

Project Highlights
AI-Ready Data Products to Facilitate Discovery and Use

Project No. ADC-AIRD-24-N22

Sponsored by the National Secure Data Service Demonstration Project

Reporting Period: 11/01/2024 - 4/01/2026
Project Agreement Holder Jose Plehn-Dujowich
Other Project Team Members
Dan Nguyen
Hari Pulijal
Sriram Bajrang Bulusu

BrightQuery, Inc.
200 Spectrum Center Drive, Suite 300
Irvine, CA 92618
(310) 448-3157
jplehn@brightquery.com

Original Final Report Submitted: 3/23/2026

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 above do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. This product has been reviewed for unauthorized disclosure of confidential information under DRN: NCSES-DRN26-068.

Project Background and Goals

The AI-Ready Data Products to Facilitate Discovery and Use (AIRD) project was initiated to address a critical challenge in federal government data ecosystems, namely transforming fragmented, heterogeneous public data assets into machine-understandable, AI-ready resources. The project is funded under the [National Secure Data Service Demonstration Project](#) and aligns with the broader vision of enabling federal agencies to leverage Generative AI technologies for improved data discovery, accessibility, and decision-making in a shared service environment.

This project addresses the following broad goals:

- **Data Discoverability:** Ensure structured and unstructured public datasets and data points are easily searchable, accessible, and appropriately categorized for AI systems to be found efficiently and utilized correctly.
- **Data & Retrieval Quality:** Establish standards for data accuracy, completeness, and consistency, ensuring AI models can retrieve reliable and high-quality information across domains.
- **Insight and Actionability:** Provide metadata and semantic structures to enable AI models and agents to derive actionable insights from data.

To achieve these goals, this project designed and implemented a scalable, standardized **framework and AI readiness prototype toolkit** to enable government public data and information to be machine-readable and semantically enriched, interoperable across agencies, and easily consumed by AI systems and applications.

The project ultimately contributes to establishing a foundation for AI-driven public sector innovation, allowing agencies to transition from static data publishing to intelligent, AI-enabled data ecosystems.

AI Readiness: Major Accomplishments and Project Deliverables

Standards and Framework

This project delivered an **AI-ready standards and a framework** informed by two activities:

1) an extensive landscape analysis across multiple federal agencies to identify common challenges and best practices; and, 2) a comprehensive AI readiness assessment across three federal statistical agencies, including the Bureau of Economic Analysis (BEA), National Center for Science and Engineering Statistics (NCSES), and the U.S. Census Bureau (Census). These three agencies were selected to represent the variability in size and type of public data offerings found across the federal data landscape.

In this landscape analysis and readiness assessment, an AI readiness scoring rubric and scorecard were developed to assess the AI-readiness of government statistical agency websites and assists agencies with evaluating what functional areas on their public websites and publicly available data assets require remediation for AI-readiness.

See example scorecard below.

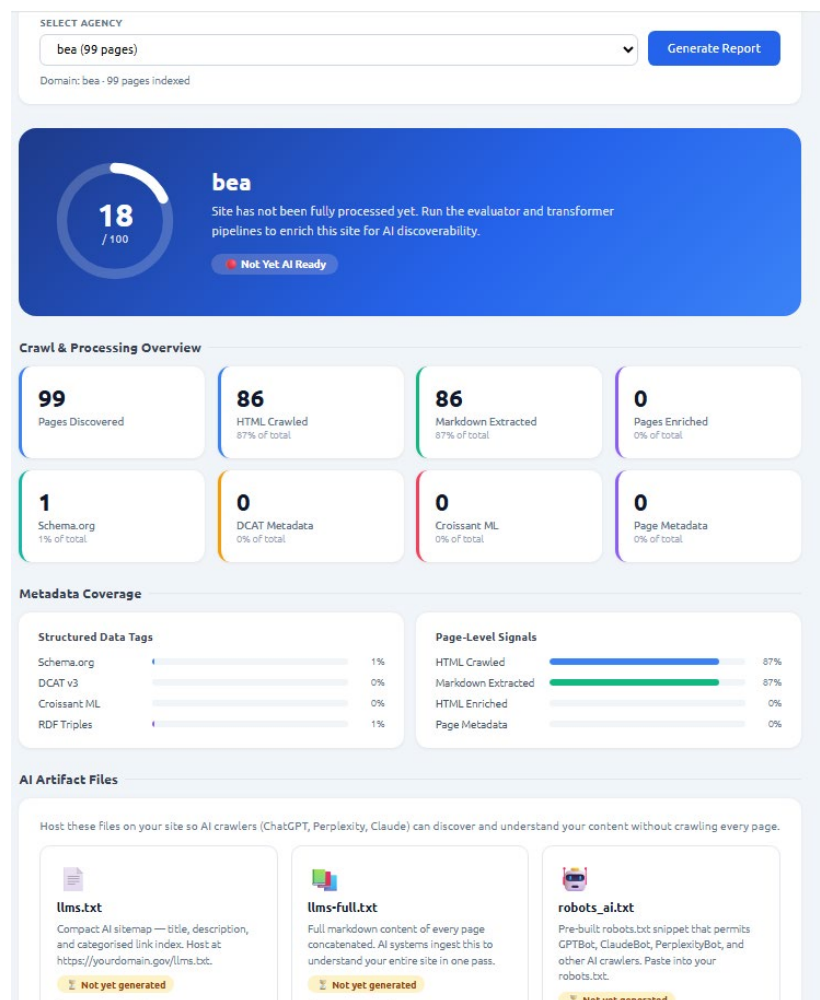
Table 1. BQ AI Readiness Scorecard					
✓	Complete / Good		✗	Missing / Not usable	
○ ● ●	Partial (slightly, somewhat, mostly)		?	Unable to determine	
NA	Not applicable				
	Functional Area				
	APIs	Reports & Unstructured Data	Tables	Interactive Charts	Website & Metadata
Scope					
Breadth	●	✓	✓	●	○
History	✓	●	✓	✓	○
Time Series Revisions Dimension	✗	✓ ≥ 1996	✓ ≥ 2002	✗	NA
AI / ML Specific Features					
Schema.org Tagging	✗	✗	✗	✗	✗
Croissant Tagging	✗	✗	✗	✗	✗
Triplification	✗	✗	✗	✗	✗
Training Data Labels	?	?	?	?	?
Bias Documentation	○	○	○	○	○
Data Structure & Format					
Digitization	✓	✓	✓	✓	●
Standardized Schema	✓	NA	●	NA	●
Standardized Layout	NA	●	NA	✓	●
Tabular Layout	NA	✓	✓	✓	✗
Field/Column Labels	●	✓	✓	✓	✗
Metadata & Context					
Units Discernable	✓	✓	✓	✓	NA
Data Dictionary	✓	●	●	●	NA
Methodology Described	●	✓	●	●	NA
Release Date/Time	○	✓	○	○	○
Interoperability & Granularity					
Geographical	●	●	●	●	NA
Industry/Sector	●	●	●	●	NA
Reporting Period	●	●	●	●	NA
Methodology	NA / ●	NA / ●	NA / ●	NA / ●	NA / ●
Technical Implementation					
Direct Access Links	✓	✓	✓	●	●
Documentation	✓	NA	NA	✓	NA
Rate Limiting	●	?	?	?	?
Authentication Methods	●	✓	✓	✓	✓
Governance & Compliance					
Version Control	NA	NA	NA	NA	✗ / ✓
Usage Rights	✓	✓	✓	✓	✓
Privacy Annotations	NA	NA	NA	NA	NA
Individual Linkages Possible	✗	✗ / ?	✗ / ?	✗	NA
PII Action Required	NA	NA	NA	NA	NA

To further test and refine the AI-ready standards and framework, a total of six federal statistical agencies, including BEA, NCSES, Census, Statistics of Income Division (SOI), National Center for Education Statistics (NCES), and the National Agriculture Statistics Service (NASS), were evaluated and received scores.

AI Readiness: Prototype toolkit

The project's AI-ready framework, assessments, and lessons learned were used to develop the *Government Data Evaluation and Transformer Tool (GDAT)*, an **AI readiness prototype toolkit**, that serves to transform public data products into machine-understandable, AI-ready data. The GDAT is a production-ready platform designed to be both self-hosted within an agency's IT environment, where agencies can choose to run 'as-is' or make modifications according to their needs, and is included as a shared service on the NSDS platform, where authentication, access, and authorization controls will be centrally implemented.

See image below which provides a high-level overview of the GDAT dashboard.



Core capabilities of the toolkit include Intelligent Asset Discovery & Classification, AI-Powered Metadata Generation, Flexible AI Deployment, and Database-Integrated Processing. The process also requires human-in-the-loop workflows for contextual accuracy. The GDAT and methodology were validated for scalability and portability across

five agencies, including BEA, NCSES, Census, Bureau of Labor Statistics (BLS), and the National Center for Health Statistics (NCHS).

In summary, the GDAT toolkit supports federal agencies' ability to:

- Define AI-ready schema and data and metadata standards
- Develop automated tools for evaluation and transformation
- Build a repeatable methodology for AI-readiness across federal agencies
- Deliver practical, deployable solutions for government adoption

Summary, Lessons Learned, and Next Steps

In support of the overall NSDS goals and objectives, this project underscored the importance of AI-driven decision-making, the continuous nature of AI-readiness development, and the foundational nature of metadata and governance. This project also identified several value propositions that support agencies' ability to move towards AI readiness for their public-facing websites. These include enhanced discoverability, interoperability, reduced manual effort, actionable insights, improved decision-making, digital infrastructure visibility, and accelerated research and innovation.

These prototype AI ready tools can be deployed to any agency with public facing data assets. Agencies should determine where these tools can fit into their data publication process. There remains a need to develop a standardized data transformation review process and implementation strategy across the federal environment. IT policy and governance efforts should also continue to leverage these AI ready tools where necessary to improve automation and efficiencies. For example, the Interagency Committee on Statistical Policy (ICSP) and other relevant federal councils can leverage these tools to provide a federal AI-readiness framework, benchmarks, and guidelines for adoption.

To build on the project's outcomes, future enhancements should include scaling, expanding and enhancing these tools capabilities, and methodologies to build a future-ready federal data ecosystem.

Technical Addendum | Solution Architecture

The AIRD solution follows a modular, scalable, AI-native architecture.

- **AI-Ready Standards Layer**
 - Schema standardization using Schema.org, Croissant and DCAT
 - Entity-Variable-Observation model
- **Evaluation Layer (GDAE)**
 - Multi-agent LLM-based system
 - Metadata detection and scoring
 - Vector-based semantic analysis
 - Technical Architecture
 - IBM Data Prep Kit-based standardized framework
 - Modular agent-based architecture with traceable reasoning
 - Hybrid AI stack (LLMs + knowledge graphs + retrieval systems)
 - RAG pipeline for grounded responses
 - Key Features
 - Flexible scoring across diverse agency architectures
 - Dynamic provenance and traceability
 - Security and governance for government deployment
 - Human-in-the-loop workflows
- **Transformation Layer (GDAT)**
 - Metadata generation
 - Schema mapping
 - Entity reconciliation
 - Knowledge graph construction
 - Technical Architecture
 - Modular agent-based architecture
 - Specialized agents for query, transformation, and response
 - RAG-based pipeline
 - Hybrid AI stack (LLMs + knowledge graphs + retrieval systems)
 - JSON-configurable workflows
 - Production Features
 - Docker-based containerized deployment
 - Secure, auditable operations
 - Source-level traceability

- Scalable performance for large datasets

- **Data Infrastructure Layer**

- Web crawling and ingestion
- Vector databases (pgvector)
- Hybrid search (semantic + keyword)

- **AI Interaction Layer**

- Natural language interfaces
- RAG and Text-to-SQL capabilities
- API-based and MCP-enabled access