FINAL REPORT
March 2026

Artificial Intelligence for Enhancing Data Quality, Standardization and Integration:

Project Activities, Findings, and Implications

Presented by:
NORC at the University of Chicago

Presented to:
National Center for Science and Engineering Statistics at the National Science Foundation and Bureau of Transportation Statistics at the U.S. Department of Transportation

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 above 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-DRN26-028).

Overview

The Artificial Intelligence for Enhancing Data Quality, Standardization, and Integration (AIDQSI) project prototyped a set of AI-assisted tools intended to demonstrate how artificial intelligence could strengthen data quality, standardization, and integration (DQSI) activities across the federal statistical system. This project informs the potential capacity of a National Secure Data Service to support the use of data in evidence-building across the federal statistical system.

Through an assessment of persistent challenges faced by federal statistical agencies when working with survey, administrative, private sector, and geospatial data the project identified and documented potential areas where DQSI needs could benefit from the introduction of AI. Federal statistical agencies increasingly rely on datasets drawn from diverse sources such as traditional surveys, administrative records, and nontraditional data such as GPS traces and web content, to build evidence and inform policy. While the Toolkit focuses on tabular and document-based inputs, many such inputs originate from non-traditional sources and enter statistical workflows via extracts, summaries, or derived indicators. Many of these data sources suffer from limited documentation, inconsistencies in formatting or schemas across jurisdictions, and a lack of machine readability for data and documentation, challenges that require increasing effort to address when these sources must be combined for evidence building or program monitoring. However, using AI to address these challenges requires careful consideration of quality, ethics, and responsible use of AI, with best practices requiring human review, transparency of approach, and careful consideration of where and how AI is applied to align with federal data quality expectations.

The AIDQSI Toolkit is designed to address DQSI needs with attention to these best practices and ethical requirements. The toolkit provides modular capabilities that assist with DQSI activities including: documenting data structure and content, organizing and labelling free-text data, harmonizing fields and values across files, imputing missing data transparently, and extracting structured information from documents.

This final report documents the design, construction, testing, and evaluation of the AI-DQSI Toolkit prototypes. It explains the rationale for tool selection, the use cases that guided development, the testing and refinement process, and best practices for responsible use. Together, the report and accompanying Toolkit are intended to inform future planning for the NSDS and to help agencies assess where AI-assisted approaches may complement existing AI-DQSI workflows.

The AI-DQSI project team included two federal partners and three external organizations. The National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation provided project oversight and leadership in conjunction with technical leadership from the Bureau of Transportation Statistics at the U.S. Department of Transportation. NORC at the University of Chicago served as prime contractor and led the framework plan, toolkit design, testing/validation, documentation, and other reporting requirements. Wolfram Research was the lead developer on the toolkit. The Discovery Partners Institute at the University of Illinois provided subject matter expertise to support the framework plan and toolkit design activities.

Activities

The AI-DQSI project was comprised of two primary workstreams. First, the team researched and created a Framework Plan report describing opportunities and challenges in using AI for DQSI tasks in the federal statistical system. Second, the team designed and developed a toolkit of resources using AI for DQSI purposes.

Framework Plan

Approach

We employed a comprehensive literature scan, targeted expert interviews, and a detailed review of available AI tools to assess the potential for AI to enhance DQSI for federal statistical data. The findings from these activities were synthesized to identify promising applications, potential challenges, and necessary safeguards for responsible AI implementation within a future NSDS.

Literature Scan. The main goal of the literature scan was to identify core DQSI challenges within the federal statistical system. The literature scan was conducted in two phases. In the first phase, we reviewed different sources categorizing data types of importance for federal statistical agencies. In this process, we identified four data types as particularly important for focus for the project: survey data, administrative data, private sector data, and geospatial data. In the second phase we identified and reviewed existing research, reports, and publications focusing on identifying existing best practices and challenges related to DQSI for each of the four data types. After our initial search, we followed additional citations from publication reference lists and followed up on citations recommended by interviewees. Insights from the literature scan informed both the design of the expert interview protocol and the subsequent review of AI tools.

Expert Interviews. We conducted fourteen semi-structured interviews with experts across nine organizations. This included eight interviews with federal agency staff, who provided insights into current data practices, challenges, and potential AI applications within their respective agencies. We also interviewed three data privacy and ethics specialists, who offered perspectives on the ethical considerations, privacy risks, and mitigation strategies related to AI implementation in data processing and integration. Finally, three subject matter experts (SMEs), provided specialized knowledge and insights into the DQSI challenges and opportunities associated with different data types of interest to the project.

Interviews lasted about 45 minutes each and followed a pre-defined interview guide, which permitted flexibility to explore themes that emerged from each conversation. Findings from the interviews are included throughout the Framework Plan report but are not attributed directly to interviewees or their agencies.

Review of AI Tools. We reviewed five existing AI tools relevant to DQSI to assess current capabilities, limitations, and potential concerns of existing tools for federal statistical agencies. Tool selection was informed by the literature scan and expert interviews, with prioritization based on the tools' relevance to DQSI challenges across the four data types. We also used AI and Large Language Models (LLMs) during the research process to identify additional use cases and recent developments. Each tool was systematically evaluated using a common set of criteria based on publicly available documentation, case studies, and user feedback to assess suitability for federal statistical applications.

Findings

The analysis of the literature scan, expert interviews, and a review of existing AI tools identified opportunities where AI may address unique DQSI challenges across the four key data types—survey data, administrative data, private sector data, and geospatial data—while also highlighting the ethical and privacy considerations that must guide responsible implementation. For survey data, AI could help by automating coding, cleaning, imputation, and the processing of open-ended responses using natural language processing. For administrative data, AI could assist with record linkage, automated processing, error detection, and anomaly identification. For private sector data, AI could support data integration, predictive modeling, bias mitigation, and pattern recognition to improve their statistical usefulness. Lastly, geospatial data can be enhanced with AI through image recognition, change detection, predictive mapping, spatial clustering, and error detection. Across all data types, the availability, completeness, and standardization of metadata emerged as foundational requirements for effective data reuse and AI-enabled workflows.

The use of AI within the federal statistical system raises legal and ethical considerations and requires responsible and trustworthy implementation. Strong privacy protections mandated by laws such as CIPSEA and Title 13 have important implications regarding how AI should be used with federal data, particularly given AI's ability to increase the risk of person or business re-identification when integrating data from multiple sources. Ethical considerations include the need to mitigate algorithmic bias, as AI systems may perpetuate or amplify existing disparities if differences in representation and coverage are not adequately addressed. High quality metadata are important to support reliable inputs and outputs for AI processes while transparency, explainability, and appropriate disclosure of methods are necessary to maintain trust. Human-in-the-loop workflows are emphasized as a key safeguard to mitigate hallucinations and inaccuracies, though they must be balanced to preserve the efficiency gains AI can offer.

A review of existing AI tools—including TurboCurator, OpenRefine, the RecordLinkage Toolkit, Google Earth Engine, and ArcGIS Pro with ArcGIS AI—demonstrates that a range of capabilities already exists to support DQSI, from metadata enhancement and data cleaning to probabilistic data linkage and large-scale geospatial analysis. These tools vary widely in suitability depending on the data type, use case, technical expertise required, and infrastructure needs, and they raise important privacy and ethical considerations related to bias, disclosure risk, security, and transparency.

The report recommends that AI can be used to enhance DQSI by automating data cleaning and validation tasks across all data types as well as integrating geospatial data into statistical applications. LLMs specifically can be used to enhance existing data documentation and metadata, thereby improving data discoverability and usability.

AI tools developed for use in a future NSDS should also consider the privacy and ethical recommendations outlined in the report. Algorithmic biases must be addressed with fairness audits, bias correction techniques, and training novel AI systems on curated, representative data, among other strategies. It is imperative that processes using AI ensure transparency of the procedures and consider explainability of the results. These priorities can be upheld by disclosing which records and variables have been used in statistical processes using AI, as well as the methods applied in data processing and integration. Finally, including humans in the loop when designing AI workflows can ensure higher accuracy and mitigate model hallucinations. AI tool design should facilitate collaboration between systems and humans.

Careful consideration of these findings is critical for the effective and responsible use of AI within the federal statistical context and specifically for AI tools developed for use in a future NSDS.

Toolkit

Overall Approach

The AI-DQSI Toolkit was developed to demonstrate how targeted, modular AI capabilities can support DQSI activities routinely performed by federal statistical agencies. The design reflects the needs identified in the project's Framework Plan, particularly the needs to: 1) improve metadata completeness; 2) address inconsistent schemas across data providers; 3) structure free text; 4) dynamically address missing data; and 5) extract structured information from documents that are not machine-readable. Across these needs, stakeholders emphasized the importance of reproducibility, transparency, and maintaining human judgement in all AI-assisted steps. The Framework Plan also highlighted a unique opportunity to leverage AI upstream of data analysis tasks, to support enhanced data exploration, understanding, and preparation. The toolkit's architecture directly responds to these priorities.

The toolkit includes tools to address each of five common, recurring tasks identified in the Framework Plan. This report's appendix includes selected screenshots of each tool for reference. The project's Documentation Report (Appendix C: Screenshots of Tools), includes detailed screenshots for each tool.

- **Metadata Extractor:** Creates and assesses metadata, which are critical when working with poorly documented survey, administrative, or vendor data.

- **Free-Text Encoder:** Structures free text, which is needed when analyzing open-ended survey responses, administrative logs, and narrative fields across program data.
- **Data Harmonizer:** Harmonizes schema, a major challenge when integrating data across providers with differences in naming conventions and formats.
- **Missing Data Assistant:** Assesses and remediates missing data, which is required to enhance data quality prior to analysis for decision-making.
- **Tabular Data Extractor:** Extracts key variables residing in PDFs and other semi-structured and non-machine-actionable documents.

Together, these tools reflect the project team's conclusion that DQSI challenges are best addressed through a suite of modular tools tailored to support analysts upstream of data analysis tasks.

Modular Design. Each tool performs a distinct function aligned to a stage of DQSI workflows. The Metadata Extractor provides a first pass inventory of variables, their types, and representative values. This enables rapid triage of datasets and supports downstream documentation and integration. For data with narrative fields, the Free-Text Encoder helps users transform unstructured or semi-structured text into structured variables using clustering or user-defined categorization. This allows users to use qualitative information in quantitative analysis and supports subsequent data harmonization tasks. The Data Harmonizer aligns schemas and normalizes values across datasets. Users can iteratively explore and refine harmonization steps for multi-source data integration until they arrive at a desired result. They can then export their workflow for auditing and repeated, future reuse. The Missing Data Assistant helps users assess missingness in their data and provides a range of simple to complex approaches to handle imputation. Imputed cells are flagged for transparency, and pre-/post-imputation diagnostics are provided after each run to help users assess data fitness-for-use. The Tabular Data Extractor retrieves structured entities and tables from semi-structured documents using retrieval-augmented generation (RAG). This supports scenarios where key information is not present in raw datasets.

This modular approach gives analysts flexibility to use one or more tools independently or in sequence to create pipelines, depending on their analytical needs. Each tool also produces artifacts – like metadata files, cluster summaries, schema mappings, imputation flags, and extracted tables – that can be passed across tools.

Prioritizing Transparency, Human-in-the-Loop Check, and Reproducibility. Across the toolkit, design decisions prioritize the values that emerged as priorities in the Framework Plan. In support of transparency, the tools generate interpretable artifacts including JSON schemas, cluster summaries, imputation profiles, and value normalization plans that help users understand the decisions made across tools. Intermediate steps are also exposed to the user for immediate review before transformations are applied. Users also provide step-wise oversight in human-in-the-loop workflows for each tool. LLM-assisted steps are exposed and always require human review and approval; they can also be augmented and overridden. Analysts have control over model usage, parameter settings, and final decisions on schema mappings, cluster label assignments, and imputation strategies.

To support reproducibility and transparency, the tools expose intermediate steps, configuration settings, and exportable artifacts for user review. The Data Harmonizer additionally stores full run artifacts in structured folders to enable reloading and reuse of workflows, while other tools rely on explicit exports and user-retained outputs. These artifacts enable analysts to recreate results, audit tool-supported decisions, and share transformation logic with collaborators.

This design aligns with findings from the Framework Plan and agency feedback, which emphasized the importance of responsible AI use. In particular, reproducible and auditable workflows, along with clear documentation, were established as priorities, and provided guiding principles for the toolkit design.

Summary of Toolkit Design and Construction Process. The development of the AI-DQSI Toolkit followed a phased process grounded in evidence gathered from the Framework Plan and informed through continuous, iterative testing and refinement. The timeline summarized in Table 1 reflects phases to identify federal DQSI needs, build and validate prototype tools, and develop accompanying documentation to support tool use. Each phase was aligned with priorities identified in the Framework Plan, from supporting cross cutting data preparation burdens to ensuring that the final tools could be deployed securely. The result is a modular toolkit designed to demonstrate how AI can be used responsibly to support realistic DQSI workflows.

Table 1. Summary of Toolkit Development Process

Phase	Timeframe	Activities	Outputs
Requirements Gathering	Framework Plan (2024 – 2025)	Literature scan; expert interviews; identification of DQSI pain points; evaluation of existing AI tools; prioritization of core functional areas	Framework Plan findings; selection of priority tool categories
Design & Prototyping	2025	Architecture definition; development of modular tool areas; creation of early prototypes; tool integration with LLMs where beneficial	Prototype versions of all five tools; preliminary documentation
Iterative Development & Testing	Late 2025 – Early 2026	Refinement through testing; dataset inventory development; definition of test cases; performance and usability assessments; debugging and UI fixes	Updated prototype tools; evaluation summaries; issue trackers
Documentation & Finalization	2026	Tool documentation; guidance on responsible AI use; performance considerations; best-practice checklists; security and privacy recommendations	Documentation Report; tool-specific best practices

Phase	Timeframe	Activities	Outputs
Final Integration & Delivery	2026	Assembly of the full toolkit; preparation of containerized builds; final revisions based on reviewer feedback; delivery to NCSES	AI-DQSI Toolkit (five tools); Final Report

Development of Use Cases

Process for Defining and Refining Use Cases. The team iteratively defined the use cases that guided the toolkit development through internal discussions, technical exploration, and consultation with agency partners. We began by synthesizing pain points identified in the Framework Plan including challenges specific to administrative, survey, private sector, and geospatial data and mapped these to analytic tasks routinely performed by federal statistical agencies.

We then identified and reviewed suitable, publicly available sample data to test with each of the five prototype tools that aligned with the use case. Publicly available datasets used to test and evaluate the toolkit are listed with direct links in the project's Documentation Report (Appendix B: Dataset Reference Table), which readers may consult for additional detail. Use cases were prioritized that were anchored in real-world data preparation burdens, were technically feasible, and were diverse enough to showcase the breadth of tools developed, from metadata profiling to schema harmonization to retrieval-augmented extraction. This approach ensured that the final set of use cases would demonstrate the toolkit's capabilities and provide practical illustrations for future NSDS workflows.

Summary of Selected Use Cases. The project implemented use cases that spanned multiple types of publicly-available data, jurisdictions, and file formats to evaluate each tool under realistic conditions.

- **Harmonizing data across jurisdictions**: Crime incident datasets from Chicago, IL, Philadelphia, PA, Las Vegas, NV, Washington D.C., Madison, WI, Cincinnati, OH, Raleigh, NC, and Scottsdale, AZ; and building permit data from Chicago, IL; Los Angeles, CA; and New York City, NY were used to test schema extraction, grouping, and value normalization within the Data Harmonizer. These data had heterogenous field names, code systems, date formats, and location fields, which made them ideal for testing semantic harmonization workflows.

- **Structuring free-text fields from survey and administrative data**: Narrative fields from an Austin Texas Cultural Centers Open Response Survey and Chicago, IL's FOIA Request Logs were used to demonstrate clustering and confidence-based assignments in the Free-Text Encoder. These fields contain heterogenous, unstructured free-text public comments and questions, which vary in length, level of detail, and formality, making them suitable for testing semantic grouping.

- **Extracting metadata from complex tabular sources:** Crime datasets (CSVs; Excel files), along with Federal Register notices containing tables and other semi-structured information were used to test the Metadata Extractor's ability to detect tables, infer variable types, interpret headers, and generate data dictionaries. In some cases, users also supplied supplemental documentation, such as TXT or JSON files, containing codebooks, variable descriptions, or schema notes to provide additional context for interpretation. These datasets varied in their layout complexity and had minimal contextual information, such as variable labels, available to support interpretation.

- **Imputing missing values in mixed-type datasets**: New York City MTA Bus Speeds and New York City Ferry Ridership transportation datasets, along with crime datasets, were used to evaluate univariate and multivariate imputation strategies, pre- and post-imputation statistics, and auditability features in the Missing Data Assistant. These real-world administrative datasets included categorical and numeric features of interest to test imputation strategies on missing values.

- **Extracting structured tables from public notices and records**: Legislative transcripts (text files, e.g., 103rd Congress, H.R. 1), Federal Register notices containing official information on Tribal Entities, and Amtrak state fact sheets (PDFs) were used to test the Tabular Data Extractor's retrieval-augmented generation pipeline, table extraction accuracy, and confidence reporting. These data provided a variety of scenarios requiring information extraction and summarization from single and multiple documents, such as information retrieval, summarization, and multi-year extraction and standardization tasks.

These use cases collectively demonstrate how the toolkit can support data preparation from initial documentation to text encoding, harmonization, imputation, and document-based extraction.

Excluded Use Cases. Several potential use cases emerged during toolkit design discussions, but were ultimately excluded due to scope, feasibility, or data access constraints.

- **Geospatial file harmonization**: Although federal agencies expressed interest in aligning shapefiles across jurisdictions or years, no publicly available datasets were identified to facilitate development of a tool for this use case.

- **Complex multi-table relational datasets**: Use cases involving longitudinal linked administrative records exceeded the prototype Data Harmonizer's planned capabilities, which focus on single-table, flat tabular inputs rather than relational schemas with foreign keys.

- **OCR-based extraction from scanned PDF documents**: While agencies expressed an interest in tools that could support the digitization and machine-actionability of static documents, such as PDFs scanned to images, this use case was deprioritized because of many existing technical solutions. Incorporating OCR workflows into the existing tools, such as to the Tabular Data Extractor, was deemed out of scope as it would require significant model development effort for marginal gains. This use case could be addressed by using other tools in tandem with the Tabular Data Extractor.

- **Large-scale, confidential data ingestion**: Given project constraints, testing on restricted data such as restricted-access microdata would require secure enclave execution was out of scope. The project focused instead on developing tests and use cases centered on publicly accessible data assets.

These exclusions reflect deliberate boundary-setting to ensure that the project delivered well-tested prototypes within available time and technical resources. They also highlight future opportunities for extension pending future-state NSDS priorities and capacity to expand.

Linking Use Cases to Toolkit Capabilities. Each use case was developed to illustrate the toolkit's modular capabilities and ability to support real-world analytic tasks. While each tool in the toolkit is modular and is designed to be used independently, several use cases span multiple tools. For example, the multi-jurisdiction crime data use case combines metadata extraction, schema harmonization, and missing-data treatment, reflecting the compound nature of real-world federal data preparation tasks. Table 2 illustrates how selected use cases map across the toolkit components.

Toolkit Construction

Tool Development, Technologies, and Data Sources Used. The AI-DQSI Toolkit was developed using a modular Python-based software stack designed for interactive data preparation workflows. Most tools rely on common Python data science libraries such as Pandas, NumPy, scikit-learn, and sentence-transformers, along with clustering algorithms (K-Means, HDBSCAN) and statistical methods for profiling and imputation. The system also uses vector search infrastructure such as FAISS for retrieval tasks and embedding pipelines for semantic similarity.

Several tools integrate a large language model stack accessed through AWS Bedrock, typically using Anthropic Claude models (e.g., Claude Sonnet or Haiku) for semantic tasks such as schema interpretation, cluster summarization, column name expansion, and document-based table extraction. Embeddings are generated using models such as Amazon Titan Text Embeddings or local sentence-transformer models. The toolkit is delivered as containerized web applications with a Streamlit interface, allowing analysts to upload datasets, configure parameters, and review results through an interactive browser-based UI that includes embedded documentation and progress indicators. During the testing phase, the tools were deployed in secure evaluation environments where agencies could either access a hosted web version or run containerized versions locally within their own infrastructure.

Testing and Evaluation Process. The toolkit was developed through an iterative testing and evaluation process conducted by the project team across NORC, Wolfram Research, and the University of Illinois Discovery Partners Institute. Testing was designed to reflect realistic federal statistical workflows and to expose tools to survey, administrative, and program data with realistic levels of variability, ambiguity, and scale. The team emphasized scenario-based testing in which analysts worked through complete use cases from initial data intake to final exports using real-world datasets drawn from public administrative sources, regulatory documents, and survey instruments.

Testing involved repeated cycles of hands-on use by data scientists, statisticians, and research methodologists with varying levels of familiarity with AI-assisted tools. This ensured feedback captured technical failures as well as usability, interpretability, and documentation gaps. Cross-organizational testing also allowed the team to evaluate how the tool performed across settings and environments.

Evaluation focused on the correctness and stability of the outputs, responsiveness and runtime for large or complex inputs, and the ease with which users could review, validate, and provide input on AI-generated suggestions. Findings from these evaluations were logged and informed subsequent development sprints through testing notes and an issue tracker shared between NORC and Wolfram.

User testing and internal evaluation informed a series of targeted improvements across the toolkit, strengthening technical completeness and tool usability. Key enhancements included:

- **Standardized export formats**: Outputs across tools were aligned to provide consistent export files and accompanying run artifacts. Additional run level artifacts like schema mapping files were exposed for download, strengthening transparency and reproducibility.
- **Improved progress indicators and feedback**: Progress bars, status messages, and clearer step level indicators were added to operations, helping users understand system state and reducing uncertainty during extended or multi-step processing.
- **Enhanced user guidance**: Additional in-tool guidance and diagnostics were added to help users understand tool outputs and options.
- **Multi-file and batch processing support**: Several tools were enhanced to improve the management of multiple input files within a single run, including clearer file-level summaries, consolidated exports, and improved handling of multiple file-type inputs.

These improvements reflect a shift from functional prototypes to analyst ready tools that emphasize clarity, auditability, and practical integration into DQSI workflows. The testing process confirmed that iterative, use-case driven evaluation was essential to aligning the toolkit with federal statistical requirements and expectations, as well as providing a foundation for future refinement and expansion.

Documentation Report

The documentation report for the AI-DQSI Toolkit outlines key considerations for responsible use and provides detailed guidance on the toolkit's intended uses, limitations, and operational context. The toolkit is designed primarily for analysts, data scientists, statisticians, and program analysts responsible for preparing datasets for research, evidence-building, and statistical integration. Documentation emphasizes that the tools assist with data preparation tasks such as metadata generation, schema harmonization, free-text structuring, missing-data imputation, and structured extraction from documents. At the same time, the report clearly identifies limitations. The tools are optimized for rectangular tabular datasets and text-based documents and are less suitable for highly complex spreadsheet structures, nested data formats, or image-based inputs. The documentation also stresses the importance of maintaining human oversight throughout the workflow. Analysts are expected to review AI-generated outputs, validate mappings and classifications, and treat automated results as exploratory until they have been verified by subject matter experts.

The report also documents decisions regarding evaluation, performance measurement, and cost monitoring across the toolkit. Each tool includes guidance on assessing output quality using task-appropriate metrics. Examples include field-mapping accuracy and schema coverage for data harmonization, clustering quality indicators such as Silhouette and Davies–Bouldin scores for text encoding, metadata completeness and type inference accuracy for metadata extraction, and statistical error measures for imputation quality. Computational performance considerations such as runtime, memory use, and scalability are documented, particularly for steps that rely on large language models. Example evaluation runs provide estimates of token usage, number of model calls, and approximate processing costs to help organizations anticipate resource requirements.

The documentation also includes sections on privacy, security, and ethical considerations. These guidelines address risks such as exposure of sensitive information in prompts, embeddings, intermediate artifacts, or extracted tables. Recommended safeguards include minimizing the amount of data provided to the tools, masking direct identifiers, restricting access to intermediate files and logs, and ensuring secure storage of outputs. The documentation also highlights risks related to algorithmic bias, potential data misuse, and adversarial attacks such as prompt injection or attempts to extract sensitive information. To mitigate these risks, the report recommends human review of outputs, subgroup performance checks to detect bias, strict governance controls in secure deployment environments, and clear documentation of model configurations and assumptions. These measures are intended to support responsible AI use and maintain transparency, accountability, and compliance with data protection requirements in federal statistical workflows.

Deployment

Project deployment planning focused on enabling agencies to adopt the toolkit within secure computing environments while maintaining flexibility in infrastructure choices. Each tool was packaged as a containerized application, typically using Docker, and delivered with a browser-based interface so that agencies could run the software locally or host it in a controlled environment. During the testing period, tools were deployed in evaluation environments where analysts could access them through web interfaces or install containerized versions within agency infrastructure. The design allows organizations to configure their own credentials for external AI services, such as AWS Bedrock, ensuring that agencies maintain control over model access and compliance with internal security policies. Deployment planning therefore emphasized compatibility with secure data environments, including scenarios where tools operate on protected datasets subject to federal privacy regulations. Throughout this process, close coordination was conducted with the NCSES, BTS and other project stakeholders to ensure that deployment considerations aligned with federal data governance requirements, security expectations, and the potential integration of these tools into future federal data service environments.

Long-term deployment also requires planning for ongoing maintenance and operational support. Because several tools depend on evolving AI services, embedding models, and Python-based machine learning libraries, maintaining compatibility will require periodic updates to model integrations, software dependencies, and security configurations. Maintenance considerations include monitoring API changes for LLM providers, updating libraries used for clustering or data processing, ensuring compatibility with container orchestration platforms, and managing the storage and lifecycle of intermediate artifacts generated during tool runs. If the toolkit were to be incorporated into a future National Secure Data Service (NSDS) environment, additional considerations would include integration with existing secure computing infrastructure, alignment with federal data governance requirements, and support for auditing and monitoring capabilities. In that context, deployment planning would also need to address scalability, user authentication, logging standards, and mechanisms for updating models or components without disrupting ongoing analytical workflows.

Lessons Learned and Considerations for a Potential, Future National Secure Data Service

Using AI for DQSI Tasks

Lessons learned. The Framework Plan findings highlighted the critical importance of data quality within the federal statistical system. Data must be objective, reliable, and accurate. They must also be coherent, reflecting consistent definitions, and interpretable, with usable and detailed metadata. When survey, administrative, and third-party data are used for purposes other than those for which they were originally designed, their fitness for those new analytical needs tends to decline.

These quality needs and challenges create opportunities for the use of AI. The processes of standardizing, documenting, and validating datasets can be manual and time intensive, suggesting opportunities for AI support to increase capacity, with potentially positive implications for the breadth and quality of the resulting data. LLMs in particular can be leveraged to perform data cleaning and documentation tasks that require basic interpretation and contextual knowledge, such as developing data dictionaries, identifying likely correspondences across files, or standardizing datasets that reflect similar underlying concepts. Prior to recent advances in AI, programmatic solutions in these areas were limited by the need for human reviewers to supply contextual understanding directly, constraining scalability.

At the same time, AI implementations in activities such as data assessment, processing, standardization, and linkage must meet a high bar for quality, accuracy, and trustworthiness. Our Framework Plan findings emphasize that transparency about AI-assisted steps and human review of AI performance are essential for this purpose. AI users in these areas must have some understanding of the methods that are applied, assumptions embedded in those methods, and the potential limitations on model performance. During toolkit development, however, we routinely encountered a tension between maximizing transparency and maintaining ease of use. Many of the same features that support auditability, such as exposed parameters, detailed logs, and quality metrics also increased the interface complexity and cognitive load for tool users.

One way we ultimately bridged this gap was in targeting the tools toward the initial exploratory steps in a task—making a first pass at a data dictionary or suggesting clusters within open-ended data, for example—where requirements for precision were lower and imperfect products still yielded efficiency gains, such as valuable time savings for users. Interactive validation and revision steps, such as those used in the Data Harmonizer, further support this approach by enabling users to review, adjust, and approve AI-generated suggestions, increasing user engagement, analytical rigor, and confidence in results.

Considerations for a potential NSDS. The project's findings suggest that the best opportunities to use AI in DQSI tasks are in areas that augment, rather than replace human effort and expertise. AI tools can support the iterative development of products by anticipating many of the edge cases that historically have made manual DQSI activities slow, inconsistent, and resource intensive. However, AI-developed outputs cannot be assured of reaching the necessary quality thresholds for use in the federal statistical system without additional human review and clear pathways from automated steps to human decision-support.

AI tools used for DQSI activities should include features that explicitly log AI-assisted actions, preserve intermediate artifacts, and support reproducibility and auditability. For example, the Data Harmonizer uses AI to optimize schema alignment and value normalization and then generates both JSON and SQL documentation of alignment decisions. These outputs make AI-assisted transformations visible, easier to review, and reusable across projects over time.

These observations emerged not only from the Framework Plan but also from Toolkit testing activities, where users repeatedly expressed the need for iteration, transparency, and human-in-the-loop decision-points; these characteristics gave users necessary confidence to build trust in AI-assisted results and to support defensible use of AI in DQSI contexts.

These considerations stand in tension with, and must be understood in the context of, rapid advances in AI capabilities. The increasing prevalence of agentic AI previews further movement away from human-in-the-loop AI approaches toward strategies that give systems increasing autonomy. Although agentic AI was still little discussed at the time of the Framework Plan, it was well-known to our testers, and they continued to express the need for human feedback and transparency for these critical DQSI operations. Agentic AI likely represents an opportunity to extend the range of tasks that AI can respond to beyond the kinds of use cases this project identified, but project findings suggest it will still be best used in similar ways, creating preliminary versions of products rather than completing entire end-to-end workflows.

Building Tools with LLMs

Lessons learned. Configuring and maintaining LLM connections requires more effort and can be more challenging for average users than a Python-based program alone would be. LLM-based tools also introduce additional considerations related to authentication, credential management, model selection, and ongoing maintenance. LLMs also come with computing and usage-based costs that potential users may not consider beyond the costs of the tools themselves. Users likely have varying access to LLMs or have access to different models. These concerns are particularly salient in use cases involving sensitive data sources. For data sources requiring secure environments, LLM-based tools require access to models set up within these environments rather than external calls.

Because each LLM performs differently, we were unable to design tools that generalized across models. Instead, we optimized performance and provided specific cost guidance on certain models, recognizing that these choices may not align with all users' institutional constraints or preferences. We also considered LLM ensembling, which uses multiple LLMs together to improve the performance of a tool; while this idea held potential for performance reasons, the complexities it raised for pricing, deployment, and optimization made it unworkable for this prototype toolkit.

The tools in the toolkit prioritize hybrid systems, combining LLMs with deterministic rules and processes. This hybrid design is partially a matter of efficiency. Both latency and cost scale with LLM usage; techniques like caching, chunking, and hybrid rule-based preprocessing were critical for managing larger data volumes within resource constraints. Beyond efficiency, however, deterministic rules also constrain the variability of LLM results, allowing for more consistency and clarity in tool results. The inherently non-deterministic nature of LLMs complicates reproducibility, debugging, and auditability for data pipelines that expect stable results.

Importantly, this approach fundamentally differs from the use of general-purpose AI tools (e.g., asking a chatbot such as ChatGPT to perform metadata extraction directly). In the AI-DQSI Toolkit, traditional statistical and machine-learning methods are used deliberately to preprocess, structure, and constrain data before any LLM involvement. LLMs are then applied only at specific points in the workflow where semantic interpretation adds value, using carefully designed prompts optimized for well-defined tasks. This design reduces ambiguity, increases consistency, and produces outputs that are more transparent and auditable than those generated through ad hoc prompting of a general AI tool. Rather than requiring individual users to experiment with prompts, the toolkit embeds vetted prompts, quality checks, and human-review steps as guardrails directly into the workflow. This combination of deterministic methods and LLMs is a primary value-add of the project and enables agencies to use AI-assisted tools that are more reliable, interpretable, and suitable for statistical workloads. Prompt design is also very context sensitive, making prompts harder to version, test, and validate than traditional code. As a result, LLMs were most effective when used selectively in tasks where the LLMs' ability to interpret and apply context provided enough value to outweigh these considerations. Providing LLMs with carefully scoped "micro-context" (e.g., column names, sample values, or data dictionaries) also helped improve LLM accuracy and consistency in classification and encoding tasks. Supplying the

LLM with small, targeted metadata snippets rather than full documentation helps to constrain context, reduce variability and improve reproducibility.

Considerations for a potential NSDS. Projects that use AI technologies require more early-stage planning for deployment with specific attention to questions of access, cost, and data security. Further complicating this situation, AI technologies are constantly changing, while government deployment cycles need to allow time for review, testing, and refinement. A potential NSDS should not discount the time and attention needed for these pieces.

A central achievement of this project is the demonstration of how standalone, task-optimized AI tools can be designed, vetted, and deployed to support specific DQSI workflows within the federal statistical system. Building tools focused on well-defined DQSI enabled more disciplined and efficient use of LLMs, not only reducing cost but also better managing the inherent uncertainty of non-deterministic technologies. These tools are designed with attention to auditability and transparency best practices, and they are optimized to particular purposes—instead of individual users designing and honing prompts, we have done that activity at scale. Finally, by concentrating on specific use cases, we help potential users to translate abstract AI capabilities into practical support for the day-to-day activities of producing high-quality data and statistics without compromising quality. This approach demonstrates how powerful AI technologies can be integrated into DQSI activities without compromising expectations for data quality, transparency, and human oversight.

Building Generalized DQSI Tools

Lessons learned. This project developed generalized, modular tools to support a wide range of DQSI activities across the federal statistical system. However, identifying public use datasets and use cases that had the same data quality and standardization challenges reflected in our Framework Plan interviews was a challenge. Many of the most salient challenges identified in interviews, such as those involving complex administrative or restricted data, could not be fully replicated using public data alone.

The final list of use cases we selected for the toolkit were directly shaped by which scenarios could be reasonably approximated using public data. Additionally, a lack of diverse public data samples also limited the project team's ability to test the tools across the full breadth of potential DQSI applications identified in the Framework Plan.

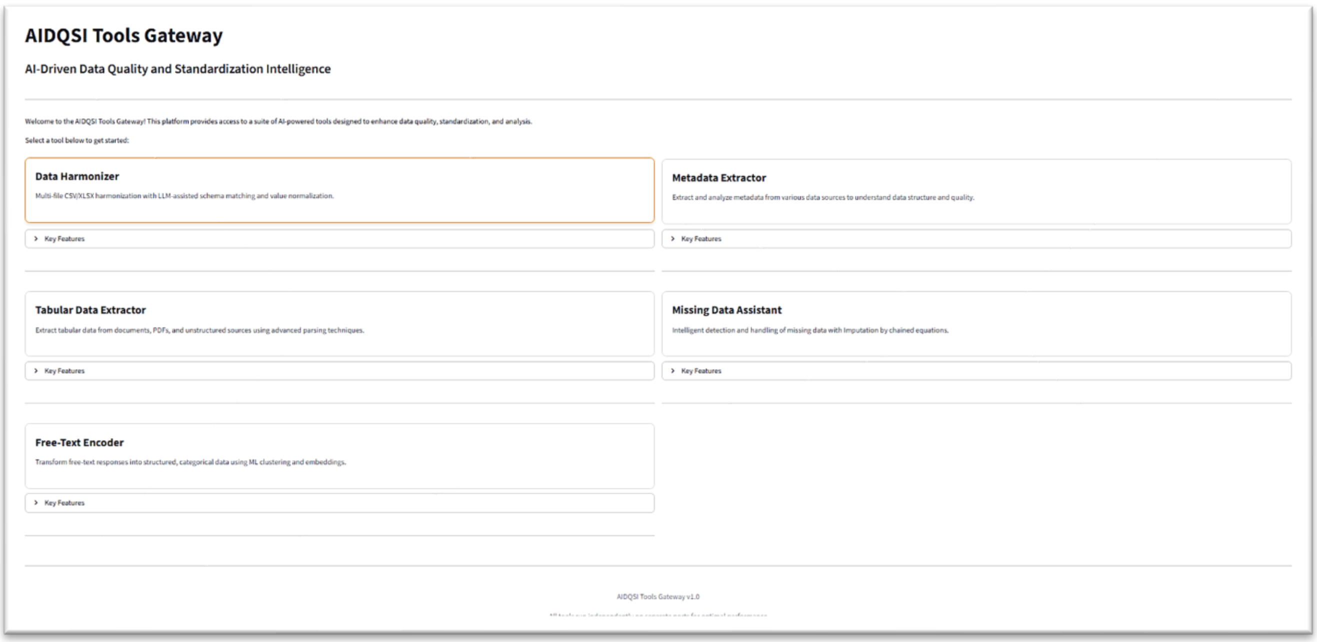
Considerations for a potential NSDS. Identifying common use cases in an area as broad as data quality, standardization, and integration requires navigating a balance between general needs, such as those identified in our Framework Plan, and the specific details needed for tool development. Developers need user personas with different capacities, data input file types, expected output format types, sample datasets, and lots of testers. We were able to create a sample of some of these detail from the data users and producers among the project partners, as well as drawing from the experiences of BTS, our technical partner. However, these inputs only capture a subset of the diversity of needs present across the federal statistical system.

For a potential future NSDS, generalized DQSI tools like these will need continued refinement and validation as additional, specific users and additional agencies attempt to apply them to new data types and use cases. Ongoing user feedback, coupled with exposure to a wider range of real-world datasets (e.g., restricted or sensitive sources within secure environments) will ensure that generalized tools remain flexible, trustworthy, and fit for purpose at scale.

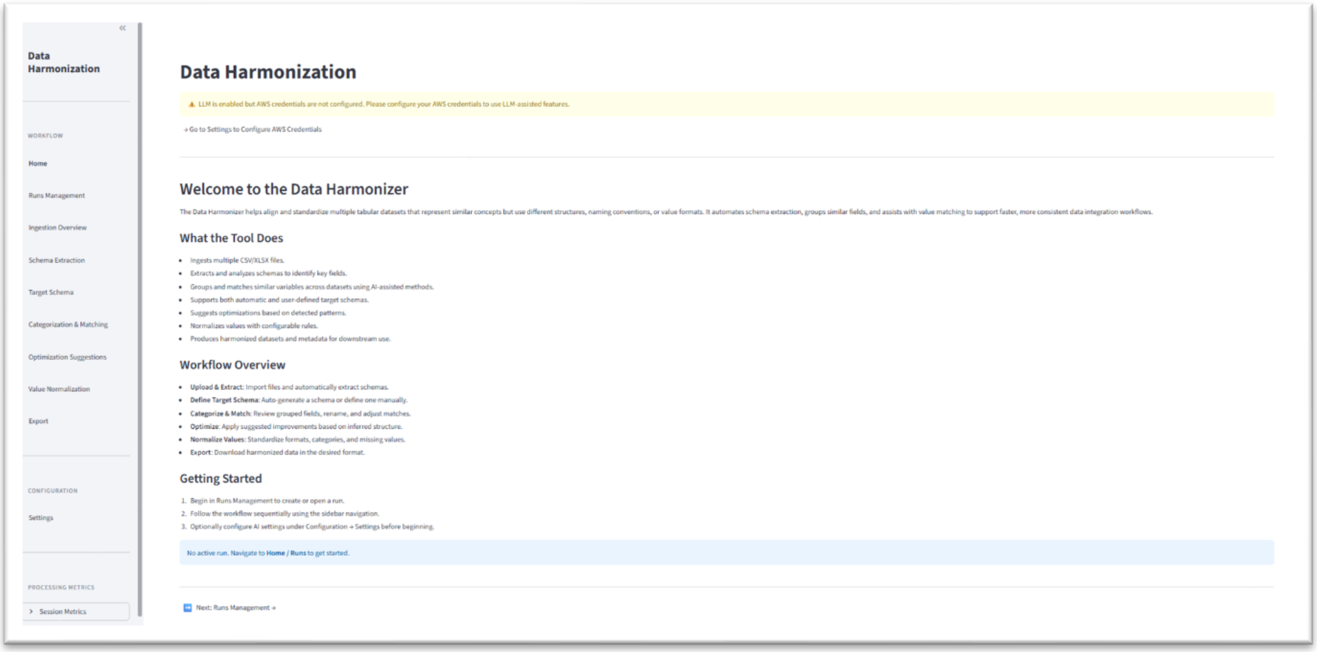
Appendix: Selected Screenshots of the AI-DQSI Toolkit

This appendix provides representative screenshots of the AI-DQSI Toolkit to illustrate the interfaces analysts encounter when using each tool. These images are intended to convey the general structure and interaction style of the prototype tools rather than to document full workflows. Full tool specific documentation is provided in the toolkit technical documentation.

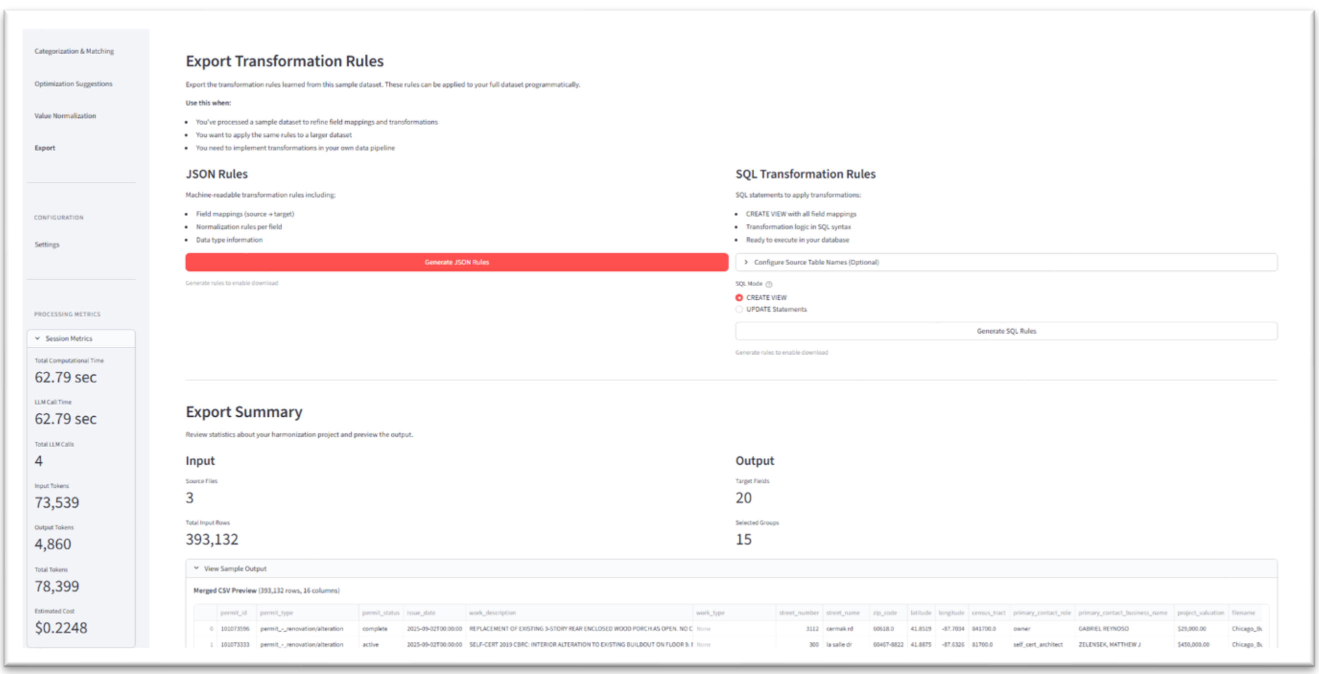
AI-DQSI Tools Gateway: Home Page



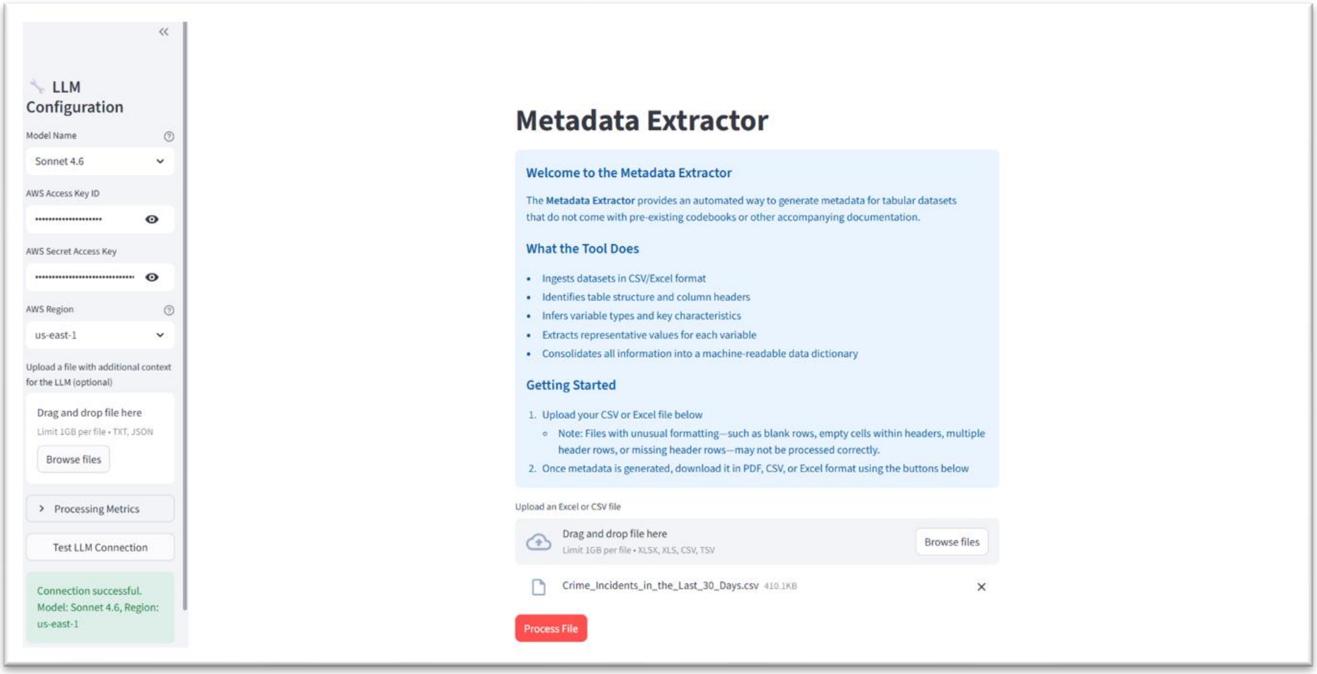
Data Harmonizer: Home Page



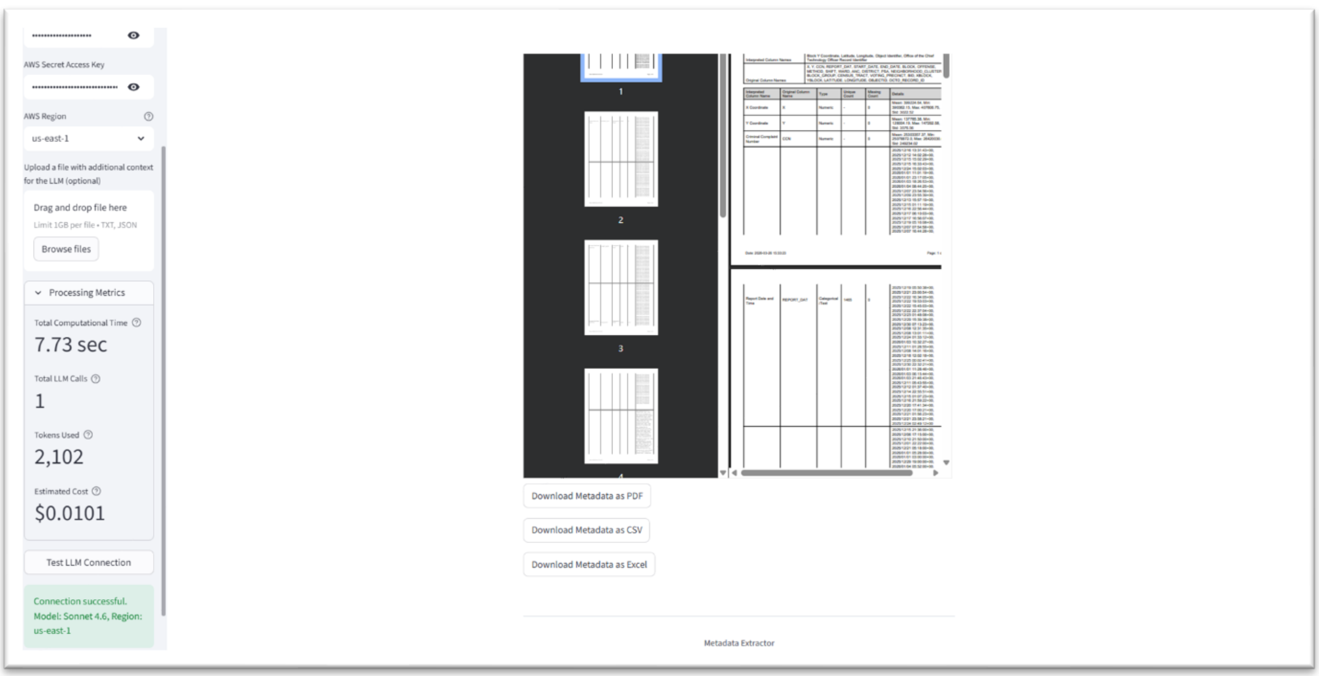
Data Harmonizer: Export & Session Metrics



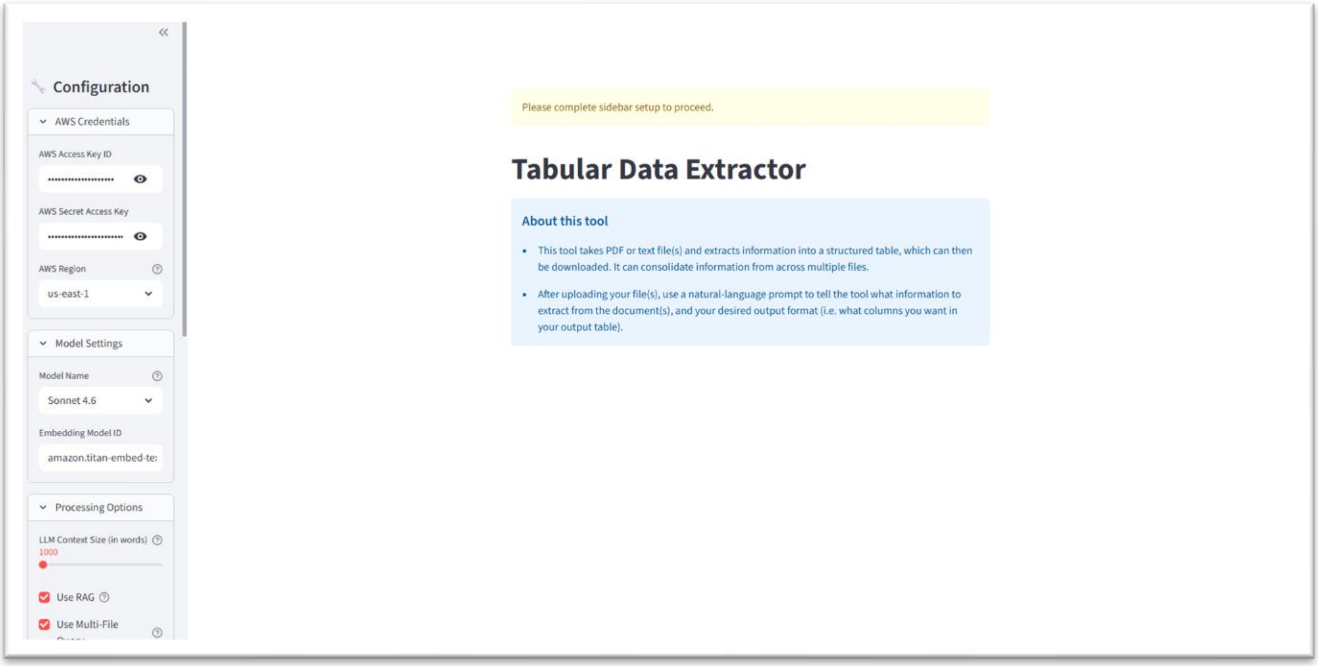
Metadata Extractor: Home Page



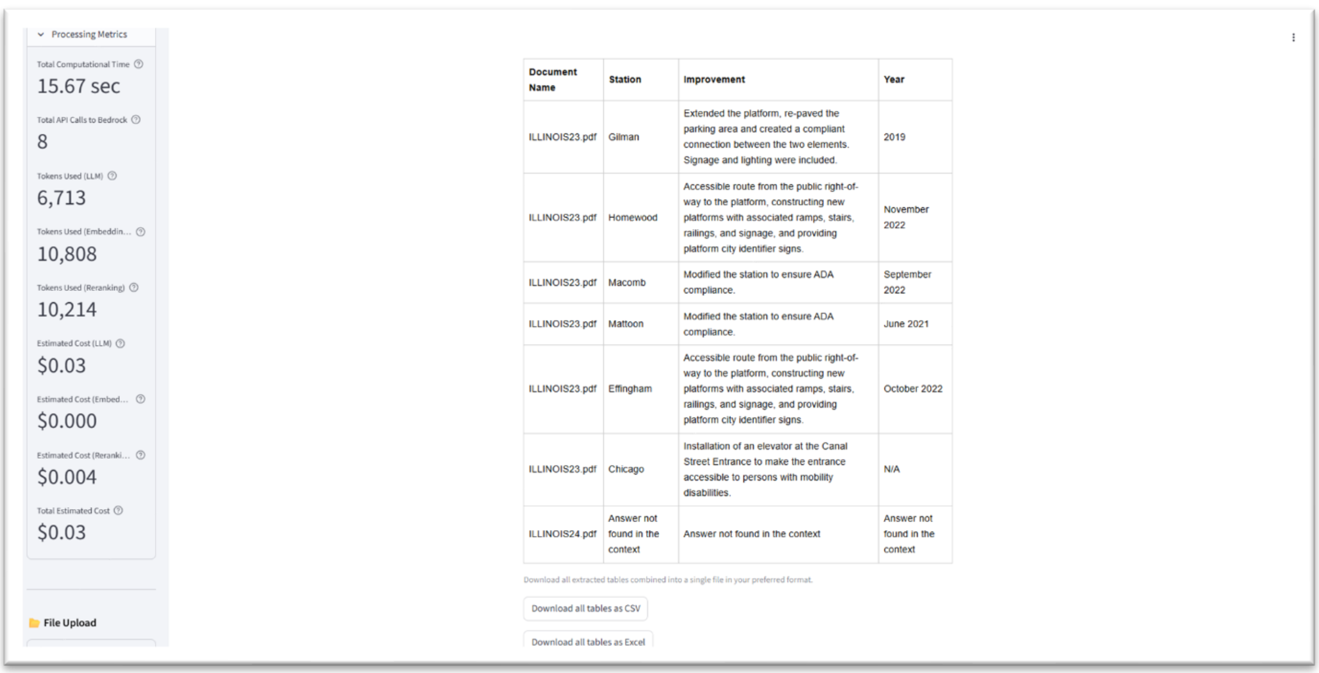
Metadata Extractor: Generated Metadata & Processing Metrics



Tabular Data Extractor: Home Page

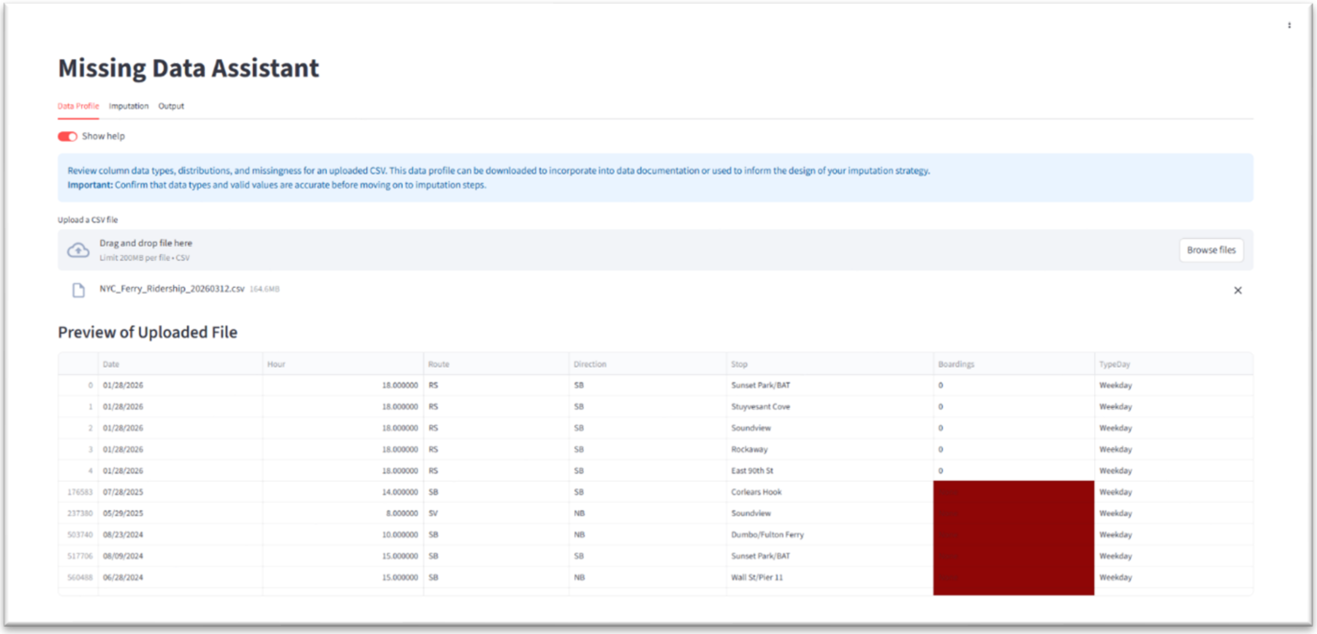


Tabular Data Extractor: Output Preview & Processing Metrics

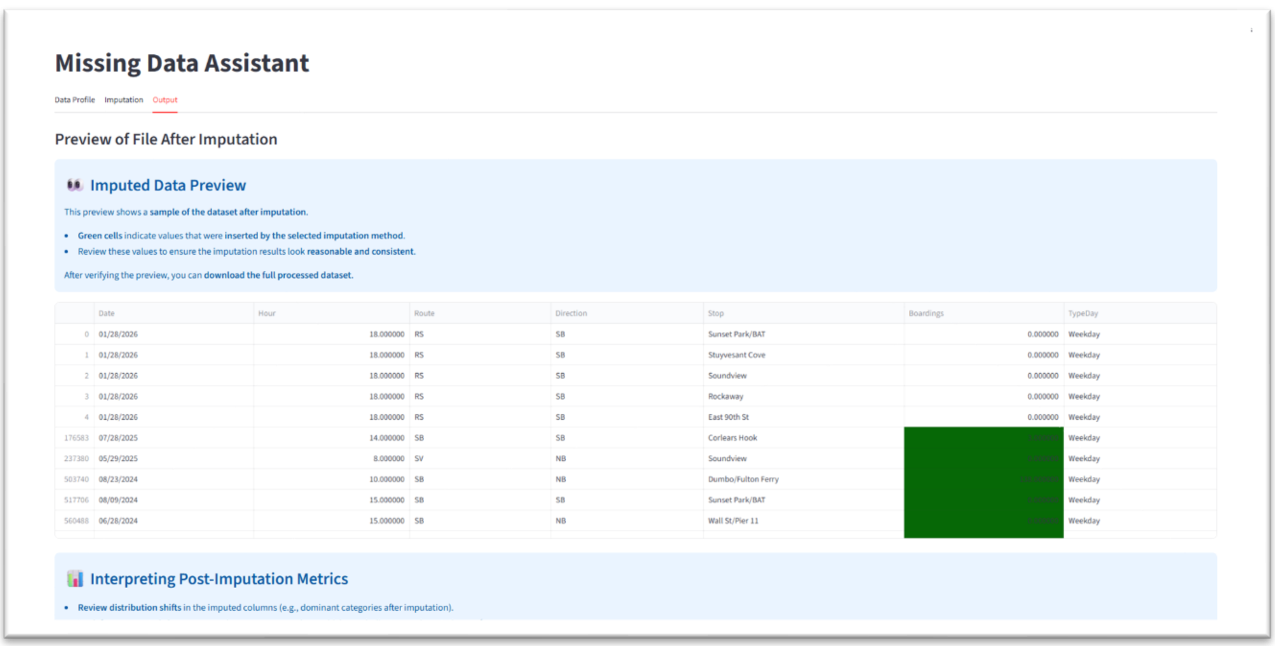


Document Name	Station	Improvement	Year
ILLINOIS23.pdf	Gilman	Extended the platform, re-paved the parking area and created a compliant connection between the two elements. Signage and lighting were included.	2019
ILLINOIS23.pdf	Homewood	Accessible route from the public right-of-way to the platform, constructing new platforms with associated ramps, stairs, railings, and signage, and providing platform city identifier signs.	November 2022
ILLINOIS23.pdf	Macomb	Modified the station to ensure ADA compliance.	September 2022
ILLINOIS23.pdf	Mattoon	Modified the station to ensure ADA compliance.	June 2021
ILLINOIS23.pdf	Effingham	Accessible route from the public right-of-way to the platform, constructing new platforms with associated ramps, stairs, railings, and signage, and providing platform city identifier signs.	October 2022
ILLINOIS23.pdf	Chicago	Installation of an elevator at the Canal Street Entrance to make the entrance accessible to persons with mobility disabilities.	N/A
ILLINOIS24.pdf	Answer not found in the context	Answer not found in the context	Answer not found in the context

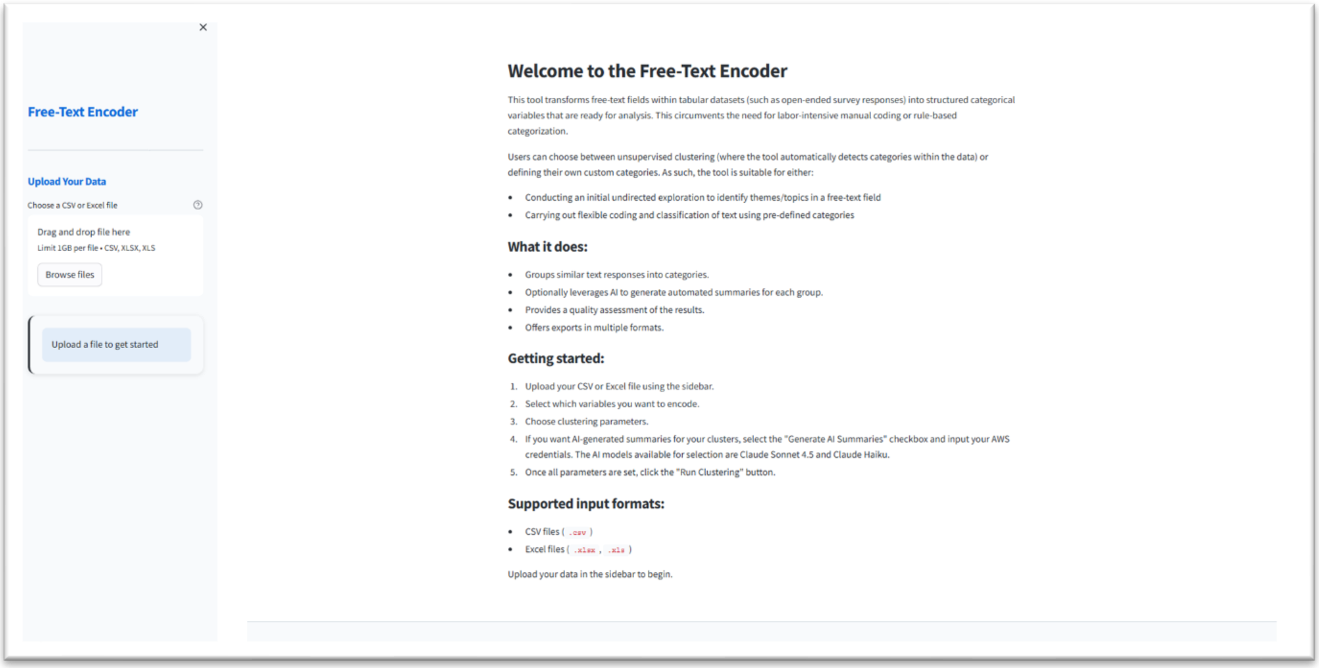
Missing Data Assistant: Home Page - Data Profile



Missing Data Assistant: Output – miceRanger



Free-Text Encoder: Home



Free-Text Encoder: Clustering Workflow – Cluster Summary