all data sets, though the manner and specifics are different. Several decades of reports and data are simply unavailable, from the NCSES or elsewhere on the internet. This is more critical for NCSES data than economic data sets (e.g., BEA) as those reports are typically archived by third parties.
- **Criteria: Time Series Consistency.** Time series are frequently "broken" as methodological changes from one year to the next render data points incomparable. This happens numerous times across several surveys, and in multiple dimensions. While we are unable to speak to the merits and tradeoffs of any individual change, the overall result makes it extremely difficult to perform meaningful analysis of trends over time or even year-over-year comparisons in the same way that is commonplace for other data sets (for example, economic data).
- **Theme: Discoverability and Navigation:** As an overall finding, there are few clear cutoff dates for which a type of data becomes available or discoverable via search functions. Data is well organized and consistently available from 2021 onward. Interactive data tools and Public Use Files typically go back to 1993. Before 2021, however, there is little discernable pattern – "what data is available in what formats and where do I find it?" seems answerable only by guess-and-check across multiple navigation, search, listings, and archive formats.
- **Criteria: AI / ML Specific Features.** Schema.org and Croissant metadata standards should be applied to enhance automated discoverability and machine understanding of NCSES data. Triplified data lowers the barrier to entry for any consumer looking to build a knowledge graph out of NCSES data.

Using BrightQuery's AI Readiness Scorecard, we evaluated each functional area across several dimensions, assigning an overall readiness score of **59%**. The individual scores below provide a clearer picture of where improvements are most needed.

Functional Areas

- **API**: No API Present.
- **Unstructured Reports**: Recent data scored **70%**, while historical data scored **60%**, reflecting challenges in availability, discoverability, and interoperability/granularity.
- **Aggregated Tables**: Recent data scored **67%**, while historical data scored **55%**, reflecting challenges in availability, discoverability, and interoperability/granularity.
- **Public Use Files**: Recent data scored **63%**, while historical data scored **52%**, reflecting challenges in availability, discoverability, interoperability/granularity, and the SAS file format.
- **Interactive Charts**: Recent data scored **61%**, while historical data scored **58%**, reflecting challenges in availability, discoverability, and interoperability/granularity.
- **Website & Metadata**: Recent data scored **62%**, while historical data scored **54%**, reflecting challenges in availability, discoverability, and interoperability/granularity.

Criteria Groups

- **Scope.** The score of **49%** primarily reflects the limited availability of historical data.
- **AI/ML Specific Features**: The score of **22%** primarily reflects the lack of structured metadata.
- **Data Structure & Format**: Recent data scored **83%**, while historical data scored **81%**, reflecting strong digitization, but schema and navigation challenges.

- **Metadata & Context**: Recent data scored **93%**, while historical data scored **38%**, reflecting the historical deterioration in data availability mentioned above.
- **Interoperability, Granularity, & Time Series Continuity**: The score of **25%** may reflect an opportunity to release data that's more granular and consistent, or it may reflect an unavoidable reality of tradeoffs that must be made.
- **Technical Implementation**: The score of **85%** reflects a website with few barriers to access and navigating.
- **Governance & Compliance**: The score of **81%** reflects the existence of a well-defined process to access individual records, which is often not available from government agencies. Addition of version control would round this out.

BrightQuery's focused analysis of NCSES's datasets has highlighted their potential for transformation into AI-ready assets, enabling deeper insights and broader accessibility. This detailed scoring approach not only identifies areas for enhancement but also provides a roadmap for aligning NCSES's data ecosystem with advanced AI and ML frameworks.

Recommendations and Action Items

Key recommendations focus on enriching metadata with **Schema.org** and **Croissant tagging** to improve **searchability** and **machine understanding**, while aligning datasets with industry standards for **interoperability** and **accessibility**. BrightQuery's **Government Data Agent (GDA) Evaluator** and **Transformer** will automate these enhancements, reducing manual efforts and enabling the NCSES to efficiently meet the needs of generative AI systems. Priorities include expanding historical data availability by incorporating scanned archival reports with essential metadata and creating or exposing version control and revision history. Providing a structured, unified, standardized, and complete **API** is also recommended. Across the board standardization of terminology and navigation – survey names, variable names, and file listings would greatly enhance usability and discoverability. These targeted actions will position NCSES's data as a benchmark for AI-ready, interoperable government datasets. An effort to make data more interoperable – by releasing more granular aggregations, aligning survey response timing, and reviewing year-over-year methodological changes – is suggested, though we aren't able to determine the scope and outcome of such an investigation.

Broader Impacts

The NCSES's commitment to AI readiness can serve as a model for other federal agencies, demonstrating the feasibility and value of transitioning datasets into machine-understandable formats. This work aligns with the OPEN Government Data Act and the Department of Commerce's AI and Open Government Data Assets Working Group, showcasing NCSES as a leader in leveraging federal data to drive innovation and evidence-based decision-making.

The Federal statistical system comprises over 125 agencies or units that engage in statistical activities. A substantial portion of official statistics are produced by the 13 agencies that have statistical work as their principal mission. BrightQuery is building a comprehensive framework to support these agencies in transitioning their datasets to AI-ready formats, ensuring that statistical products are enriched with metadata, interoperable, and optimized for machine learning applications. This initiative will enhance the accessibility, accuracy, and utility of federal statistics for policymakers, researchers, and the public.

—Bureau of Economic Analysis (Department of Commerce).
—Bureau of Justice Statistics (Department of Justice).
—Bureau of Labor Statistics (Department of Labor).
—Bureau of Transportation Statistics (Department of Transportation).
—Census Bureau (Department of Commerce).
—Economic Research Service (Department of Agriculture).
—Energy Information Administration (Department of Energy).

—National Agricultural Statistics Service (Department of Agriculture).
—National Center for Education Statistics (Department of Education).
—National Center for Health Statistics (Department of Health and Human Services).
—National Center for Science and Engineering Statistics (National Science Foundation).
—Office of Research, Evaluation, and Statistics (Social Security Administration).
—Statistics of Income Division (Department of the Treasury).
—Microeconomic Surveys Unit, (Board of Directors of the Federal Reserve System).
—Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration (Department of Health and Human Services); and
—National Animal Health Monitoring System, Animal and Plant Health Inspection Service (Department of Agriculture).

Appeal

To realize this vision, BrightQuery recommends that NCSES should implement the outlined recommendations, pilot the GDA tools, and engage with BrightQuery in continued collaboration. Expanding the scope of this transformation to include replicability testing with other federal agencies will further validate and refine the methodologies and tools developed in this project. Establishing a roadmap for long-term AI readiness will ensure sustained progress and position NCSES at the forefront of AI-enhanced data dissemination.

Looking Ahead

BrightQuery is ready to support the NCSES to enhance high-priority datasets and build scalable solutions to ensure data integrity, accessibility, and utility. BrightQuery believes that together, we can transform NCSES 's data ecosystem into a beacon of innovation and accessibility, setting a new standard for public data in the age of artificial intelligence.

Appendix A: Triplification

Triplification refers to the process of converting structured or semi-structured data into a format suitable for the **Semantic Web** or other AI-driven systems work with **knowledge graphs**. This typically involves organizing data into **triples**, which are the foundational units of a **Resource Description Framework (RDF)**. A triple consists of three components and are structured as **"subject-predicate-object"** statements. For example, “John is 30 years old” would triplify to:

1. **Subject**: The entity being described (e.g., "John").
2. **Predicate**: The relationship or property of the subject (e.g., "hasAge").
3. **Object**: The value or another entity related to the subject (e.g., "30").

Learn More: EAV (entity-attribute-value) data models

Why Triplification for AI?

AI systems, especially those leveraging **natural language understanding (NLU)** or **knowledge graphs**, benefit from structured, interconnected data because it enables:

- **Semantic understanding**: Triplification enriches data with explicit meaning by defining entities and their relationships.
- **Interoperability**: Standardized triples can integrate seamlessly with other datasets and systems.
- **Machine reasoning**: AI can infer new facts or relationships using rules and ontologies applied to triples.

Steps in Triplification

1. **Data Extraction**: Identify structured, semi-structured, or unstructured data sources (e.g., relational databases, spreadsheets, XML, JSON, etc.).
2. **Entity Mapping**: Map data fields to entities, predicates, and values based on ontology or schema.
3. **Triple Creation**: Transform data into the "subject-predicate-object" format.
4. **Serialization**: Store triples in RDF or compatible formats (e.g., Turtle, JSON-LD, N-Triples).
5. **Integration**: Merge triples into a knowledge graph or database for AI consumption.

Benefits of Triplification for AI

- Facilitates **context-aware AI systems**.
- Supports advanced **querying and reasoning**.
- Makes data reusable across domains and applications.
- Enhances AI's ability to generalize knowledge and make decisions.

By adopting triplification, you prepare data in a way that aligns with how AI systems understand and process information, creating a foundation for smarter, more connected applications.

Example 1a: Naïve Triplification of a single NCSES Data Table

Table 1-1. College graduates, by level of highest degree, minor field of highest degree, and labor force status: 2023

(Number)

Level and field of highest degree	Total	Employed				Unemployed[a]	Not in labor force[b]
		Total	S&E occupations	S&E-related occupations	Non-S&E occupations		
All degrees	71,697,000	56,061,000	8,711,000	11,253,000	36,097,000	1,771,000	13,865,000
S&E fields	22,109,000	17,566,000	6,518,000	2,955,000	8,093,000	602,000	3,941,000
Biological, agricultural, and environmental life sciences	3,537,000	2,721,000	755,000	760,000	1,206,000	88,000	728,000
Agricultural and food sciences	471,000	370,000	53,000	59,000	259,000	15,000	86,000
Biological sciences	2,633,000	2,012,000	625,000	651,000	736,000	65,000	556,000
Environmental life sciences	432,000	338,000	77,000	50,000	211,000	8,000	86,000
Computer and mathematical sciences	4,361,000	3,563,000	2,110,000	524,000	928,000	156,000	642,000
Computer and information sciences	3,332,000	2,781,000	1,811,000	407,000	562,000	122,000	430,000
Mathematics and statistics	1,029,000	782,000	299,000	117,000	366,000	34,000	212,000

Triplification Output:

- AllDegrees hasTotal 71,697,000
- AllDegrees hasEmployedTotal 56,061,000
- …
- S&EFields hasTotal 22,109,000
- …
- S&EFields hasNotInLaborForce 3,941,000
- …
- BiologicalSciences hasEmployedInS&EOccupation 625,000
- …
- BiologicalSciences isSubCategoryOf BiologicalAgriculturalAndEnvironmentalLifeSciences
- BiologicalAgriculturalAndEnvironmentalLifeSciences isSubCategoryOf S&EFields
- S&EFields isSubCategoryOf AllDegrees
- …

This representation enables AI to understand relationships and infer new knowledge, such as "The unemployment rate for Biological Sciences is 2.5%" or "Economics and Psychology are both Social Sciences"

Note that this simplified example doesn't (yet) enable us to represent data across multiple years!

Example 1b: Including Dates in Triplification

- StatInstance1 asOfYear 2023
- StatInstance1 isForTypeOfDegree AllDegrees
- StatInstance1 isCountOf Total
- StatInstance1 hasValue 71,697,000
- StatInstance2 asOfYear 2023
- StatInstance2 isForTypeOfDegree AllDegrees
- StatInstance2 isCountOf EmployedTotal
- StatInstance2 hasValue 56,061,000
- …
- StatInstance3 asOfYear 2021
- StatInstance3 isForTypeOfDegree AllDegrees
- StatInstance3 isCountOf Total
- StatInstance3 hasValue 68,593,000
- StatInstance4 asOfYear 2021
- StatInstance4 isForTypeOfDegree AllDegrees
- StatInstance4 isCountOf EmployedTotal
- StatInstance4 hasValue 51,764,000
- …

Note that, were the “tree” of degree categories to shift from year-to-year, that would introduce further complexity in the graph.

Some Tools for Triplification

Triplifying unstructured text data involves extracting meaningful entities, relationships, and facts from the text and transforming them into subject-predicate-object triples. This process requires tools and techniques from natural language processing (NLP), knowledge graph construction, and semantic web technologies. Here are some commonly used tools and approaches for this purpose:

1. Pre-trained NLP Frameworks

These frameworks extract entities and relationships from unstructured text and can be configured to output triples.

a. spaCy with extensions

- Description: A powerful Python NLP library for tokenization, named entity recognition (NER), part-of-speech tagging, and dependency parsing.
- How to use: Combine spaCy's NER and dependency parsing with custom rules or extensions (like spaCy-Transformers) to identify subject-predicate-object triples.
- Output: Use the parsed relationships to generate RDF triples.

b. Stanford CoreNLP

- Description: Java-based NLP toolkit offering NER, sentiment analysis, and relation extraction.
- How to use: Apply its OpenIE module to extract facts and relations, which can be directly converted to triples.
- Output Example: From "John works at Google," it might extract (John, worksAt, Google).

2. Open Information Extraction (OpenIE) Tools

These tools are specifically designed to extract structured relationships (facts) from unstructured text.

a. OpenIE

- Description: A Stanford NLP tool for extracting subject-predicate-object triples directly from sentences.
- Features:
    - Handles complex sentences.
    - Outputs triples in text or RDF format.
- Example: From "Microsoft was founded by Bill Gates in 1975," OpenIE extracts:
  (Microsoft, was founded by, Bill Gates)
  (Microsoft, was founded in, 1975)

b. OLLIE (Open Language Learning for Information Extraction)

- Description: Focuses on extracting triples from sentences, including ones with implicit relationships.

3. Semantic Annotation and RDF Conversion Tools

These tools annotate text with semantic tags and convert them into RDF triples.

a. Apache Jena

- Description: A framework for building and querying RDF data.
- How to use: Pair it with NLP tools to store extracted triples in a semantic graph.

b. DBpedia Spotlight

- Description: An open-source tool for annotating and linking entities in text to DBpedia.
- How to use:
    - Input unstructured text.
    - Spotlight links entities in the text to concepts in DBpedia.
    - Output RDF triples that describe the linked entities and their relationships.

c. FRED (Semantic Graph Extraction)

- Description: A tool that converts text into RDF/OWL graphs.
- Features:
    - Combines semantic role labeling, NER, and ontology linking.

    - Outputs triples in multiple RDF formats.
- Example: From "Barack Obama was born in Hawaii," FRED generates:(Barack_Obama, bornIn, Hawaii)

4. Machine Learning-based Relation Extraction Tools

These use pre-trained or custom models to extract relationships for triplification.

a. Hugging Face Transformers

- Description: Provides pre-trained language models like BERT, RoBERTa, and T5 for relation extraction and entity recognition.
- How to use:
    - Fine-tune a transformer for relation extraction tasks.
    - Pair it with a triple generator script to convert detected relationships into triples.

b. DeepDive

- Description: A framework for statistical relational learning which extracts structured information from text.
- Use Case: Ideal for large-scale information extraction pipelines.

5. Knowledge Graph Construction Platforms

These platforms integrate NLP pipelines and RDF triple generation.

a. Neo4j with NLP Plugins

- Description: A graph database with NLP integration for knowledge graph creation.
- How to use:
    - Extract entities and relationships with NLP libraries.
    - Store the results as triples in Neo4j's graph format.

b. Amazon Comprehend

- Description: AWS NLP service for entity recognition and relationship extraction.
- Output: Results can be transformed into triples for storage in RDF-compatible formats.

c. IBM Watson Knowledge Studio

- Description: A cloud-based tool for training custom models to extract entities and relationships.
- How to use: Train the system to detect domain-specific entities and relationships, then export as triples.

6. Ontology-driven Tools

a. Protégé with Plugins

- Description: Ontology editing and management tool which can annotate and structure text-based data.
- How to use:
    - Define an ontology.
    - Annotate text with the ontology terms to extract triples.

b. AlchemyAPI (discontinued but alternatives exist)

- Description: Previously provided entity and relation extraction services. Alternatives like Google Cloud NLP now offer similar capabilities.

7. Custom Pipelines

For domain-specific triplification, you might need to build a custom pipeline combining:

- Pre-processing: Tokenization, entity recognition (e.g., spaCy, NLTK).
- Relation extraction: Using tools like Stanford OpenIE or custom models.
- Triple generation: Scripts to transform extracted relationships into RDF format.

Example Pipeline Using OpenIE and Jena

Input Text:	"Alice joined Acme Corp in 2021 as a designer."
OpenIE Output:	(Alice, joined, Acme Corp) (Alice, in, 2021)

	(Alice, as, designer)
RDF Triple Generation:	:Alice :joined :AcmeCorp . :Alice :joinedIn "2021"^^xsd:date . :Alice :hasRole "designer" .

Meta Content Framework (MCF) Format in Data Commons

Triplification is a generic term and the above examples should be treated as pseudo-code; ultimately, any triplified data must be inserted into a knowledge graph before it can be used. For a specific example, reference the format used by Data Commons: https://github.com/datacommonsorg/data/blob/master/docs/README.md

```
# Places and their relationships with one another
Node: USCounty_06085
typeOf: dcs:County
dcid: "geoId/06085"
name: "Santa Clara County"
containedInPlace: l:USState_06

Node: USState_06
typeOf: schema:State
dcid: "geoId/06"
name: "California"
containedInPlace: dcid:country/USA

# Statistical variables
Node: dcid:Median_Age_Person_Female
typeOf: dcs:StatisticalVariable
populationType: schema:Person
measuredProperty: dcs:age
statType: dcs:medianValue
gender: dcs:Female

# Statistical variable observation
Node: Observation_Median_Age_Person_Female_SanAntonio_TX_2014
typeOf: dcs:StatVarObservation
variableMeasured: dcs:Median_Age_Person_Female
observationAbout: dcid:geoId/4865000
observationDate: "2014"
value: 34.4
unit: dcs:Year
ss dcs:CensusACS5yrSurvey
```

Appendix B: Modeling Time Series data with Revisions

NOTE: This appendix does not apply to most NCSES data sets. NCSES treats corrections as isolated incidents with one-off fixes, rather than a timeline of expected estimates and revisions as first-order objects. However, we saw fit to include this appendix, as the concept is crucial across many government data sets – there is nothing *inherent* in the NCSES remit or coverage that precludes estimates or revisions, it's just not *currently* done.

Appendix B explores the challenges and opportunities in triplifying time-series data, focusing on the unique complexities of incorporating temporal dimensions into AI-ready datasets. While there are numerous tools available for approximating triplification from unstructured text, integrating the time component presents distinct difficulties that require innovative approaches.

One promising strategy involves treating the **reporting time period** as a first-order entity in the data model. Time in this context is inherently two-dimensional: it encompasses the time at which the data was produced and the time to which the data refers. Furthermore, time-series datasets often include multiple data points—such as estimates, revisions, and finalized values—that pertain to the same reference period. These intricacies make reporting periods a unique challenge, often more complex than a simple date stamp.

In the financial data domain, time-series modeling incorporates these nuances to account for revisions and overlapping periods effectively. Given the limitations of current large language models (LLMs) in handling time-series data, testing this approach within an AI framework may yield insights and improvements, especially as LLMs currently struggle to manage these complexities. This methodology aligns with strategies used in financial analysis and could serve as a benchmark for improving AI systems' interaction with temporal data.

What do we aim to solve?

A relational database offers robust capabilities for managing and querying time-series data, enabling precise insights and historical context. Below are some of the key use cases that such a database could support:

1. **Identify Data Within a Specific Time Range**
   Easily retrieve data points that fall within a defined time range, facilitating focused analyses such as quarterly or yearly trends.
2. **Generate Line Graphs Across Time**
   Use structured time-series data to create line graphs that illustrate changes over time, providing clear visualizations of trends and patterns.
3. **Identify the Latest Revision for Each Time Period**
   Quickly pinpoint the most up-to-date data for any given time period (e.g., determining the final GDP figure for Q2 2020 after all revisions).
4. **Roll Back Charts to a Historical Data View**
   Recreate what the dataset looked like at a specific historical point in time (e.g., showing the GDP figure reported in July 2020 for Q2 2020 before subsequent revisions).
5. **Align Data Across Multiple Time Series**
   Compare and analyze related data points from different time series that overlap in time (e.g., correlating an unemployment rate of 4.2% with a GDP value of $29 trillion for the same period).
6. **Approximate Like-for-Like Comparisons Across Asynchronous Data**
   Adjust for misaligned reporting schedules to generate accurate comparisons (e.g., aligning federal government fiscal year-end debt with calendar year-end GDP, accounting for the three-month offset).

By structuring time-series data in a relational database, these use cases can be addressed with precision, providing analysts and decision-makers with accurate and actionable insights. This approach also facilitates consistent handling of revisions and temporal relationships, ensuring a comprehensive understanding of historical and current trends.

Example 3: Time Series Modeling

In a relational database, we'd make each of these its own table, of course, but it's a similar concept.

STAT1: hasName: "Unemployment Rate"
STAT2: hasName: "GDP"
STAT3: hasName: "Unemployment Rate, Men"
STAT3: hasParent: STAT1
STAT4: hasName: "Unemployment Rate, Women"
STAT4: hasParent: STAT1

LOC1: hasName: United States
LOC2: hasName: Florida
LOC2: isIn: LOC1

PERIOD1: hasYear: 2024
PERIOD1: hasMonth: November
PERIOD1: hasType: Calendar Month
PERIOD1: precedes: PERIOD2
PERIOD2: hasYear: 2024
PERIOD2: hasMonth: December
PERIOD2: hasType: Calendar Month
PERIOD2: follows: PERIOD1

STATINSTANCE1: isForPeriod: PERIOD1
STATINSTANCE1: isStat: STAT1
STATINSTANCE1: isForLocation: LOC1
STATINSTANCE1: hasValue: 4.1%

STATINSTANCE2: isForPeriod: PERIOD2
STATINSTANCE2: isStat: STAT1
STATINSTANCE2: isForLocation: LOC1
STATINSTANCE2: hasValue: 4.2%
(each period + stat + location has a value)

We could capture estimates/actuals/revisions with:

STATINSTANCE1: isForPeriod: PERIOD1
STATINSTANCE1: isStat: STAT1
STATINSTANCE1: isForLocation: LOC1
STATINSTANCE1: hasValue: 4.1%
STATINSTANCE1: releaseDate: 2024-Dec-06
STATINSTANCE1: revises: None
STATINSTANCE1: revisedBy: STATINSTANCE1a

STATINSTANCE1a: isForPeriod: PERIOD1
STATINSTANCE1a: isStat: STAT1
STATINSTANCE1a: isForLocation: LOC1
STATINSTANCE1a: hasValue: 4.0%
STATINSTANCE1a: releaseDate: 2024-Jan-06
STATINSTANCE1a: revises: STATINSTANCE1
STATINSTANCE1: revisedBy: None

More on Period Types

Here is a (mostly complete) list of period types:

- Calendar Year [periodic]
- Calendar Quarter [periodic]
- Calendar Month [periodic]
- Fiscal Year [periodic]
- Fiscal Quarter [periodic]
- LTM (last twelve months), also referred to as TTM (trailing twelve months) [periodic]
- NTM (next twelve months) [periodic]
- Point in time [*not* periodic]

Why?

Understanding and accurately modeling time concepts is essential for interpreting data effectively. Various temporal dimensions, such as calendar periods, fiscal periods, and point-in-time versus periodic metrics, influence how data is structured, reported, and analyzed. Below are key distinctions and their implications:

1. Calendar Periods

These are the most intuitive timeframes, corresponding to standard calendar dates like "December 2024" or "the year 2000." They are what most people think of when referencing time and are universally understood. Calendar periods serve as the baseline for many datasets, providing a common framework for comparing data.

2. Fiscal Periods

Fiscal periods are tailored to organizational needs, diverging from calendar years to suit financial reporting purposes.

- **Corporate Examples**: Retailers often end their fiscal year on January 31 of the following year to account for holiday sales and returns. For instance, a retailer avoiding a December 31 year-end mitigates arbitrary financial fluctuations caused by factors like snowstorms affecting returns in the Northeast.
- **Government Examples**: The U.S. Federal Government begins its fiscal year in October, allowing time for Congress and the President to finalize the budget. This timing ensures that reporting aligns with the fiscal decision-making process rather than the calendar year.

3. LTM (Last Twelve Months) and NTM (Next Twelve Months)

LTM and NTM metrics offer a dynamic approach to annualized reporting, smoothing seasonal variations by providing rolling or forward-looking annual estimates.

- **Example**: If you want to understand “how many jobs the U.S. is creating annually” in July, relying solely on data from January-December of the previous year could misrepresent trends. LTM captures July-to-June data for a more timely and representative analysis.

4. Point-in-Time Metrics

These metrics capture discrete values at specific moments, akin to balance sheet data in accounting.

- **Examples**:
  - "What is the total household debt in 2024?" depends on whether you’re asking about January 1 or December 31, as the answers will differ.
  - Stock prices, which fluctuate intra-second, exemplify the granularity possible with point-in-time data, though most government datasets are unlikely to report with such frequency.

5. Periodic Metrics

Periodic data aggregates values over a span of time, such as annual GDP, which represents the sum of goods and services produced during a year.

- **Examples**:
    - "What is U.S. GDP in 2024?" includes every day of the year, from January 1 to December 31, summed up.
    - Some questions don't semantically align with periodic data, such as "How many new jobs were created at 2pm on Friday?" These metrics inherently represent a broader timeframe, like monthly or quarterly data, even if interpreted as momentary.

Why These Distinctions Matter

Accurately distinguishing between these temporal dimensions helps avoid misinterpretations and ensures consistency in analysis. For example:

- A point-in-time metric like "temperature at 2pm" may refer to the exact value at 2:00, while a periodic metric like "rainfall at 2pm" implicitly spans a range (e.g., 1:45 to 2:00).
- Reporting GDP as an annual figure makes semantic sense but attributing it to a single moment within the year, such as 2pm on December 31, does not.

By modeling time appropriately, analysts can ensure that the data reflects its true context and meaning, enabling more accurate and insightful interpretations across various use cases.

Estimates and Revisions

Estimates and revisions share a common framework: both are modeled with a **reference period** (the time the data pertains to) and a **reporting timestamp** (when the estimate or revision is published). This dual-timestamp system captures the temporal relationship between when the data is analyzed or revised and the period it describes.

Example Timeline: US GDP for 2024

1. **Initial Estimates**:
    - *January 7, 2019*: An economist predicts, "US GDP will be $a by 2024."
    - *July 14, 2020*: A revised estimate states, "US GDP will be $b by 2024."
    - *December 20, 2023*: Another revision predicts, "US GDP will be $c by 2024."
2. **Near-Term Estimates**:
    - *January 7, 2024*: An economist estimates, "US GDP was $d in 2024," anticipating BEA's official compilation.
3. **Initial BEA Report**:
    - *January 25, 2025*: BEA releases its first official figure: "US GDP was $e in 2024."
4. **Subsequent Revisions**:
    - *February 28, 2025*: BEA revises its figure to: "US GDP was $f in 2024."
    - *March 28, 2025*: BEA further refines its estimate: "US GDP was $g in 2024."

The Difference Between Estimates and Revisions

- **Estimates**: Typically published **before** the publisher releases an official value. These are often created by third parties, such as economists or financial analysts, and may use incomplete or alternative data sources to project outcomes.
- **Revisions**: Published **after** an initial official value is released, often by the publisher itself (e.g., BEA). Revisions refine the original data based on new information or methodologies.

The Complexity of Definitions

Even this distinction can blur. For example, the BEA refers to its March 2025 revision of 2024 GDP as an "estimate" because, in their words, "the real value of GDP is unknowable for certain." This highlights the inherent uncertainty in economic data, where even finalized values are approximations informed by the best available evidence and assumptions.

By understanding this nuanced relationship between estimates and revisions, analysts can better interpret how data evolves over time and account for its inherent uncertainties in decision-making.

Why do we even care about that second time component?

For most natural language queries, the second time component—the reporting timestamp—doesn't matter. If someone asks, "What was the US GDP in 1980?" they generally expect the latest revision unless they specifically request the value as it was reported at a particular moment in the past.
However, this time component becomes critically important in scenarios like **back-testing strategies**, where the accuracy of historical data as it was originally reported is essential.

Example: Back Testing a Gold Investment Strategy

Imagine a strategy that dictates, "Buy gold if inflation exceeds **2.5%."**

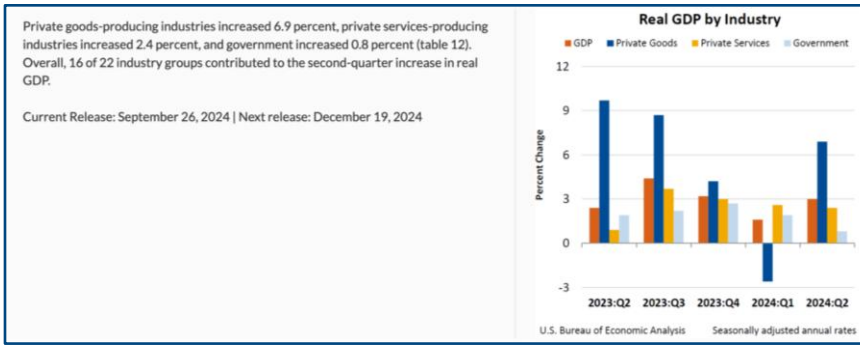
1. **Hypothetical Scenario**:
    - March: Inflation is reported as **2.6%** in the revised data. Following the strategy, you would buy gold in April.
    - April: Gold prices rise in May. The strategy appears successful.
2. **What Actually Happened**:
    - In April, the **original report** stated inflation was only **2.4%,** not **2.6%**. Based on this initial report, you wouldn't have bought gold in April.
    - The inflation figure wasn't revised to **2.6%** until September, months after you would have acted on the original report.
3. **Outcome**:
    - The back-test falsely assumes you would have acted on the revised inflation data **(2.6%)** rather than the initially reported figure **(2.4%)**.
    - As a result, you might conclude that your strategy was successful when, in reality, you made the wrong decision based on the information available at the time.

The Importance of Historical Context

This example illustrates why it's crucial to differentiate between **original reports** and **subsequent revisions**. Using only the latest data for back-testing can create a false sense of confidence in a strategy that wouldn't have worked in real-time. For scenarios requiring historical accuracy, such as financial modeling or policy evaluation, the second time component ensures decisions are evaluated against the data as it was known at the time, preserving fidelity to real-world conditions.

How?

The good news is that, at least for a human, making this distinction is pretty easy because it's one that the statisticians all tend to care to make. https://www.bea.gov/data/gdp/gdp-industry



It's very clear

- **releaseDate:** Sept 26, 2024. You can even get a timestamp if you click into the documents. 8:30am EDT
- Data in the chart
  - **isForPeriod:** 2023 Q2
  - **isForPeriod:** 2023 Q3
  - **isForPeriod:** 2023 Q4
  - **isForPeriod:** 2024 Q1
  - **isForPeriod:** 2024 Q2

Appendix C: Primer on Schema.org and Croissant Standards

Schema.org

Summary

Schema.org tags are a way of describing to a search engine what the contents of HTML are about, in a way that's more digestible.

Schema.org focuses on identifying entities (people, companies, places, creative works, etc.), describing properties (names, descriptions, published dates, etc) of those entities, relationships between them (authors, social graphs, etc.), and tying those to the HTML contents of web pages.

In their words:

"Your web pages have an underlying meaning that people understand when they read the web pages. But search engines have a limited understanding of what is being discussed on those pages. By adding additional tags to the HTML of your web pages—tags that say, "Hey search engine, this information describes this specific movie, or place, or person, or video"—you can help search engines and other applications better understand your content and display it in a useful, relevant way. Microdata is a set of tags, introduced with HTML5, that allows you to do this."

Full documentation: https://schema.org/docs/documents.html

Example 1

From the official documentation

```
<div itemscope itemtype ="https://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="https://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span> (born <span
itemprop="birthDate">August 16, 1954</span>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

(Schema.org metadata in bold)

The way to read this is:

- This block refers to a movie.
    - This heading is the movie name.
    - This block refers to the movie's director, a person.
        - This span is his name.
        - This span is his birth date.
    - This span is the movie genre.

Example 2

Illustrative example of Schema.org metadata applied to NCSES content

```
<div itemscope itemtype =" https://schema.org/Dataset">
  <h1 itemprop="name"> National Survey of College Graduates (NSCG)</h1>
  <p itemprop="description">
The NSCG is a unique source for examining the relationship of degree field and
occupation in addition to other characteristics of college-educated individuals,
including work activities, salary, and demographic information.
  </p>
  <h2>Survey Administration</h2>
  <p itemprop="measurementMethod">
    This survey was conducted by the
    <span itemprop="author" itemscope itemtype="https://schema.org/Organization">
      <span itemprop="name">
        Census Bureau
      </span>
    </span>
    in partnership with the
    <span itemprop="author" itemscope itemtype="https://schema.org/Organization">
      <span itemprop="name">
        National Center for Science and Engineering Statistics
      </span>
    </span>
    Within the
    <span itemprop="author" itemscope itemtype="https://schema.org/Organization">
      <span itemprop="name">
        U.S. National Science Foundation
      </span>
    </span>
    .
  </p>
  <a href=" https://ncses.nsf.gov/surveys/national-survey-college-graduates/"
itemprop="url">NSCG Home</a>
</div>
```

Consortium

Schema.org counts Google, Microsoft, and Yahoo among its founders. It is widely used and supported.

Limitations

Schema.org is designed largely for discoverability, and some limited semantic understanding of basic facts about those things by search engines and AI tools (e.g., triplification).

It doesn't handle much complexity:

- Tabular data
- Time series
- Numerical relationships

For example, you'd be able to tag a company, name, address, employees … but the financial statement is just a document (a link to a PDF or HTML page) or an "additionalProperty".

It only covers object types and properties within a defined vocabulary, which is very limited once you get into any level of detail in any technical field (finance, medicine, law, etc.)

Vocabulary: https://schema.org/docs/full.html

It provides limited options for disambiguation - URL, name, your own identifiers - but there's not a clear and consistent way to ensure that "Jimmy Carter" and "James Earl Carter Jr." are recognized as the same person, and *not* the same person as "James Earl Carter" (his father), James Carter the singer, Jimmy Carter the singer, Jim Carter the English actor, or any number of other people who aren't blessed with Wikipedia pages.

Extensibility

Schema.org *is* very extensible. Essentially, anyone can make up their own entity types, properties, relationships and it's "valid". That doesn't mean that search engines will interpret it correctly (or at all), of course, but there *is* a process for having things added to the standard, and Schema.org extensions may be adopted by narrower groups as their own standards.

Croissant is one such extension.

Croissant

Summary

The Croissant spec itself is fairly small and digestible. Therefore, this document should be read as an executive summary and our own takeaways.

It can be read about in full detail here: https://docs.mlcommons.org/croissant/docs/croissant-spec.html

While the Schema.org standard focuses on entities, properties, and relationships in structured and semi-structured data (HTML), Croissant focuses on files, tabular data, and images.

Croissant is specified in JSON and is an *extension of* Schema.org metadata. (you will see, for example, sc:name and sc: integer nested inside of cr: JSON nodes)

It provides for **discoverability** and **interpretability** of data sets (rather than web pages) - ingestion and discovery code by the consortium members.

Example 1

From the official documentation

The ratings RecordSet below corresponds to a CSV table, declared as a ratings table FileObject. Each field specifies as a source the corresponding column of the CSV file.

```
{
  "@type": "cr:FileObject",
```

```
  "@id": "ratings-table",
  "contentUrl": "ratings.csv",
  "containedIn": { "@id": "ml-25m.zip" },
  "encodingFormat": "text/csv"
},
```

The ratings RecordSet below defines the fields user_id, movie_id, rating and timestamp. The movie_id field is a reference to another RecordSet, movies.

```
{
  "@type": "cr:RecordSet",
  "@id": "ratings",
  "key": [{ "@id": "ratings/user_id" }, { "@id": "ratings/movie_id" }],
  "field": [
    {
      "@type": "cr:Field",
      "@id": "ratings/user_id",
      "dataType": "sc:Integer",
      "source": {
        "fileObject": { "@id": "ratings-table" },
        "extract": {
          "column": "userId"
        }
      }
    },
    {
      "@type": "cr:Field",
      "@id": "ratings/movie_id",
      "dataType": "sc:Integer",
      "source": {
        "fileObject": { "@id": "ratings-table" },
        "extract": {
          "column": "movieId"
        }
      },
      "references": {
        "@idfield": "movies/movie_id"
      }
    },
    {
      "@type": "cr:Field",
      "@id": "ratings/rating",
      "description": "The score of the rating on a five-star scale.",
      "dataType": "sc:Float",
      "source": {
        "fileObject": { "@id": "ratings-table" },
        "extract": {
          "column": "rating"
        }
      }
    },
    {
      "@type": "cr:Field",
      "@id": "ratings/timestamp",
```

```
      "dataType": "sc:Integer",
      "source": {
        "fileObject": { "@id": "ratings-table" },
        "extract": {
          "column": "timestamp"
        }
      }
    }
  ]
}
```

Example 2

Illustrative example of Croissant metadata applied to NCSES content

The ratings RecordSet below corresponds to an Excel file, declared as a FileObject. Each field specifies as a source the corresponding column of the Excel file.

```
{
  "@type": "cr:FileObject",
  "@id": "nsf25322-tab001-001-excel",
  "contentUrl": "nsf25322-tab001-001.xlsx",
  "encodingFormat": "application/vnd.openxmlformats-
officedocument.spreadsheetml.sheet"
},
```

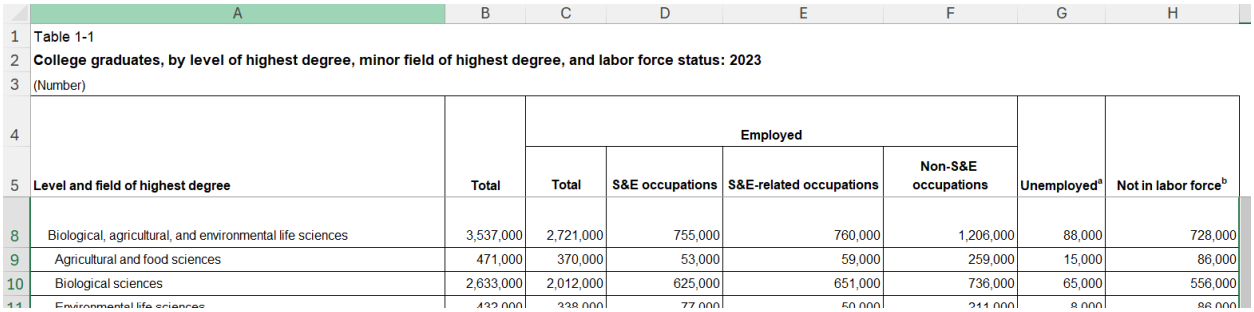
The RecordSet below defines the fields “Level and field of highest degree”, “Total” , “Employed – Total” , “Employed – S&E Occupations” , “Employed – S&E-related Occupations” , “Employed – Non-S&E Occupations”, “Unemployed”, “Not in Labor Force”.

```
{
  "@type": "cr:RecordSet",
  "@id": "nsf25322-tab001-001",
  "name": "Table 1-1",
  "description": "College graduates, by level of highest degree, minor field of
highest degree, and labor force status: 2023",
  "cr:path": "Table 1-1",
  "field": [
    {
      "@type": "cr:Field",
      "@id": " nsf25322-tab001-001/Level and field of highest degree",
      "dataType": "sc:Text",
      "source": {
        "fileObject": { "@id": "nsf25322-tab001-001-excel" },
        "extract": {
          "column": "Level and field of highest degree"
        }
      }
    },
    {
      "@type": "cr:Field",
      "@id": " nsf25322-tab001-001/Total",
      "dataType": "sc:Integer",
      "source": {
```

```
        "fileObject": { "@id": "nsf25322-tab001-001-excel" },
        "extract": {
          "column": "Total"
        }
      }
    },
    {
      "@type": "cr:Field",
      "@id": " nsf25322-tab001-001/Employed - Total",
      "dataType": "sc:Integer",
      "source": {
        "fileObject": { "@id": "nsf25322-tab001-001-excel" },
        "extract": {
          "column": "Employed - Total"
        }
      }
    },
    ...
  ],
}
```

Note that the Croissant standard doesn't currently contain a method to handle "skipping" rows, multiple header rows, non-data rows (i.e. footnotes), non-string column specifications (i.e. specifying column B and C rather than "Total" and "Employed – Total"), or duplicated strings.

As such, Croissant alone is not able to direct AI/ML tools to properly ingest NCSES data! Manual intervention and some custom ETL is needed.



	A	B	C	D	E	F	G	H
1	Table 1-1							
2	**College graduates, by level of highest degree, minor field of highest degree, and labor force status: 2023**							
3	(Number)							
4			Employed					
5	**Level and field of highest degree**	**Total**	**Total**	**S&E occupations**	**S&E-related occupations**	**Non-S&E occupations**	**Unemployed[a]**	**Not in labor force[b]**
8	Biological, agricultural, and environmental life sciences	3,537,000	2,721,000	755,000	760,000	1,206,000	88,000	728,000
9	Agricultural and food sciences	471,000	370,000	53,000	59,000	259,000	15,000	86,000
10	Biological sciences	2,633,000	2,012,000	625,000	651,000	736,000	65,000	556,000
11	Environmental life sciences	433,000	338,000	77,000	50,000	211,000	8,000	86,000

Consortium

Croissant is published by MLCommons (a consortium including Google, Meta, Microsoft, among others https://mlcommons.org/our-members/). Discovery and ingestion code belonging to the consortium's members is likely to be designed to look for Croissant metadata, but it's an open standard - anyone can use it.

What this means, simply, is that anyone who has "adopted the standard" is much more likely to be able to recognize a particular CSV as a data set, understand the contents (fields, data types) correctly (and without manual intervention), and the description, authors, and other metadata. Without Croissant metadata, a web crawler is much more likely to see a CSV file as just a CSV file, along with whatever context it may or may not be able to glean from the HTML surrounding the link.

Goals

- Discoverability
- Portability

- Reliability
- Responsible AI

Discoverability

Crawlers will look for the Croissant metadata to determine that a CSV is a data set suitable for machine learning, rather than just trawling for CSVs (in HTML links, zip archives, etc.) and guessing as to the contents.

Portability / Reliability

Most of this comes down to adding metadata.

- General metadata
    - Name of data set
    - Description
    - Authors
    - License information
    - Revision history
    - "File sets" - logical groupings of files
    - Etc.

- Field labels
    - Name
    - Description
    - Data type (integer, float, boolean, text, etc.)
    - Reference data (enums; i.e. if a column is "multiple choice" what are the choices and what do they mean)
    - Examples
- Images
    - A description of the image
    - Bounding boxes (outlines drawn around key elements)
    - Image Hash

Responsible AI

The treatment of responsible AI in Croissant boils down to a series of rai: tags, mostly in text format, mostly for human consumption (or machine consumption for a human audience).

- How the data is collected (e.g., web scraping, interviews, surveys, etc.)
- Expected use cases.
- Data limitations (e.g., who was left out of a survey?)
- Bias (what sort of systemic miscalculations may be made if data is taken naively? e.g., we weren't able to survey anyone without a permanent home address)
- Sensitive Information (does it contain PII?)

Full details available here: https://docs.mlcommons.org/croissant/docs/croissant-rai-spec.html

Limitations

Understanding Croissant's limitations is largely a matter of understanding what it does and does not include. It does not provide conceptual understanding or meaning in the human sense. It's just the basics: "here's a file, description, its authors, license information, and what columns are in it". While it will help ML tools find data sets and save ML engineers some time in answering basic questions and ingesting the data, it doesn't do any of the work of interpreting it for them.

The RAI part is even more limited. It does not provide a rigorous definition of bias (for example), a way to detect it, or any sort of rubric to score or measure it. It's simply asking data set publishers to tag their data with "rai:dataBiases: <please write something here>". (the same goes for all of the other RAI tags)

Again, we'd consider this **necessary but not sufficient.** While it may be second-nature to a statistician that there are (for example) people that can't or won't respond to a phone survey, and that those people are meaningfully different from respondents in a meaningful non-random way, it may not be to a machine learning engineer or the systems they design. A paragraph of explanation is not a substitute for that expertise, nor will it (of course) fix their code to properly account for those facts.

Appendix D: Data Sets

Government Funding for Science and Engineering

Survey of Federal Funds for Research and Development.

The Survey of Federal Funds for R&D is an annual census completed by the federal agencies that conduct R&D programs and serves as the primary source of information about federal funding for R&D in the United States.

Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions.

The Federal Science and Engineering Support Survey is a congressionally mandated survey and the only source of comprehensive data on federal science and engineering (S&E) funding to individual academic and nonprofit institutions.

Survey of State Government Research and Development.

The Survey of State Government R&D provides comprehensive, uniform statistics regarding the extent of R&D activity performed and funded by departments and agencies in each of the nation's 50 states, the District of Columbia, and Puerto Rico.

Higher Education Research and Development

Higher Education Research and Development (HERD) Survey.

The HERD Survey is the primary source of information on R&D expenditures at U.S. colleges and universities that expended at least $150,000 in separately accounted for R&D in the fiscal year.

Survey of Science and Engineering Research Facilities.

The Survey of Science and Engineering Research Facilities is a congressionally mandated, biennial survey that collects data on the amount, construction, repair, renovation, and funding of research facilities at U.S. colleges and universities.

Research and Development

Annual Business Survey (ABS).

The ABS is the primary source of information on R&D among for-profit businesses operating in the United States with one to nine employees. The ABS also collects data on innovation, technology, intellectual property, and financing from U.S.-based companies of all sizes.

Business Enterprise Research and Development (BERD) Survey.

The BERD Survey is the primary source of information on R&D expenditures and R&D employees of for-profit, publicly or privately held, nonfarm businesses with 10 or more employees in the United States that performed or funded R&D domestically or abroad.

Federal Facilities Research and Development Survey.

The Federal Facilities Research and Development Survey collects information on R&D performed by facilities owned and operated by the federal government.

FFRDC Research and Development Survey.

The FFRDC R&D Survey is the primary source of information on separately accounted for R&D expenditures at federally funded research and development centers in the United States.

Nonprofit Research Activities (NPRA) Survey.

The NPRA Survey collects information on research and experimental development performed by 501(c) nonprofit organizations in the United States.

Science and Engineering Workforce

Early Career Doctorates Survey (ECDS).

The ECDS provides statistics on demographics, labor market experiences, and jobs held by early career doctorates working at academic institutions and FFRDCs and covers all doctoral degree fields regardless of degree origin.

National Survey of College Graduates (NSCG).

The NSCG is a biennial survey that provides data on the characteristics of the nation's college graduates, with a focus on those in the science and engineering workforce.

National Training, Education, and Workforce Survey (NTEWS).

The NTEWS provides data on the educational and training characteristics of the nation's workforce ages 16 through 75, with a focus on those in the skilled technical workforce.

Survey of Doctorate Recipients (SDR).

The SDR provides data on the characteristics of science, engineering, and health research doctorate degree holders from U.S. academic institutions who are under the age of 76.

Survey of Postdocs at Federally Funded Research and Development Centers.

The FFRDC Postdoc Survey collects aggregate information on the demographic characteristics and fields of research of postdoctoral researchers employed at a FFRDC as of 1 October of the survey year.

STEM Education

Survey of Earned Doctorates (SED).

The SED is an annual census of research doctorate recipients from U.S. academic institutions that collects information on educational history, demographic characteristics, and postgraduation plans.

Survey of Graduate Students and Post Doctorates in Science and Engineering (GSS).

The GSS is an annual census of all academic institutions in the United States and its territories granting research-based master's degrees or doctorates in science, engineering, or selected health fields as of the fall of the survey year.

END OF REPORT

Learn more here:

BrightQuery.com

BrightQuery.ai