NORC at the University of Chicago

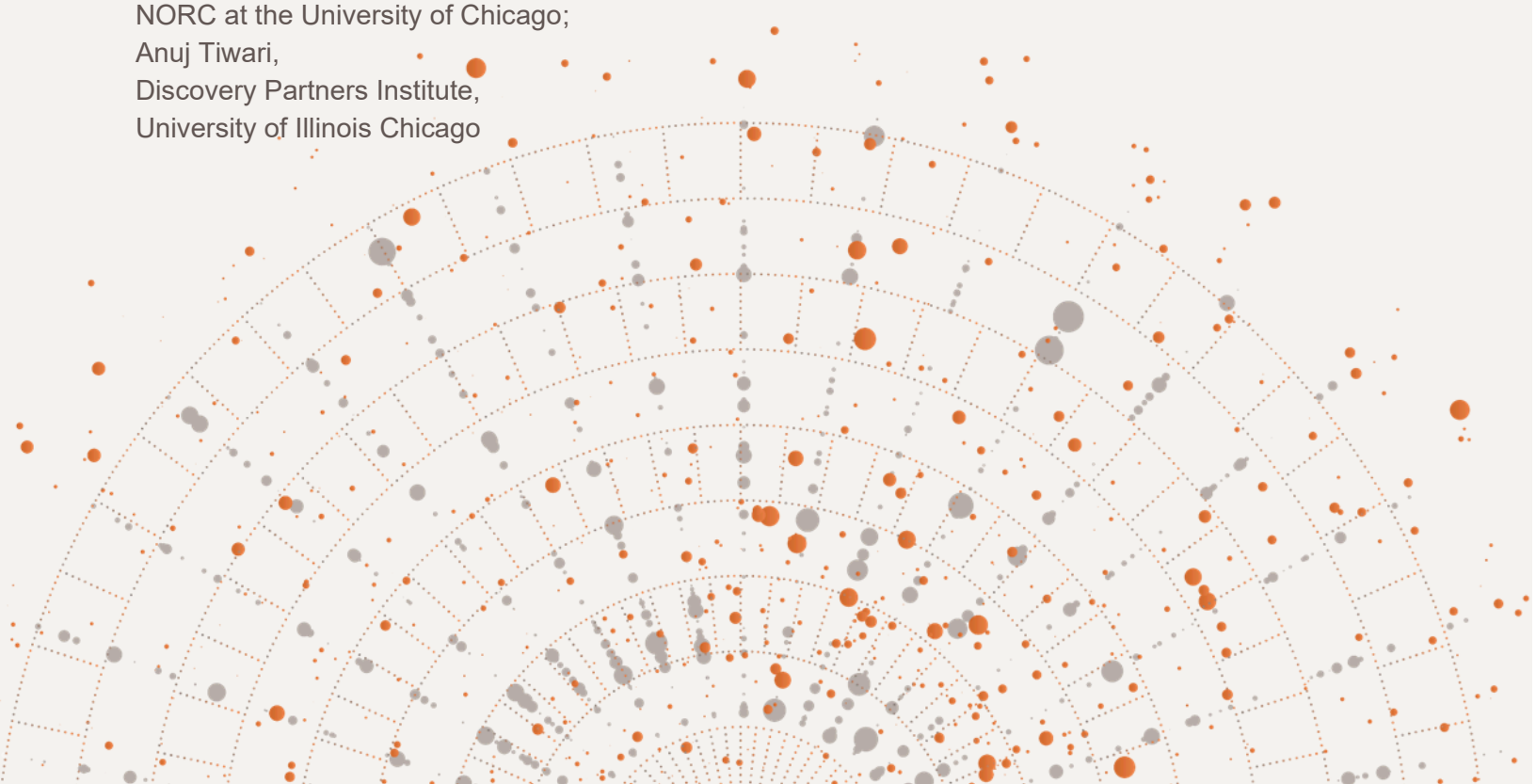
**FINAL REPORT**
**September 2025**

# AI-DQSI Framework Plan:

## Artificial Intelligence for Enhancing Data Quality, Standardization, and Integration

**Presented by:**
Sara Lafia,
Zachary H. Seeskin,
Julia Dennis,
Emily Wiegand,
NORC at the University of Chicago;
Anuj Tiwari,
Discovery Partners Institute,
University of Illinois Chicago

**Presented to:**
National Center for Science and Engineering Statistics at the National Science Foundation:
Project AI-DQSI-24



# 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-DRN25-053).

# Executive Summary

This report investigates the potential of Artificial Intelligence (AI) to enhance data quality, standardization, and integration (DQSI) within the context of the federal statistical system. It focuses on areas in which AI has the potential to improve DQSI, especially for non-traditional data sources. In this report, we use “non-traditional data sources” to refer to data not collected for statistical purposes. Non-traditional data sources offer analytical opportunities but pose unique challenges due to limited metadata, inconsistent formatting, and challenges for variable harmonization when combined with other data sources. This report identifies promising application areas where AI may assist with DQSI challenges and concludes with recommendations to inform the development of an AI toolkit for a future National Secure Data Service (NSDS). The toolkit will enhance data quality while promoting data standardization practices, improving data integration methodologies, and strategically leveraging the capabilities of AI to address common data issues.

This report includes findings from a literature scan, expert interviews, and an AI tool review. The project team first conducted a literature scan to identify existing research, reports, and publications relevant to the challenges of preparing and using distinct types of data identified as important in the federal statistical system. We then led fourteen semi-structured interviews with experts across nine organizations to gain perspective into these challenges, including eight interviews with federal agency staff, three data privacy and ethics specialists, and three subject matter experts. Finally, insights from the literature scan and interviews informed the selection of five AI tools for an in-depth review to examine the capabilities and limitations of some currently available AI tools.

We identify four main types of data relevant to federal statistics – survey data, administrative data, private sector data, and geospatial data – each with unique DQSI considerations. Survey data, which are primarily designed to address specific research and policy questions, face challenges related to nonresponse bias, timeliness, coverage issues, and measurement errors. AI opportunities to improve DQSI in survey data include automating coding, cleaning, and imputation, as well as using natural language processing to format open-ended responses. Administrative data, which are collected as a byproduct of public program administration, present challenges related to data linkage, terminology differences, coverage limitations, and data lags. AI can assist with record linkage, error detection, automated data processing and cleaning, and anomaly detection. The use of private sector data, which can offer timely or locally relevant indicators for trend detection, introduces additional concerns about coverage, privacy, access, and documentation quality. AI can facilitate data integration, bias mitigation, predictive modeling, and pattern recognition for these data. Geospatial data, which capture location-based attributes, come with issues such as positional accuracy, spatial and temporal resolution, metadata gaps, and formatting inconsistencies. AI can support image recognition, change detection, predictive mapping, spatial clustering, and error detection. Across these data types, the availability and completeness of metadata stand out as critical factors for effective data reuse.

Our research finds that AI use raises legal and ethical considerations and requires responsible and trustworthy implementation within the federal statistical system. Privacy concerns, especially the risk of person or business re-identification, are amplified by AI capabilities. Laws ensuring consent and privacy protections for input data sources must be maintained across integrated data sources. Regarding ethical considerations, statistical agencies must be cautious regarding how algorithmic biases from AI could lead to differential treatment of subgroups in a dataset. High quality metadata are important to support reliable inputs and outputs for AI processes. Further, care must be taken for AI to address differences in representation and coverage across datasets while seeking fairness in outputs. Finally, oversight through human-in-the-loop workflows can help avoid hallucinations and inaccuracies in AI output. Sufficient human review of AI output can be critical to avoid harmful results from mistaken output. However, there should be a balance, as involving humans too much may offset the efficiencies gained through AI adoption.

The team's review of currently available AI tools (TurboCurator, OpenRefine, the RecordLinkage Toolkit, Google Earth Engine, and ArcGIS Pro with ArcGIS AI) showcased diverse capabilities for enhancing DQSI. The TurboCurator tool uses ChatGPT to improve metadata, while OpenRefine cleans unstructured data. The RecordLinkage Toolkit links variables across datasets using probabilistic methods. Google Earth Engine processes large-scale geospatial datasets, and ArcGIS Pro applies AI to prepare and analyze spatial data. Each tool offered strengths and limitations across application areas, scalability, usability, and costs. The tools reviewed also raised important privacy and ethical considerations related to potential biases, disclosure risks, security vulnerabilities, the challenges of transparency, and the need for user expertise. Careful consideration of these factors is critical for the effective and responsible use of AI tools within the federal statistical context.

This report concludes by identifying promising application areas where AI may assist with data quality, standardization, and integration (DQSI) challenges and makes several key recommendations to inform the development of an AI toolkit for a National Secure Data Service (NSDS):

Recommendations for using AI to enhance DQSI include that:

1. AI may be leveraged to automate data cleaning and validation tasks across all data types, such as identifying and correcting errors, inconsistencies, and outliers.

2. AI may simplify the integration of geospatial data into statistical applications by automating and standardizing the extraction of structured features, such as road networks and their attributes from satellite imagery.

3. Large Language Models can be leveraged to enhance existing data documentation and metadata, thereby improving data discoverability and usability.

Recommendations for protecting privacy and ensuring ethical use of AI include that:

4. To address algorithmic biases, strategies for agencies to consider and explore include fairness audits, bias correction techniques, and training novel AI systems on curated, representative data.

5. Transparency and explainability when using AI can be upheld by disclosing which records and variables have been used, as well as the methods applied in data processing and integration, allowing users to understand potential limitations that may influence model performance.

6. 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.

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| CIPSEA | Confidential Information Protection and Statistical Efficiency Act |
| DQSI | Data Quality, Standardization, and Integration |
| FCSM | Federal Committee on Statistical Methodology |
| GEE | Google Earth Engine |
| GIS | Geographic Information System |
| LLM | Large Language Model |
| NSDS | National Secure Data Service |
| PII | Personally Identifiable Information |
| SME | Subject Matter Expert |

# Introduction

Artificial Intelligence (AI) has the potential to enhance work undertaken by federal statistical agencies to process, format, standardize, and integrate data from various sources. A 2020 canvas of AI use across 142 federal departments and agencies found that nearly half (45%) of these agencies had already experimented with AI-related tools and were incorporating them to improve daily agency operations at that point (Engstrom et al., 2020). This report addresses the role that AI can play to support the integration of traditional data sources, like survey data, with non-traditional data sources[1], such as administrative records, private sector data, and geospatial information, to support cross-agency work and increase organizational capacity (National Academies of Sciences, Engineering, and Medicine, 2023). While data integration enables unique analytical opportunities that promise to increase efficiency, it also introduces new challenges and potential threats to data quality that must be addressed (Yarkoni et al., 2021; O'Toole et al., 2024).

## Goals

Many sources of non-traditional data have analytical potential, yet challenges like limited metadata, formatting inconsistencies across reporting entities, and unharmonized variables across diverse sources limit their usability. This report explores how AI might mitigate these challenges by augmenting human expertise in data curation and data transformation activities. The report also explores challenges for using AI in the federal statistical system, including the extra care needed to ensure data privacy and the importance of transparency and explainability of statistical processes.

This report informs the development of a toolkit that leverages AI to enhance data quality, standardization, and integration (DQSI) activities for statistical agencies, including within a future National Secure Data Service (NSDS). The toolkit will promote data standardization best practices, improve data integration methodologies, and strategically leverage AI capabilities to address persistent data-related issues.

## Terminology

Throughout this report, the term Artificial Intelligence, or AI, is used broadly to describe a range of emerging technologies, from narrow machine learning applications to general-purpose Large Language Models (LLMs), that use computational methods to perform tasks involving problem solving, learning, or decision-making based on a set of human objectives (National Artificial Intelligence Initiative Act, 2020). What distinguishes AI systems is their ability to learn based on inputs, which makes them particularly promising for automating and scaling complex data curation, integration, and analysis tasks involving diverse data types in a potential future NSDS. This report addresses key data quality considerations for

[1] In this report, we refer to data not collected for statistical purposes as non-traditional data sources.

AI applications across the domains of utility, objectivity, and integrity from the Federal Committee of Statistical Methodology's (FCSM's) Framework for Data Quality, such as granularity, accuracy and reliability, and confidentiality (FCSM, 2020).

## Report Structure

The report is structured to provide an overview of the research process and main findings. We begin with a discussion of the outreach and data collection methods. These methods include a detailed literature scan, expert interviews, and a review of available AI tools. We then present findings on key data types and the relevant DQSI considerations for each data type. Next, we detail ethical and privacy considerations for AI use. Finally, we present the results of our evaluation of selected AI tools, assessing their current capabilities and limitations in the context of DQSI for federal statistical purposes. We conclude with recommendations for AI adoption in standardization and integration workflows, intended to help guide tool development in the next phase of the project.

# Methods

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.

The second phase of the literature scan focused on identifying existing best practices and challenges related to DQSI for each data type. The information gathered also informed the design and planning of the expert interview protocol as well as the subsequent review of existing AI tools. We identified and reviewed existing research, reports, and publications relevant to each of the four data types. We searched using tailored keywords and phrases related to each data type, focusing on DQSI considerations for the use of AI. Key search terms for each data type are included in **Appendix 1**. After our initial search, we followed additional citations from publication reference lists and followed up on citations recommended by interviewees.

The research team used ChatGPT to explore and organize themes emerging from publicly available materials, augment conventional search methods, and ensure a thorough exploration of the relevant literature. More information about the specific models used and prompts issued is available in **Appendix 2.**

## Expert Interviews

We conducted fourteen semi-structured interviews with experts across nine organizations including eight interviews with federal agency staff, three data privacy and ethics specialists, and three subject matter experts (SMEs) summarized in **Table 1**. We recruited SMEs with expertise in curating diverse social science data and integrating data across providers for evidence-building. The SMEs had experience working across stakeholders including state and local governments, nonprofit organizations, and research organizations including universities, and provided valuable insights that complemented the expertise of the federal agency staff and data privacy and ethics experts.

**Table 1.** Summary of Interviews

| Interview Type | Number Interviewed | Description of Interviewee Expertise |
|---|---|---|
| **Federal agency staff** | 8 | Provided insights into current data practices, challenges, and potential AI applications within their respective agencies |
| **Data privacy and ethics experts** | 3 | Offered perspectives on the ethical considerations, privacy risks, and mitigation strategies related to AI implementation in data processing and integration |
| **Subject matter experts** | 3 | 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 and followed a pre-defined interview guide, which permitted flexibility to explore themes that emerged from each conversation. Three of the federal agency staff interviews were conducted as panels with multiple staff members from the Bureau of Transportation Statistics (BTS) across different offices. The interview questions for each interview type are included in **Appendix 3**. Findings from the interviews are included throughout the report but are not attributed directly to interviewees or their agencies.

## Review of AI Tools

Our work included a review of five existing AI tools for DQSI to investigate the capabilities of current tools and understand the limitations and any potential areas of concern, particularly for the needs of federal statistical agencies. A goal was to motivate issues for focus in further development and refinement of AI tools for DQSI needs across different data types.

Our identification of AI tools for review was informed by the literature scan and expert interviews. The team prioritized tools based on their potential to address issues that the team identified related to DQSI challenges across the four data types. In addition, we ourselves used AI and LLMs in this research process to search public materials for use cases, AI applications, and recent updates in tool capabilities that may have been missed in conventional searches. More information describing the models used and prompts issued is included in the **Appendix 2.**

We also developed evaluation criteria to systematically review and compare the selected tools. Each tool was evaluated based on publicly available information including documentation, case studies, and

user reviews to determine capabilities, limitations, and potential suitability for enhancing DQSI for the work of federal statistical agencies.

# Findings: Data Types

The following findings regarding data types, their statistical uses, DQSI considerations, and AI opportunities are informed by insights that the research team gathered from the literature scan and expert interviews. The literature scan provided a broad understanding of data types and their associated challenges in the federal statistical context, while the expert interviews offered detailed, domain-specific perspectives from practitioners and experts working in this context.

## Taxonomy of Data Types

To inform the review of opportunities for AI across data types, the research team developed a taxonomy of data types relevant to federal statistics, accounting for differences in data collection methods and data structures. A recent National Academies report distinguishing data types and methods provided a key starting point for developing the taxonomy (National Academies, 2023). The report draws distinctions among probability surveys or censuses; administrative records collected through the administration of government programs; commercial data collected by the private sector; geospatial data including that derived from sensors, satellites, and other location data; nonprobability or convenience samples; and data from social media, webscraping, and crowdsourcing.

Similar discussions of data types were identified in earlier reports issued by the National Academies (Groves & Harris-Kojetin, 2017a, 2017b; Harris-Kojetin & Citro, 2021). These reports distinguished types of data that support the integration and use of federal statistical data; for example, potential sources of error vary by data type and determine whether data can support data integration activities intended to improve the quality of official statistics (Biemer et al., 2014; Citro, 2014).

We prioritized four of these data types (survey, administrative, private sector, and geospatial) for further investigation in this project. These four data types were determined to be more relevant for the project and more mature for use in federal statistics. While webscraping has uses in the federal statistical system, webscraping presents significant data quality challenges due to the lack of centralized control and vulnerability to manipulation by external actors; similarly, nonprobability or convenience samples pose methodological challenges for statistical inference as it can be difficult to draw reliable and generalizable conclusions from samples that are not chosen at random.

**Table 2** summarizes our findings regarding statistical uses and DQSI considerations for each of the four data types that were the focus of this research. These findings are discussed in more detail in the following sections.

**Table 2.** Summary of Common DQSI Challenges and AI Opportunities by Data Type

| Data Type | Description | Statistical Uses | DQSI Challenges | AI Opportunities |
|---|---|---|---|---|
| **Survey** | Information collected through probability-based samples or censuses designed to represent populations of interest | Benchmarking, official statistics, understanding population characteristics, trend analysis, program evaluation | Timeliness, granularity, sampling errors, non-response biases, measurement errors, coverage issues, reporting errors, terminology or definition differences, instrument changes over time | Automated coding, cleaning, imputation, natural language processing for open-ended responses, bias detection |
| **Administrative** | Records collected by government agencies on individuals or groups as part of the routine administrative procedures for a program | Supplementing or replacing survey data, auxiliary data for small area estimation, longitudinal analysis, program monitoring, creating population registers | Data linkage challenges, terminology or definition differences, coverage limitations, data lags, errors, or inconsistencies | Record linkage and de-duplication, error detection, data cleaning, anomaly detection |
| **Private Sector** | Data collected and managed by private sector organizations, including from data aggregators and single companies | Real-time indicators, trend detection, tracking of consumer behavior, study of service delivery | Coverage and representation, privacy concerns, access restrictions, lack of transparency about data processing | Data integration and harmonization, bias mitigation, predictive modeling, pattern recognition |
| **Geospatial** | Information that includes location-based attributes, capturing the geographic dimensions of features, phenomena, or events | Spatial sampling, spatial linkage, exposure mapping, predictive modeling | Spatial accuracy, metadata gaps, formatting inconsistencies, varying coordinate systems, integration across spatial units | Error detection, feature extraction from imagery, interpolation, schema alignment, predictive modeling, real-time monitoring |

# Survey Data

Survey data, in the context of this report, refer to information collected through probability-based samples or censuses designed to represent populations of interest. These surveys employ well-defined variables and questions, contributing to higher data quality and comparability over time (National Academies of Sciences, Engineering, and Medicine, 2023). This structured information, gathered directly from entities such as individuals or businesses, can capture both objective behaviors and subjective information like sentiment (Laaksonen, 2018).

## Statistical Uses

Surveys are primarily designed and collected to address specific research and policy questions (Lohr & Raghunathan, 2017). This includes studying relationships among variables, assessing the impact of policy decisions, creating statistical classifications and standards, and developing key economic indicators. Survey data are also used for estimating population characteristics, such as poverty rates, consumer spending patterns, and health insurance coverage rates. Study of survey data is also used to monitor societal conditions and trends (Laaksonen, 2018).

While surveys have historically been the main source of data for federal government statistics, rising costs, declining response rates, a lack of detail for state and local levels, and lags in timely updates have made the administration and use of survey data more complex in recent years (Groves & Harris-Kojetin, 2017b). Interviewees mentioned the need to consider additional data sources to supplement and offset some of these limitations with survey data.

Examples of well-known survey data identified in literature and discussed in the expert interviews include demographic surveys like the American Community Survey and the Current Population Survey and economic surveys like the Current Employment Statistics program. Survey designs include cross-sectional studies collecting data at a single point in time and longitudinal or panel studies which follow the same units over time.

## Needs Regarding DQSI

Interviewees described several challenges associated with maintaining high-quality, standardized survey data and enabling its integration with other data types. Quality considerations related to accuracy and reliability include addressing nonresponse (cases where respondents do not participate or skip certain questions), coverage issues, and measurement errors (FCSM, 2020). Interviewees talked about timeliness and frequency as important considerations, given the often-high cost associated with conducting high-quality surveys. A trade-off often exists between investing more time and resources to obtain higher quality, more representative data and the practical constraints of budget and deadlines. Granularity can also be a limitation, as sample sizes of probability surveys are often insufficient to provide accurate estimates for small subgroups and/or small geographic areas within the population (National Academies, 2023).

Data integration efforts often involve combining survey data with other data sources, such as administrative records or private sector data, to improve overall data quality and add additional variables for analysis. Many of the experts we interviewed described undertaking data integration activities to enable the analysis of survey data alongside other data sources and types. The harmonization of survey data itself is also essential to ensure comparability across different data sources and time periods (Lohr & Raghunathan, 2017).

### Opportunities for AI

AI offers many opportunities to enhance survey data across its lifecycle. Several interviewees described opportunities for leveraging AI techniques such as natural language processing to help review and structure text, including open-ended responses from transcripts or audio recordings. AI can also expedite data entry and coding by assigning categories or labels to text. Interviewees described the potential to automate data processing tasks such as coding, validation, consistency checks, cleaning, and categorization using AI. Furthermore, interviewees also described how LLMs can be leveraged to make documentation, such as codebooks, and metadata more detailed, helping future users find and reuse survey data. In addition to direct data manipulation, interviewees also mentioned that AI could help monitor issue logs to identify data quality issues or privacy concerns. Finally, AI can facilitate analysis by enabling the linking of survey data to other sources and harmonizing variables across different datasets (Yarkoni et al., 2021).

## Administrative Data

Administrative data are records collected by government agencies on individuals or groups as part of the routine administrative procedures for a program (Vale, 2011). These records are not collected for explicit research purposes but rather are a byproduct of program administration and regulatory activities (Iwig et al., 2013). Examples of administrative data include information from income tax forms, social security data, health and education records, and death certificates. This category of data includes both federal and state/local administrative sources.

### Statistical Uses

Administrative data play a crucial role in various statistical applications. Administrative data are often used to estimate or monitor trends, particularly for specific subpopulations or small geographic areas, where survey data may be limited (Cole et al., 2020). The analysis of administrative data also provides insights into long-term outcomes, such as multi-generational effects of policies and interventions, providing a critical link between social science research and policy (Penner & Dodge, 2019). Compared with survey data, they also offer much larger sample sizes and have fewer challenges related to attrition, non-response, and measurement error than traditional data from surveys (Card et al., 2010). These features, combined with the increasing availability of large administrative datasets, enable more rigorous testing of economic theories (Einav & Levin, 2013). Administrative data sharing across federal,

state, and local entities supports applications of linked administrative data, program evaluation, policy development, and the creation of sample frames for surveys (Prell et al., 2009). Interviewees described analyzing administrative data on its own or using it to augment surveys or other data sources in several ways, including survey frame construction, data imputation, calibration weighting, providing auxiliary data for small area estimation, and facilitating record linkage. Utilizing administrative data in these ways may improve data quality, reduce data collection costs, reduce respondent survey burden, and improve evidence-building for policy and program evaluation (Prell, 2019). Analyzing administrative data provides novel research opportunities, helps fill gaps in primary data, and may also provide savings over primary data collection (Connelly et al., 2016).

Many of the interviewees we spoke with described leveraging administrative data to enhance their data products. Selected examples of data integration from literature include the integration of Census data with Internal Revenue Service and Centers for Medicare & Medicaid Services records at the Bureau of Economic Analysis to help measure the economy, integration of crime data from the Bureau of Justice Statistics with federal prison facilities data to assess operations, and integration of data from the Social Security Administration and Department of Housing and Urban Development to develop sample frames (O'Hara & Medalia, 2018). Administrative records have also been used to create and assess the quality of novel research products, such as in the Decennial Census Digitization and Linkage Project, which links historical census files using administrative and non-routine survey data (Alexander & Genadek, 2023).

## Needs Regarding DQSI

In OMB memorandum M-19-15, the Office of Management and Budget encourages federal agencies to leverage available administrative data for statistical purposes (Vought, 2019). Data infrastructure that upholds reputation, reciprocity, and trust is needed to support the use of administrative data by policymakers, researchers, and programmatic agencies for operational and research purposes (Lane, 2018). To enable this vision, agencies must document and share information about data quality to help secondary data users determine data fitness for use in new statistical application areas (FCSM, 2020).

Interviewees described how analyzing administrative data, especially when linking data, can present challenges due to underlying data quality constraints. Administrative data linkage quality can be determined using gold standard reference data, performing post-linkage validation, performing sensitivity analysis, and comparing characteristics of linked and unlinked administrative data (Harron et al., 2017). Given that administrative data have been collected for a particular purpose, adapting them for use for new purposes raises critical data quality issues (Iwig et al., 2013).

Factors such as relevance, accuracy, completeness, timeliness, accessibility, clarity/interpretability, coherence/consistency, and comparability help assess data fitness for purpose (Seeskin et al., 2019). Interviewees mentioned that since administrative data are not collected for research purposes, they often require substantial manipulation to be usable for statistical analysis. Interviewees also described that coverage issues can also arise, as the populations included in administrative data may not fully represent non-program participants. Differences in formatting and standardization, particularly across

state and local data sources, along with the limited availability of unique identifiers for linking records pose integration challenges. Lastly, administrative datasets often contain outliers and nonsensical values that require careful attention.

In terms of usability, interviewees talked about the difficulty of gaining access to or sharing administrative records. Data sharing introduces data quality issues, which arise when data are reused outside of the original context or purpose for which they were collected; sharing administrative data across federal, state, and local contexts also necessitates standardization and requires shared understanding of data definitions across agencies (Prell, 2019; Prell et al., 2009). One recurring challenge includes privacy concerns posed by Personally Identifiable Information (PII) (Jarmin & O'Hara, 2016). Missing data or other inconsistencies can also make it difficult to use administrative data in combination with other data sources for novel research purposes (Groves & Harris-Kojetin, 2017b). Interviewees described inadequate documentation of data limitations as a related challenge, making it difficult to reuse data responsibly. As an example, administrative data and electronic health records across different systems have been shown to have substantial issues with missing data for race and ethnicity, which can have downstream consequences for understanding outcomes for different demographic groups (O'Hara & Rhodes, 2023). Decisions about how to process administrative data, such as deduplicating individuals, imputing missing values, and cleaning identifiers, also have consequences for analysis (O'Hara & Medalia, 2018). Interviewees cited inadequate resources to prepare data documentation, extract data, and transmit data as barriers to administrative data sharing and use.

### Opportunities for AI

AI and related methods may offer promising solutions to address administrative data quality challenges. Many interviewees expressed interest in applying AI to improve data quality through cleaning tasks such as formatting, deduplication, outlier detection, imputation, validation, and cross-checking. AI can also enhance analysis by facilitating data integration through processes like record linkage to add context to extant data and by detecting changes over time across files. Finally, AI can assist in data documentation efforts by enhancing inconsistent or incomplete metadata.

## Private Sector Data

Private sector data, also referred to as corporate or vendor data, encompass data collected and managed by private sector organizations; these organizations include data aggregators as well as single companies that collect and provide data on their customers (Groves & Harris-Kojetin, 2017b). Examples include commercial property tax data from aggregators (Seeskin, 2018), transaction data like retail scanner data from market research firms like IRI (Muth et al., 2016), and credit records collected and shared by information brokers like Dun & Bradstreet (Kallberg & Udell, 2003). Private sector data are increasingly relevant to policymaking and governance as they provide timely and granular insights into social and economic dynamics that complement traditional survey or administrative data collected by agencies (Einav & Levin, 2013; Future of Privacy Forum, 2017).

## Statistical Uses

Like administrative data, private sector data offer a range of statistical, analytical, and logistical utility. The real-time and granular nature of some private sector data also provides opportunities for use as timely economic or behavioral indicators (Einav & Levin, 2013). The availability of large-scale private sector datasets, such as commercial scanner data and online transaction data, is changing econometric analysis, enabling more granular and real-time insights into consumer behavior (Varian, 2014). For example, commercial scanner data provides the Department of Agriculture's Economic Research Service with insights into health and nutrition research based on retail food purchases (Muth et al., 2016). The Consumer Price Index also relies on private sector scanner data to improve its accuracy; however, obtaining this kind of commercial data often requires negotiations with data providers or purchasing data products through third-party vendors and processing large, messy datasets that often lack detailed documentation (International Monetary Fund, 2020). Products like commercial property tax data can be broadly applicable to improve survey estimates when they are of sufficient quality; for example, they may be considered for use to reduce respondent burden by eliminating the need to ask survey questions through record linkage to survey respondents (Seeskin, 2016). Private sector entities also collaborate with statistical agencies to develop complementary statistics, such as economic indicators, and use official statistics as benchmarks for comparison (Groshen, 2021). While a large share of official statistics are based on official surveys, public and private non-survey data collected for non-statistical purposes have been used to construct economic activity indicators; for example, Ward's/JD Powers/Polk private source data on auto sales, price, and registration may be used in the development of early Gross Domestic Product estimates when official monthly indicators are not available (Landefeld, 2014).

## Needs Regarding DQSI

Data quality considerations for private sector data overlap with considerations for administrative data. Interviewees who had experience working with private sector data cited data quality issues as a major limiting factor. Like administrative data, many sources of private sector data are "found data", which are not primarily collected for statistical purposes (Seeskin, 2018). This can lead to generally lower levels of trust and perceptions of veracity compared to government-collected data (Seeskin et al., 2018).

A recent survey of private sector data use across federal agencies conducted by the Interagency Council on Statistical Policy identified several main challenges that agencies encounter when working with private sector data (Interagency Council on Statistical Policy Private Data Steering Group, 2023). Poor quality documentation of products and methods, such as data collection and preprocessing procedures, is a common challenge. This opinion was echoed by several interviewees who found the quality of private sector metadata limiting to the point that the data could not be reliably used. Verifying and conducting quality control on private sector data can also be difficult due to a lack of transparency and access to source data (Landefeld, 2014). Without this information, it can be difficult to assess data completeness and suitability for a given task.

Steps to enhance data quality can include cleaning, sampling, reweighting, and documentation of known issues or lags in delivery or timeliness of data; however, remediating quality issues requires an understanding of the data limitations prior to use (U.S. Bureau of Labor Statistics, 2024). Private sector data contain potential biases, measurement errors, and require standardization across data providers to enable comparative analysis, which is not possible if high quality documentation is not provided (Varian, 2014).

Usability issues include privacy concerns, especially for location-based data (Cyphers & Gebhart, 2019), and concerns about perceptions surrounding the use of personal data collected with or without individuals' awareness (Auxier et al., 2019). Interviewees expressed concerns about additional constraints, including the risk of vendor lock-in due to licensing costs, data product longevity, and product sustainability. For example, if companies are acquired or experience changes in leadership, the design of data products may change, leading to inconsistencies and discontinuities in access to data.

### Opportunities for AI

Opportunities for AI to enhance private sector data are similar to those for administrative data. Interviewees described the potential for AI to improve data collection and preprocessing by enhancing data delivery and ensuring consistent data formatting and organization. The prospects for administrative data use in the federal system are generally more favorable than for private sector data, however, because trust in administrative data is generally higher given that it is subject to higher levels of transparency and that data are generated through the administration of public programs (Groves & Harris-Kojetin, 2017b).

## Geospatial Data

Geospatial data refer to information that includes location-based attributes, capturing the geographic dimensions of features, phenomena, or events (Goodchild, 2007). Geospatial data include raster files such as GeoTIFFs, vector files such as shapefiles, and geo-enabled tabular datasets (Longley et al., 2015). Geospatial data may originate from satellite systems (e.g., Landsat, MODIS), government mapping initiatives (e.g., TIGER/Line files from the U.S. Census Bureau), agricultural mapping products (e.g., the National Agricultural Statistics Service Cropland Data Layer), and transportation or environmental monitoring (Boryan et al., 2011; U.S. Census Bureau, 2022).

Many federal agencies, local governments, research institutions, or private sector sources produce authoritative geospatial data (National Research Council, 2007). Several interviewees described using geospatial data to support planning, monitoring, and operational decision-making activities. Geospatial data are used to map population distributions, track environmental changes, identify infrastructure vulnerabilities, and model spatial relationships across sectors such as health, transportation, agriculture, and housing (Miller & Goodchild, 2015). The availability of consistent geospatial identifiers also allows these datasets to be integrated with survey, administrative, or private sector data to support spatial analyses (Selmy et al., 2024).

## Statistical Uses

Geospatial data support a wide range of statistical applications, with a particularly important role in survey planning and analysis. At the planning stage, geospatial data are used in location-based sampling methods by enabling the definition of stratified or clustered areas based on geography (Haining, 2003). Geospatial data also support the delineation of study boundaries, service areas, and target regions, helping improve the accuracy of sampling (Reibel, 2007). Several interviewees described how location information in geospatial data enables data integration and linkage. In data aggregation and weighting, geospatial data are used to group and summarize information within defined geographic areas such as census tracts, ZIP codes, or custom boundaries (Goodchild, 2007). Geospatial data allow for adjusting results based on population size, land use, or access to services to make estimates more accurate (Brunsdon et al., 1998).

In addition to supporting surveys, geospatial data play a critical role in linking and enriching other types of datasets, including administrative records and private sector data. For example, geographic references such as coordinates, postal codes, or address ranges allow for the spatial integration of health records, school enrollment data, social service usage, and retail activity (Longley et al., 2015). Interviewees highlighted how spatial linkage enables the combination of disparate datasets, survey data with hospital visits, or consumer behavior with climate exposure, to support cross-sector statistical analysis and policymaking. Geospatial data also support a range of spatial statistical methods, including spatial autocorrelation, geographically weighted regression, and hotspot analysis (Anselin, 1995; Fotheringham et al., 2003). Spatial interpolation to estimate values in places without direct data supports prediction models that account for both temporal and spatial variation (Cressie, 1993). Finally, geospatial data enhance data communication through maps and visual tools like heat maps, density plots, or area-based charts. These tools help identify patterns, highlight gaps, and support planning or decision-making (MacEachren, 2004).

## Needs Regarding DQSI

Geospatial data present several challenges related to DQSI. Data quality is determined by spatial, positional, and temporal accuracy (Goodchild, 2007). Several interviewees mentioned how errors in location coordinates, outdated timestamps, and misaligned features can affect the reliability of spatial analysis. Gaps in geographic coverage or missing features such as roads, buildings, or land parcels can also limit completeness, while topological inconsistencies, such as disconnected segments or overlapping polygons, reduce data usability (Burrough et al., 2015).

Interviewees described that lack of comprehensive metadata is another key issue. Incomplete documentation can make it difficult to assess whether a dataset is suitable for reuse, especially when integrating data from multiple sources (Guptill & Morrison, 2013). For example, if a land cover dataset lacks details on classification methods or spatial resolution, it may not be appropriate for modeling or cross-regional comparisons.

Even when data are complete and well-documented, inconsistent standards across sources introduce additional barriers. Interviewees cited differences in coordinate systems, file formats, spatial resolution, and geospatial identifiers such as place names or administrative codes as examples of inconsistencies that can complicate comparisons. Definitions of geographic concepts like “urban” or “rural” may also vary among data sources being linked or blended, leading to classification mismatches (Harvey, 2015).

These challenges become complex when combining datasets from different systems. Interviewees described that integration is difficult when datasets use varying spatial units, such as points versus administrative boundaries. Aligning such data often requires adjusting the spatial resolution, aggregating data to common units, or transforming coordinates, all of which may introduce uncertainty (Goodchild & Li, 2012). Privacy concerns may also arise when linking detailed location data with personal or sensitive information (Curry, 1997). Effective integration depends on consistent spatial hierarchies, clear data lineage, and accurate geocoding (Longley et al., 2015).

### Opportunities for AI

AI can enhance the processing and analysis of geospatial data. Several interviewees described how AI could automate data cleaning and detect spatial errors, including positional inaccuracies, topological issues, and inconsistent attributes (Zhu et al., 2017). One example is the use of deep learning techniques to analyze imagery, identify missing features such as roads or buildings, and support automated digitization through image matching, boundary delineation, and enhanced address geocoding (Ma et al., 2019). AI can also assist with validating geographic identifiers, such as FIPS codes, to ensure location accuracy. Other examples include automated crosswalk creation between different geographic units, such as ZIP codes and census tracts. To support standardization, AI can align mismatched formats, coordinate systems, and classifications, and convert between geospatial file types using schema matching or language models that extract and harmonize metadata (Zhu et al., 2018). Several interviewees expressed interest in using AI to help detect and correct outliers and anomalies with smoothing or denoising rather than simply discarding data. Other interviewees mentioned using AI to support inferences like making urban and rural designations.

AI can also enable integration by linking datasets across different spatial units, recommending appropriate transformations, and reducing uncertainty (Gao, 2021). Privacy-aware methods can further help combine sensitive location data while minimizing disclosure risks (Boutet, 2024). AI also supports predictive modeling and real-time monitoring of data streams from satellites and sensors, enabling continuous quality control in dynamic environments (Reichstein et al., 2019). When thoughtfully applied, these tools make DQSI workflows more efficient, scalable, and responsive to evolving spatial data needs.

## Summary

In summary, the review of survey, administrative, private sector, and geospatial data revealed common threads and distinct challenges concerning DQSI. Survey data necessitates addressing nonresponse,

ensuring timeliness, and balancing granularity with cost, while also pursuing harmonization with other sources. Administrative and private sector data share concerns regarding manipulation for usability, coverage limitations, varying data formats, and PII access restrictions, further complicated by issues of trust, documentation quality, and vendor lock-in for private sector data. Geospatial data present challenges centered on spatial and temporal accuracy, inconsistencies in data formats and reference systems, and the complexities of data interoperability and aggregation. In general, AI holds promise for facilitating work across data types, for example by facilitating code templates in multiple languages, enabling users to work with data in their languages of choice. Metadata completeness also stands out as a critical factor for effective data reuse across data types.

# Findings: Privacy and Ethical Considerations for AI Use

This section discussed privacy and ethical considerations for applying AI in DQSI applications. The findings are drawn from the research team's literature scan and interviews with data privacy experts and are summarized in **Table 3**. These considerations are focused on responsible and trustworthy use of AI in the context of the federal statistical system. Here, privacy addresses the rights of individuals to control the collection, use, and disclosure of their personal information. Privacy focuses on safeguarding data and preventing unauthorized access or identification (National Academies of Sciences, Engineering, and Medicine, 2024). Ethics refers to the moral principles and values that guide the responsible development and deployment of AI, ensuring fairness, transparency, and accountability in its application (Federal Chief Data Officer Council, 2019). It encompasses broader societal implications and the potential for bias or harm.

**Table 3**. Summary of Privacy and Ethics Concerns and Mitigation Strategies

| Category | Consideration | Mitigation Strategy |
|---|---|---|
| **Privacy** | **AI Amplifying Risk of Person or Business Re-identification** | Use of secure methods and environments, such as private LLMs deployed within secure enclaves behind firewalls |
| | **Potential for Data Misuse** | Data governance processes that ensure consent and privacy protections for input data sources are maintained across integrated data sources |
| **Ethics** | **Algorithmic Biases from AI Leading to Unfair Outcomes** | Best practices and principles outlined in existing ethical frameworks from American Statistical Association and the Association for Computing Machinery, potential employment of fairness audits and bias correction techniques, consideration of handling datasets with differences in representation and coverage |
| | **Metadata Quality for AI-Readiness** | Focus on improving metadata practices to ensure reliable AI inputs and outputs |
| | **Hallucinations and Inaccuracies in AI Outputs** | Human oversight to mitigate hallucinations and ensure accuracy |

# Privacy Considerations

AI use in data integration is subject to laws designed to protect privacy and confidentiality. Multiple interviewees highlighted the importance of different laws directing statistical agencies to ensure the privacy of the individuals and entities included in a data file and placing specific constraints on how AI can be applied to integrated data sources. For instance, AI algorithms must be implemented in ways that adhere to the purpose limitations outlined in CIPSEA, ensuring that data integrated using AI is used solely for statistical purposes and not for administrative, regulatory, or enforcement actions against individuals or entities. Similarly, Title 13's stringent confidentiality requirements necessitate that AI models do not inadvertently reveal identifiable information from census data during integration or analysis. Agencies must always respect the different legal requirements applying to different input datasets. For example, in the handling of education records the Family Educational Rights and Privacy Act applies, while for health records the Health Insurance Portability and Accountability Act applies. These laws mandate that AI tools and processes are designed and operated in a manner that upholds the confidentiality promises made to data providers.

Beyond general cybersecurity concerns, AI introduces unique security considerations in the context of data integration. Several interviewees described how the complexity and interconnectedness of AI systems can create new vulnerabilities. For example, the risk of adversarial attacks on AI models trained on integrated sensitive data could lead to the extraction of private information or the manipulation of statistical outputs in ways that traditional security measures might not fully address. The deployment of AI, especially large models like LLMs, requires secure environments such as private enclaves behind firewalls, as emphasized by interviewees. Guidance on secure AI use should strictly limit data sharing to specific statistical purposes and implement the principle of least privilege for data access at appropriate tiers based on the user's specific analytical needs. Techniques like privacy-preserving record linkage, as mentioned by interviewees, become critical for enabling secure AI-driven data integration while minimizing disclosure risks (Ranbaduge et al., 2024).

General privacy risks and surveillance concerns are significantly amplified by AI's advanced analytical capabilities in data integration. The combination of datasets using AI increases the risk of person or business re-identification, as traditional anonymization methods may prove insufficient against sophisticated AI techniques (Landau et al., 2024).

Interviewees stressed that privacy considerations must be integrated from the very beginning of any project, ideally before data collection, to anticipate future AI-driven integration uses. As one interviewee noted, AI's ability to find correlations across integrated datasets can lead to the re-identification of previously classified information. This challenge highlights the inherent tradeoffs between data utility and individual privacy when employing AI for integration. Furthermore, unintended data uses, such as AI-powered linking of employment and medical records, pose significant contextual privacy risks. The potential leakage of sensitive training data embedded within AI models and their outputs (Landau et al., 2024) represents another unique privacy challenge. Data governance processes must ensure that the consent and privacy protections associated with original data sources are rigorously maintained

throughout the AI-driven integration process (National Security Council, 2024). A foundational principle is to collect data with privacy in mind from the outset. Strict adherence to government regulations protecting confidentiality is paramount when using AI tools and LLMs for data integration.

## Ethical Considerations

While the previous section addressed legal and procedural safeguards for data privacy, ethical considerations in applying AI to DQSI delve into the principles guiding its responsible use. Ethical considerations surrounding the application of AI in DQSI must account for sources of error and bias in both the data and in the AI development and implementation. Bias can be introduced in sampling and when unrepresentative data are used to train AI. When AI supports decision making processes for DQSI activities, such as automated data cleaning rule generation or anomaly detection, it has often been shown to produce different results across different subgroups (Yeung et al., 2021). Multiple interviewees described risks such as inadvertent perpetuation and reinforcement of algorithmic biases in decision support tasks, leading to unfair or discriminatory outcomes. One interviewee went further, stressing how the risks of false matches with person-level data can have severe consequences, such as erroneously identifying insurance fraud.

A key concern shared by interviewees is that end-users, either humans or machines, will lack the domain knowledge necessary to reuse data properly and may misinterpret uncertainty in data products (Schwabe et al., 2024). Representativeness and coverage are also critical concerns, as administrative and third-party data often lack detailed information on the populations they capture, making it difficult to assess who is missing from the data. Interviewees explained that data quality issues can be compounded when data are integrated, and several interviewees described how many datasets are not "AI-ready" without metadata explaining limitations and processing steps, such as imputation or PII masking.

Given the potential for AI outputs to introduce quality concerns such as inconsistency, misattribution of statistics, generation of synthetic statistics ("hallucinations"), or outdated or incorrect information (Prem, 2024), most interviewees agreed that AI use requires human input and oversight; no interviewees endorsed using AI to automate decision-making processes. Rather, interviewees envisioned limited uses of AI to support data curation and preprocessing tasks with expected outcomes, which should be guided by human input. Generally, the lack of transparency in AI algorithms was a key concern for use for federal statistical agency products identified across interviews and the literature scan. Interviewees mentioned that a lack of transparency and model explainability can hinder AI adoption and use, especially when observable quality and utility gains are marginal. Furthermore, accountability becomes a concern when AI systems operate unexpectedly, and the explainability of AI methods is crucial for verification and validation of model outputs (Prem, 2024).

To mitigate the ethical risks posed by AI use in DQSI, interviewees proposed a range of strategies. One interviewee pointed to relevant best practices and principles outlined in existing ethical frameworks, such as those provided by the American Statistical Association and the Association for Computing Machinery, which can provide guidance. Fair Information Practice Principles can also provide correctives by helping ensure that personal data are used for the purposes for which they were

collected (Borgesius et al., 2016). Other strategies include fairness audits using existing tools and software, which can also help identify and address potential biases in input data and analyses.

Ethical AI use should involve disclosing the records and variables used from data sources; disclosing the methods applied in data processing and integration; training AI systems on curated, representative data optimized for specific tasks; and designing workflows with humans in the AI loop to mitigate hallucinations and ensure accuracy. One direction worth exploring is that of Explainable AI frameworks that help make AI decision-making processes more understandable, allowing users to see why and how a model produced its result (Adadi & Berrada, 2018; Deeks, 2019). Several interviewees described how human-in-the-loop frameworks (Mazzolin, 2020) can help ensure human involvement in sensitive or critical AI-driven decisions. For example, in a human-in-the-loop system, an AI model might flag potential data anomalies, but a human analyst would review and approve any final changes before they are applied. Human review and involvement can be critical to ensure high-quality output. However, the inclusion of humans-in-the-loop should be balanced against efficiency considerations as the involvement of manual processes performed by humans can offset the benefits of time and effort saved by adopting AI.

One interviewee noted the importance of capacity building as a mitigation strategy to foster a deeper understanding of DQSI practices and responsible data integration when using AI. For example, training programs focused on data literacy in K-12 and higher education or building data capacity at state and local agencies can increase understanding of the importance of responsible use of data. Multiple interviewees identified comprehensive data documentation as another important element of AI ethics. Spelling out limitations related to data collection and identifying potential sources of error introduced through data processing is crucial for ensuring ethical data use in integration activities.

# Findings: AI Tools

The following section presents findings from our review of identified AI tools, informed by the literature scan and expert interviews. The literature scan provided examples of relevant tools for specific data types, such as Geographic Information System (GIS) software intended to support DQSI for geospatial data. Expert interviews did not focus on examples of specific tools, but did provide DQSI use cases to inform which tool capabilities, such as record linkage, to focus on. Each tool was considered based on:

- application areas,
- associated risks and mitigation strategies,
- ownership and licensing,
- scalability,
- costs and computing requirements,
- audience and usability,
- how it evaluates its results, and
- its documentation for users.

We further considered the explainability and interpretability of the algorithms and the tools' capacity for quantifying uncertainty.

This review of AI tools investigates tool capabilities and potential areas of concern, particularly given the needs of federal statistical agencies from some currently available AI tools we identified. The review identifies AI tools that address DQSI needs and pinpoints areas in need of further development across data types. The tool review offered several main takeaways. First, AI offers diverse capabilities ranging from metadata generation to complex geospatial analysis. Second, the adoption of these tools presents important considerations around data privacy, algorithmic bias, and the needs for transparency and for user expertise. Third, the suitability of a tool often depends on the specific data type, desired application, and the technical infrastructure and skills available within an agency.

The tools we reviewed included the following:

- TurboCurator, a tool that uses ChatGPT, an LLM, to generate metadata;
- Open Refine, a data cleaning engine;
- the Record Linkage Toolkit, a Python library for identifying and connecting records across datasets;
- Google Earth Engine, a cloud-based library of satellite imagery and analysis tools; and
- ArcGIS Pro with AI, desktop software for geospatial processing with AI-enabled workflows.

A summary of these tools and high-level findings by the criteria we used to evaluate them are provided in **Table 4**.

**Table 4.** AI Tools Reviewed and Summary of Findings by Evaluation Criteria

| Evaluation Criteria | Turbo Curator | Open Refine | Record Linkage Toolkit | Google Earth Engine | ArcGIS Pro with AI |
|---|---|---|---|---|---|
| **Data Type** | Any metadata (Survey, Administrative, Private Sector, Geospatial) | Any tabular data (Survey, Administrative, Private Sector) | Any tabular data (Survey, Administrative, Private Sector) | Geospatial | Geospatial |
| **Application** | Data curation | Data cleaning | Data integration | Geospatial processing | Geospatial processing |
| **Risks** | Data sharing, disclosure risk | Human and algorithmic error | Algorithmic error | Model bias, gaps in data coverage | Model bias, tool limitations |
| **Licensing** | Open source | Open source | Open source | Free (limits apply) | Proprietary |
| **Scalability** | High (accessed via API endpoint) | Low (local deployment) | Modest (not suited for big data) | High (cloud based; API integration) | Moderate (local deployment) |
| **Costs** | Incurred by service providers | Free for analyst to use locally | Free for analyst to use locally | Free and paid tiers | License required |
| **Audience** | General public | General public | Python users | Technical Users | Employees working with GIS |
| **How It Evaluates Its Results** | Compares metadata quality | Compares data pre- and post-process | Provides record linkage metrics | Accuracy of outputs | Accuracy of outputs |
| **Documentation** | Low (little to no disclosure of how tool works) | High (user manual, community forum available) | High (documentation, release notes available) | High (cookbook, user guides available) | Medium (tool help and guides available; less is available for AI features) |

# Tools

## TurboCurator

The TurboCurator tool, developed by ICPSR at the University of Michigan, uses ChatGPT (an LLM) to assist data depositors in enhancing their metadata before they publish and share their data (ICPSR, 2023). When a user deposits data into a system like ICPSR, TurboCurator analyzes the initial metadata they provide and recommends improvements for titles, descriptions, and keywords. This aims to make the metadata FAIR (Findable, Accessible, Interoperable, Reusable). The tool uses ChatGPT in

conjunction with controlled social science subject terms to generate suggestions for more descriptive keywords, titles, and abstracts in metadata. The controlled terms act as a safeguard to reduce "hallucinations," where the AI system invents information or provides irrelevant content. TurboCurator is designed for broad public use and does not require specialized knowledge of data curation or social science, making it accessible to anyone contributing data. However, users seeking detailed information about the underlying LLM or potential disclosure risks will find that the tool lacks comprehensive documentation, potentially leading to user hesitancy or improper use. Furthermore, the tool transmits user-entered information to a public LLM, which introduces potential disclosure risks for any sensitive information a user might input, although this risk is somewhat mitigated because the transmitted information is metadata rather than the raw data itself. Therefore, while TurboCurator employs AI in the form of an LLM to generate metadata suggestions, it does not offer traditional data cleaning or other DQSI functionalities.

Federal statistical agencies might find a tool like TurboCurator useful for augmenting incomplete or missing metadata across various data types, including administrative, survey, private sector, and geospatial data, regardless of whether the data are being deposited into ICPSR. The tool is distributed under an open-source license, although the code for the underlying model or service is not publicly accessible. TurboCurator is inherently scalable, operating by calling a model endpoint, which allows it to potentially review metadata for entire collections rather than individual datasets. The costs associated with TurboCurator appear to be covered by the service provider, Dataverse, which integrates the tool into its data self-deposit workflows. Its user-friendly design targets a general audience, specifically data depositors who can benefit from AI-driven suggestions to improve their metadata before submission. A "data depositor" in this context is anyone who is submitting data to a repository or system for archiving and sharing.

TurboCurator's functionality is limited to metadata enhancement. It does not directly manipulate the data itself (e.g., perform quality checks, identify and tag variables) nor does it streamline other curation tasks such as disclosure risk review, quality checks, codebook production, or automate repetitive data transformations. The suitability of TurboCurator for use with sensitive data remains unclear, and biases present in the training data of the underlying LLM could be reflected in the generated metadata descriptions, potentially leading to uneven representation across subject areas or the inclusion of incorrect or misleading information, which could ultimately hinder data usability. Additionally, LLMs are known to produce varied outputs even with identical prompts, leading to inconsistencies in the generated metadata. The only way to evaluate the model's output is through a direct comparison of the suggested terms and descriptions to the input, requiring the user to make a subjective quality assessment and decide whether to accept or reject the AI's recommendations. This lack of transparency in the metadata generation process may erode user trust and limit adoption, and the tool's lack of repeatability makes it challenging to objectively evaluate the quality of the generated metadata over time.

## OpenRefine

OpenRefine is a data cleaning tool designed for unstructured or semi-structured data. It provides a graphical user interface that allows analysts to systematically clean data without writing code and supports exploratory data analysis. (OpenRefine, 2022). OpenRefine primarily uses rule-based transformations and pattern recognition algorithms for tasks like clustering and reconciliation. It does not inherently incorporate LLMs or advanced machine learning for its core data cleaning functionalities. As an open-source tool that runs locally on a user's machine, it is well-suited for handling sensitive data, provided that users avoid connecting to external services. The tool offers extensive and regularly updated documentation, along with a moderated user forum where community members can ask and answer questions. The output of OpenRefine is a cleaned and transformed dataset, reflecting the various cleaning operations the user has applied.

Federal statistical agencies might find a tool like OpenRefine valuable because it offers a range of replicable and documented subroutines for addressing messy data. Users can record and reapply the sequence of steps they use to manipulate data, ensuring consistency and transparency in the cleaning process, similar to creating and running macros in Excel. As an open-source project, OpenRefine benefits from community development and maintenance. It is free for analysts to install and use locally; however, its local operation means it is not designed for large-scale, centralized projects. Developed with a general audience in mind, OpenRefine does not require specialized technical expertise. However, users should possess a good understanding of their input data to effectively evaluate and apply cleaning routines.

While OpenRefine is powerful for data cleaning, it is not a comprehensive extract-transform-load or database management tool and lacks the capacity for complex analytical tasks. Although its local operation supports work with sensitive data, caution is needed when using plugins or connecting to external reconciliation services, as this could potentially expose or leak information. Cybersecurity policies within federal agencies may restrict local software installations. It is also important to note that cleaning steps like creating new fields or transforming existing ones can introduce re-identification risks, and the underlying algorithms for clustering and transforming records might introduce biases, particularly affecting sensitive populations. Poorly designed cleaning workflows can also degrade data quality, leading to compounded errors during subsequent analysis.

## Record Linkage Toolkit

The Python Record Linkage Toolkit is a Python library that enables analysts to connect records across different data files using probabilistic and fuzzy matching techniques (de Bruin, 2025). This toolkit primarily employs statistical algorithms for record linkage and does not currently integrate LLMs or advanced machine learning techniques for its core functionalities. It enhances efficiency when dealing with large datasets through methods like blocking (a way to reduce the number of record pairs to compare) and facilitates the standardization of records and the comparison of their similarity based on defined features. The library is well-documented, including change notes to inform users about updates

across different versions. The primary output of this toolkit is a set of identified links between records from different datasets, along with associated weights or probabilities indicating the strength of the match.

Federal statistical agencies might find a tool like the Python Record Linkage Toolkit valuable because it offers an accessible Python implementation of widely used record linkage methods that can be seamlessly integrated into existing Python-based data workflows, providing a solid foundation for data integration efforts. As an open-source tool, the Record Linkage Toolkit makes its underlying algorithms transparent for user inspection. Its free distribution, coupled with example data and code, makes it relatively straightforward for analysts to begin using the toolkit. The toolkit is built upon the Python Pandas library, which has known limitations when handling very large datasets; however, these limitations can potentially be mitigated by using alternative libraries like Polars for data loading. Given its implementation as a Python library, the target audience for this tool is primarily users with technical proficiency and comfort in Python scripting.

In terms of its limitations, the toolkit does not support dimensionality reduction for records with numerous attributes and does not incorporate neural networks or other advanced AI methods to improve record-matching performance. When using the toolkit with data containing sensitive PII, stringent measures must be taken to ensure compliance with relevant privacy laws. Record linkage analyses themselves should be conducted in secure environments to prevent unauthorized access to sensitive data. Furthermore, linking records inherently increases the risk of disclosure. Record linkage algorithms employed may sometimes lead to unequal linkage success rates for certain demographic groups; research has indicated, for instance, that identifiers like surnames associated with specific groups can have varying success rates in entity resolution tasks (Imai et al., 2022). Additionally, users may find it challenging to interpret the linkage quality metrics provided by the library; while these metrics (e.g., accuracy, precision, recall, F1 score) are commonly used to evaluate machine learning models, their interpretation may not be intuitive for a general audience. Guidance may be needed to interpret the linkage quality metrics produced by the toolkit.

## Google Earth Engine

Google Earth Engine (GEE) is a cloud-based platform designed for processing, visualizing, and analyzing large-scale geospatial datasets. (Google, 2025). GEE integrates AI through built-in machine learning and deep learning tools that support both supervised and unsupervised learning tasks. These AI capabilities are applied to diverse DQSI-related operations, such as image classification (e.g., support vector machines, random forests), object detection (e.g., detecting buildings or water bodies), and regression analysis on geospatial data (e.g., predicting biomass or yield). While GEE does not directly employ LLMs, its AI backbone enables automated feature extraction, pattern recognition, and predictive modeling on petabyte-scale Earth observation data.

It provides access to an extensive catalog of data. Users interact with GEE through a browser-based JavaScript editor or a Python API, supported by comprehensive documentation and user resources. GEE offers significant capabilities for federal statistical agencies, particularly in deriving geospatial

features (e.g., identifying land cover types), conducting time series analyses (e.g., tracking changes in vegetation over time), performing change detection (e.g., mapping deforestation), and enabling machine learning applications for classification (e.g., categorizing land use) and regression (e.g., predicting crop yields). Additionally, unsupervised AI methods such as k-means clustering are commonly used to identify latent spatial structures and segment remote sensing data. The output from GEE can take various forms, including processed imagery, statistical maps, time series charts, and the results of machine learning models applied to geospatial data (e.g., classified maps showing different land cover types). GEE is widely applied for environmental monitoring, urban growth analysis, deforestation tracking, and trend forecasting using Earth observation data. The platform handles the complexities of cloud infrastructure, allowing users to process a high volume of data without needing local computing resources. It also supports the integration of user-supplied training datasets and external libraries to further extend its functionality. Technical users can import models into GEE for deployment on satellite imagery at scale.

However, GEE also has limitations that users must consider. Continuous internet access is essential, and usage quotas may restrict high-frequency or computationally intensive operations. The platform's programming environment requires proficiency specifically in JavaScript or Python, which can be a hurdle for users. Privacy concerns can also arise when working with high-resolution imagery in populated or sensitive areas. Biases in satellite data coverage, such as missing time steps or underrepresented regions, can affect the fairness and reliability of analyses. GEE also lacks integrated tools for comprehensive ground truth validation and offers limited transparency regarding the assumptions embedded in its remote sensing models and AI algorithms. From an ethical and operational standpoint, users must exercise caution when using GEE to develop AI-driven models for public or policy-facing outputs. Thorough documentation of data sources, rigorous validation procedures, and a keen awareness of representativeness in model training are crucial for ensuring accountability and mitigating potential risks. While GEE is a powerful tool, its responsible application, especially within federal statistical contexts, demands expertise in both technical and ethical best practices.

## ArcGIS Pro with ArcGIS AI

ArcGIS Pro, a desktop-based GIS platform developed by Esri, offers advanced spatial analysis, 3D visualization, and map production (Esri, 2024). When used with its suite of ArcGIS AI tools, including ArcGIS Image Analyst, GeoAI solutions, and deep learning toolboxes, ArcGIS Pro enables users to apply various Artificial Intelligence and machine learning techniques to enhance geospatial DQSI. These techniques work with a wide range of spatial datasets, such as satellite and aerial imagery; points, lines, and polygons (vector features); and 3D data (point clouds). While Esri provides extensive documentation for its core GIS functionalities, information on the newer AI integration features is comparatively less detailed.

Many federal statistical agencies already utilize ArcGIS Pro for a broad spectrum of geospatial analysis needs. The AI capabilities within ArcGIS Pro support several tasks relevant to DQSI, including object detection (identifying specific features like buildings or roads in imagery), image segmentation (dividing

an image into meaningful regions), classification (categorizing areas in imagery based on land cover), clustering (grouping similar spatial features), and spatial prediction (forecasting values across geographic areas). The output of these AI-powered DQSI tasks can include maps highlighting detected objects, segmented imagery showing different land types, classified maps illustrating land use, clusters of similar geographic features, and predictive maps showing forecasted values. These AI models can be pre-trained deep learning models or custom-trained models developed using labeled imagery. Users can train these models locally or access AI workflows through ArcGIS Notebooks, model builders, or by integrating with platforms like TensorFlow and PyTorch. ArcGIS Pro and its AI tools find applications in areas like land use monitoring, infrastructure mapping, asset management, disaster response, and tracking changes in the natural world, such as forests or coastlines. These tools are particularly valuable for local and regional applications where precise spatial information and controlled data labeling are essential. For instance, deep learning in ArcGIS Pro has been used to automatically identify building footprints from aerial photos, assess damage after natural disasters, and map impermeable surfaces for city planning related to flooding.

However, ArcGIS Pro has limitations. It requires a software license and sufficient local computer resources, especially for intensive deep learning tasks involving large image datasets. Utilizing AI workflows often necessitates a strong understanding of GIS principles, spatial data management, and basic programming skills in Python. Unlike fully cloud-based tools, its ability to handle very large datasets is limited by the capacity of desktop or organizational computer systems. Ethical and operational considerations include the quality of the training data used for AI models and the accuracy of the labeled features, both of which significantly impact model performance. Without proper ground-truth verification (checking AI results against real-world observations) or bias checks, AI models can produce inaccurate classifications, particularly when applied across varied environmental contexts such as dense urban areas or heterogeneous natural landscapes. Furthermore, the proprietary nature of the software can restrict transparency, and reproducing results may be challenging without access to the same licenses, datasets, or specific software configurations.

# Summary

We conducted a review of several existing AI tools for DQSI to identify areas for future development tailored to the needs of federal statistical agencies. This review of AI tools (TurboCurator, OpenRefine, the RecordLinkage Toolkit, GEE, and ArcGIS Pro with ArcGIS AI) reveals a set of diverse tool capabilities for enhancing DQSI. AI offers a broad range of functionalities for DQSI, ranging from automated metadata generation (TurboCurator) to sophisticated geospatial analysis (ArcGIS). However, the review underscores that the adoption of AI tools necessitates careful consideration of challenges related to data privacy, the potential for algorithmic bias, and the need for transparency and user expertise. We found that the selection of AI tools for DQSI is context-dependent, varying by specific data type, intended application, and existing technical infrastructure and skill sets within an agency. The strengths and limitations of each tool across application areas, scalability, usability, and costs further emphasize the importance of context in guiding tool selection. A comprehensive understanding of these factors is crucial for effective and responsible integration of AI tools within the

federal statistical system to avoid redundancy (e.g., developing tools that already exist to address DQSI challenges) and guiding the development of solutions that address recurring DQSI challenges.

# Conclusions and Recommendations

This report explores the potential of AI to enhance data processing, formatting, standardization, and integration within the context of the federal statistical system. Through a comprehensive literature scan, expert interviews, and a review of AI tools, we identify areas where AI could help improve DQSI challenges and enable more effective data integration across a range of data types, including survey data, administrative records, private sector data, and geospatial information. This report also highlights ethical and privacy considerations that must be addressed to ensure responsible and trustworthy AI implementation for federal statistical agencies, including within a future NSDS.

Our analysis reveals several key areas where AI could offer benefits for DQSI. AI can automate data cleaning and validation tasks across diverse data types, addressing errors, inconsistencies, and outliers. LLMs can enhance data discoverability and usability by improving the creation and standardization of metadata. Furthermore, AI can streamline the processing and integration of complex data, such as extracting structured features from geospatial imagery. While specific AI applications may vary across data types, the overarching potential for automation, enhanced metadata, and improved integration is evident.

The federal statistical system presents unique legal and ethical considerations for AI adoption. Paramount among these are the stringent privacy protections mandated by laws such as CIPSEA and Title 13, which necessitate careful attention to data security and the risk of re-identification. Additionally, the potential for algorithmic biases to perpetuate and amplify existing biases requires proactive mitigation strategies. Finally, the need for data to be "AI-ready," with adequate metadata and an understanding of data representativeness and quality, is a critical prerequisite for successful AI implementation.

Our review of existing AI tools, including TurboCurator, OpenRefine, the RecordLinkage Toolkit, Google Earth Engine, and ArcGIS Pro with ArcGIS AI, demonstrates the diverse capabilities currently available for enhancing DQSI. However, the suitability of each tool varies significantly based on the specific DQSI task, data type, technical expertise required, and infrastructure needs. Moreover, issues related to potential biases, disclosure risks, security vulnerabilities, and the transparency of AI processes must be carefully addressed for their responsible use within the federal statistical context.

To ensure the responsible and ethical implementation of AI for statistical applications, the following recommendations should guide the development of an AI toolkit for a future NSDS:

1. AI may be leveraged to automate data cleaning and validation tasks across all data types, such as identifying and correcting errors, inconsistencies, and outliers.
2. AI may simplify the integration of geospatial data into statistical applications by automating and standardizing the extraction of structured features, such as road networks and their attributes from satellite imagery.

3. LLMs can be leveraged to enhance existing data documentation and metadata, thereby improving data discoverability and usability.
4. Algorithmic biases can be addressed by conducting fairness audits, applying bias correction techniques, and training novel AI systems on curated, representative data.
5. Transparency and explainability when using AI can be upheld by disclosing which records and variables have been used, as well as the methods applied in data processing and integration, allowing users to understand potential limitations that may influence model performance.
6. Including humans in the loop when designing AI workflows has the potential to be explored for yielding more accurate results and mitigating model hallucinations. AI tool design should facilitate balanced collaboration between systems and humans.

AI tools developed for use in a future NSDS should consider the recommendations outlined in this report to enhance DQSI while safeguarding privacy, upholding ethical practices, and promoting user trust.

# Acknowledgements

The research team thanks the interviewees for their time to participate in interviews and for their expertise and insights. We also thank our partners at both the National Center for Science and Engineering Statistics at the National Science Foundation and the Bureau of Transportation Statistics for their support for the project, guidance for the work, and for facilitating input from experts at different federal agencies. We additionally thank our team members at NORC at the University of Chicago Lisa Blumerman, Mehmet Celepkolu, Beth Fisher, Bob Goerge, and Martha Stapleton for valuable direction informing this project and report.

# References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alexander, J., & Genadek, K. (2023). Using administrative records to support the linkage of census data: Protocol for building a longitudinal infrastructure of U.S. census records. *International Journal of Population Data Science*, *7*(4). https://doi.org/10.23889/ijpds.v7i4.1764

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). *Americans and Privacy: Concerned, Confused, and Feeling Lack of Control Over Their Personal Information*. Pew Research Center.

Biemer, P., Trewin, D., Bergdahl, H., & Japec, L. (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics*, *30*(3), 381–415. https://doi.org/10.2478/jos-2014-0022

Borgesius, F. Z., Gray, J., & van Eechoud, M. (2016). Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework. *Berkeley Technology Law Journal*. https://doi.org/10.15779/Z389S18

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, *26*(5), 341–358. https://doi.org/10.1080/10106049.2011.562309

Boutet, A. (2024). *Privacy issues in AI and geolocation: From data protection to user awareness* [Thesis, Insa Lyon]. https://hal.science/tel-04909989

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically Weighted Regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(3), 431–443. https://doi.org/10.1111/1467-9884.00145

Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principles of Geographical Information Systems*. OUP Oxford.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). *Expanding Access to Administrative Data for Research in the United States* (SSRN Scholarly Paper No. 1888586). Social Science Research Network. https://doi.org/10.2139/ssrn.1888586

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*(2), 137–161.

Cole, S., Dhaliwal, I., Sautmann, A., & Vilhuber, L. (Eds.). (2020). *Handbook on Using Administrative Data for Research and Evidence-based Policy* (1st ed.). Abdul Latif Jameel Poverty Action Lab. https://doi.org/10.31485/admindatahandbook.1.0

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Curry, M. R. (1997). The Digital Individual and the Private Realm. *Annals of the Association of American Geographers*, *87*(4), 681–699. https://doi.org/10.1111/1467-8306.00073

Cyphers, B., & Gebhart, G. (2019). *Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance*. Electronic Frontier Foundation.

de Bruin, J. (2025). *Python Record Linkage Toolkit Documentation*. https://recordlinkage.readthedocs.io/en/latest/

Deeks, A. (2019). The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review*, *119*(7), 1829–1850.

Einav, L., & Levin, J. (2013). *The Data Revolution and Economic Analysis* (No. Working Paper 19035). NBER.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (SSRN Scholarly Paper No. 3551505). Social Science Research Network. https://doi.org/10.2139/ssrn.3551505

Esri. (2024). AI-Enhanced User Experiences in ArcGIS Pro 3.3. *ArcGIS Blog*. https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/ai-in-arcgis-pro-3-3

Federal Chief Data Officer Council. (2019). *Federal Data Strategy Data Ethics Framework*. President's Management Agenda. https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality* (No. FCSM-20-04). FCSM. https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Future of Privacy Forum. (2017). *Understanding Corporate Data Sharing Decisions*. https://fpf.org/wp-content/uploads/2017/11/FPF_Data_Sharing_Report_FINAL.pdf

Gao, S. (2021). Geospatial Artificial Intelligence (GeoAI). In *Geography* (Vol. 10). Oxford University Press. https://doi.org/10.1093/obo/9780199874002-0228

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Google. (2025). *Google Earth Engine* [API Reference]. Google for Developers. https://developers.google.com/earth-engine/apidocs

Groshen, E. L. (2021). The Future of Official Statistics. *Harvard Data Science Review*, *3*(4). https://doi.org/10.1162/99608f92.591917c6

Groves, R. M., & Harris-Kojetin, B. A. (Eds.) (with Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine). (2017a). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. National Academies Press. https://doi.org/10.17226/24893

Groves, R. M., & Harris-Kojetin, B. A. (Eds.) (with Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine). (2017b). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. National Academies Press. https://doi.org/10.17226/24652

Guptill, S. C., & Morrison, J. L. (2013). *Elements of Spatial Data Quality*. Elsevier.

Haining, R. P. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press.

Harris-Kojetin, B. A., & Citro, C. F. (Eds.) (with Committee on National Statistics, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine). (2021). *Principles and Practices for a Federal Statistical Agency: Seventh Edition*. National Academies Press. https://doi.org/10.17226/25885

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, *4*(2), 2053951717745678. https://doi.org/10.1177/2053951717745678

Harvey, F. (2015). *A Primer of GIS: Fundamental Geographic and Cartographic Concepts*. Guilford Publications.

ICPSR. (2023). *TurboCurator*. Dataverse Administrator About. https://turbocurator.icpsr.umich.edu/tc/adminabout

Imai, K., Olivella, S., & Rosenman, E. T. R. (2022). Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. *Science Advances*, *8*(49), eadc9824. https://doi.org/10.1126/sciadv.adc9824

Interagency Council on Statistical Policy Private Data Steering Group. (2023). *The Use of Private Datasets by Federal Statistical Programs: Extent, Challenges, and Lessons Learned*. Interagency Council on Statistical Policy.

International Monetary Fund. (2020). Scanner Data. In *Consumer Price Index Manual: Concepts and Methods*. https://www.imf.org/en/Data/Statistics/cpi-manual

Iwig, W., Berning, M., Marck, P., & Prell, M. (2013). *Data Quality Assessment Tool for Administrative Data*.

Jarmin, R. S., & O'Hara, A. B. (2016). Big Data and the Transformation of Public Policy Analysis. *Journal of Policy Analysis and Management*, *35*(3), 715–721. https://doi.org/10.1002/pam.21925

Kallberg, J. G., & Udell, G. F. (2003). The value of private sector business credit information sharing: The US case. *Journal of Banking & Finance*, *27*(3), 449–469. https://doi.org/10.1016/S0378-4266(02)00387-4

Laaksonen, S. (2018). *Survey Methodology and Missing Data*. Springer International Publishing. https://doi.org/10.1007/978-3-319-79011-4

Landau, S., Dempsey, J. X., Kamar, E., Bellovin, S. M., & Pool, R. (2024). *Challenging the Machine: Contestability in Government AI Systems* (No. arXiv:2406.10430). arXiv. https://doi.org/10.48550/arXiv.2406.10430

Landefeld, S. (2014). *Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues*. International Conference on Big Data for Official Statistics, Beijing, China.

Lane, J. (2018). Building an Infrastructure to Support the Use of Government Administrative Data for Program Performance and Social Science Research. *The ANNALS of the American Academy of Political and Social Science*, *675*(1), 240–252. https://doi.org/10.1177/0002716217746652

Lohr, S. L., & Raghunathan, T. E. (2017). Combining Survey Data with Other Data Sources. *Statistical Science*, *32*(2). https://doi.org/10.1214/16-STS584

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic Information Science and Systems*. John Wiley & Sons.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

MacEachren, A. M. (2004). *How Maps Work: Representation, Visualization, and Design*. Guilford Press.

Mazzolin, R. (2020). *Artificial Intelligence and Keeping Humans "in the Loop"* (Modern Conflict and Artificial Intelligence, pp. 48–54). Centre for International Governance Innovation. https://www.jstor.org/stable/resrep27510.10

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Muth, M. K., Sweitzer, M., Brown, D., Capogrossi, K., Karns, S., Levin, D., Okrent, A., Siegel, P., & Zhen, C. (2016). *Understanding IRI Household-Based and Store-Based Scanner Data* (No. Technical Bulletin 1942). United States Department of Agriculture Economic Research Service.

National Academies of Sciences, Engineering, and Medicine. (2023). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26804

National Academies of Sciences, Engineering, and Medicine. (2024). *Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data* (J. P. Reiter & J. Park, Eds.; p. 27335). National Academies Press. https://doi.org/10.17226/27335

National Artificial Intelligence Initiative Act of 2020, No. H.R.6216 (2020). https://www.congress.gov/bill/116th-congress/house-bill/6216

National Research Council. (2007). *Successful Response Starts with a Map: Improving Geospatial Support for Disaster Management* (p. 11793). National Academies Press. https://doi.org/10.17226/11793

National Security Council. (2024). *Framework to Advance AI Governance and Risk Management in National Security*. White House. https://ai.gov/wp-content/uploads/2024/10/NSM-Framework-to-Advance-AI-Governance-and-Risk-Management-in-National-Security.pdf

O'Hara, A., & Medalia, C. (2018). Data Sharing in the Federal Statistical System: Impediments and Possibilities. *The Annals of the American Academy of Political and Social Science*, *675*, 138–150. https://doi.org/10.1177/0002716217740863

O'Hara, A., & Rhodes, R. (2023). The Federal Agencies' Hidden Efforts to Produce Equitable Data. *American Journal of Public Health*, *113*(12), 1278–1282. https://doi.org/10.2105/AJPH.2023.307465

OpenRefine. (2022). *OpenRefine User Manual*. https://openrefine.org/docs

O'Toole, K., Turbes, C., & Freeman, A. (2024). *Data Policy in the Age of AI: A Guide to Using Data for Artificial Intelligence*. Data Foundation.

Penner, A. M., & Dodge, K. A. (2019). Using Administrative Data for Social Science and Policy. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *5*(2), 1–18. https://doi.org/10.7758/RSF.2019.5.2.01

Prell, M. (2019). *Transparent Reporting for Integrated Data Quality: Practices of Seven Federal Statistical Agencies* (No. FCSM-19-01).

Prell, M., Bradsher-Fredrick, H., Comisarow, C., Cornman, S., & United States. Federal Committee on Statistical Methodology. (2009). *Profiles in Success of Statistical Uses of Administrative Data*. https://doi.org/10.21949/1529872

Prem, E. (2024). Approaches to Ethical AI. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to Digital Humanism: A Textbook* (pp. 225–239). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45304-5_15

Ranbaduge, T., Vatsalan, D., & Ding, M. (2024). Privacy-Preserving Deep Learning Based Record Linkage. *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 6839–6850. https://doi.org/10.1109/TKDE.2023.3342757

Reibel, M. (2007). Geographic Information Systems and Spatial Data Processing in Demography: A Review. *Population Research and Policy Review*, *26*(5–6), 601–618. https://doi.org/10.1007/s11113-007-9046-5

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: A systematic review. *NPJ Digital Medicine*, *7*(1), 203. https://doi.org/10.1038/s41746-024-01196-4

Seeskin, Z. H. (2018). Evaluating the utility of a commercial data source for estimating property tax amounts. *Statistical Journal of the IAOS*, *34*(4), 543-551.

Seeskin, Z. H., LeClere, F., Ahn, J., & Williams, J. A. (2018). Uses of Alternative Data Sources for Public Health Statistics and Policymaking: Challenges and Opportunities. G*overnment Statistics Section, JSM Proceedings*, 1822-1861.

Seeskin, Z. H., Ugarte, G., & Datta, A. R. (2019). Constructing a toolkit to evaluate quality of state and local administrative data. *International Journal of Population Data Science*, *4*(1). https://doi.org/10.23889/ijpds.v4i1.937

Selmy, S. A. H., E. Kucher, D., Yang, Y., & Jesús García-Navarro, F. (2024). Geospatial Data: Acquisition, Applications, and Challenges. In R. M. Abdalla (Ed.), *Exploring Remote Sensing—Methods and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.1006635

U.S. Bureau of Labor Statistics. (2024). *Consumer Price Index*. Division of Consumer Prices and Price Indexes.

U.S. Census Bureau. (2022). *TIGER/Line Shapefiles* [Dataset]. U.S. Census Bureau. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

Vale, S. (2011). *Using Administrative and Secondary Sources for Official Statistics*. United Nations Economic Commission for Europe.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Vought, R. T. (2019, April 24). *Improving Implementation of the Information Quality Act*. Office of Management and Budget. https://www.cdo.gov/assets/documents/OMB-Improving-Implementation-of-Info-Quality-Act-M-19-15.pdf

Yarkoni, T., Eckles, D., Heathers, J. A. J., Levenstein, M. C., Smaldino, P. E., & Lane, J. (2021). Enhancing and Accelerating Social Science via Automation: Challenges and Opportunities. *Harvard Data Science Review*, *3*(2). https://doi.org/10.1162/99608f92.df2262f5

Yeung, D., Khan, I., Kalra, N., & Osoba, O. A. (2021). *Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications*. RAND Corporation. https://www.jstor.org/stable/resrep29576

Zhu, X., Cai, F., Tian, J., & Williams, T. (2018). Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions. *Remote Sensing*, *10*(4), 527. https://doi.org/10.3390/rs10040527

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

# Appendix

## Appendix 1. Search Terms

**Appendix Table 1.** Search Terms Used for Literature Scan

| Topic | Purpose | Query |
|---|---|---|
| **Taxonomies** | Identify existing data type taxonomies | "federal statistical" + (data structure \| type \| taxonomy \| ontology \| typology \| collection method \| source \| categories) |
| **Survey Data** | Identify literature about survey data definitions and uses | survey data \| survey methodology \| data quality \| probability sample |
| **Administrative Data** | Identify literature about administrative data definitions and uses | administrative data \| administrative records \| program data |
| **Private Sector Data** | Identify literature about private sector data definitions and uses | private sector data \| private data \| corporate data \| third party data \| vendor data |
| **Geospatial Data** | Identify literature about geospatial data definitions and uses | geospatial data \| spatial data \| GIS \| remote sensing \| spatial analysis \| location data \| satellite imagery |
| **Data Uses and Needs** | Identify literature about each data type's needs | [Data Type] + (quality \| integration \| standardization \| processing \| tools \| Artificial Intelligence \| ethics \| privacy \| compliance \| risks) |

# Appendix 2. AI Use for Research for this Report

This section of the appendix describes how AI was used in preparation of this report. The following sections indicate the models and tools used, along with the prompts issued to the models to conduct research in preparation of the report.

## Models

- ChatGPT 4o: Offered recommendations of datasets, documents, and software documentation to support the literature scan and review of AI tools; specific prompts used are included below
- KeyBERT: Provided context-aware keyword extraction
- Zoom AI: Helped summarize key points from interviews

## ChatGPT 4o Prompts

- Prompt 1: Identify and list primary datasets, geospatial applications, and operational use cases in sectors such as agriculture (e.g., NASS Cropland Data Layer), transportation (e.g., BTS spatial transportation models), and demographic analysis (e.g., Census Bureau geodatabases and urban planning systems).
- Prompt 2: Extract core recommendations, frameworks, and challenges related to geospatial data governance, data quality, and integration, with particular attention to foundational insights provided in the National Research Council (2007) report *"Successful Response Starts with a Map"*.
- Prompt 3: Scan official government open documents, technical forums, and major geospatial software repositories to extract references to AI-enabled geospatial tools.
- Prompt 4: Identify and systematically categorize AI-driven geospatial tools mentioned across U.S. government reports, white papers, GitHub repositories, industry partner websites, and technical discussion forums (e.g., GIS Stack Exchange, Esri Community).

# Appendix 3. Interview Questions

This section of the appendix includes the questions used in interviews with federal agency staff, subject matter experts, and data privacy and ethics experts.

## Federal Agency Staff

- Please describe your professional background and your role with producing or using data at your agency.
- What types of data do you work with at your agency?
  - What are the common features of these data?
  - What are the common uses and opportunities for these data sources?
  - What are common data quality challenges with the data you encounter?
  - How do you address these data quality challenges?
- Do you or your team conduct data integration and/or use data integrated from multiple sources?
  - If so, what are the common uses and/or opportunities for using integrated data?
  - What challenges do you encounter with conducting data integration and/or with using integrated data?
  - Specifically, what challenges do you find with standardizing different data sources for purposes of data integration?
  - How does your team mitigate such challenges regarding data standardization and integration?
  - What kinds of data integration and standardization applications would you be useful to you or to your team?
- What data integration and standardization applications would you like to see as part of a National Secure Data Service?
- How have you or your team explored using AI or related methods to enhance data quality, standardization, and integration for the data types you commonly use?
  - If AI is used or has been considered for use, what potential applications and benefits have you found from the use of AI or related methods for these purposes?
  - What challenges or concerns have you found from using AI for these purposes? These could be challenges related to privacy, ethics, or other concerns.

- Please tell us anything else about your experiences with data quality, standardization, and integration at your agency that you think is important for us to know for this project.

## Data Privacy and Ethics Experts

- Please describe your professional background, your role at your organization/institution, and your areas of expertise including with different data types.
- What are common privacy and ethical challenges that are important to consider when integrating data from multiple data sources?
  - What are your recommendations for assessing and addressing these challenges?
  - What are common privacy and ethical challenges regarding uses of AI or related methods to enhance data quality, standardization, and integration?
- How does the use of AI impact data privacy concerns, and what are options for mitigating these concerns?
  - How do challenges with the interpretability of AI models impact recommendations for using AI for these purposes?
  - What are the challenges regarding introduction of bias from using AI for these purposes, and how do you recommend mitigating the risk for bias to ensure fairness for policy conclusions?
- Are there other privacy or ethical challenges with using AI you would like to highlight for this project?
- What privacy and ethical considerations do you recommend for data integration and standardization applications at a National Secure Data Service to address?
- Please tell us anything else about privacy and ethical considerations for uses of AI to enhance data quality, standardization, and integration that you think is important for us to know for this project.

## Subject Matter Experts

- Please describe your professional background, your role at your organization/institution, and your areas of expertise.
- What types of data do you commonly work with?
  - What are the common features of these data?
  - What are the common uses and opportunities for these data sources?
  - What are common data quality challenges with the data you encounter?
  - What are your recommendations for addressing these data quality challenges?
- How does integration of data from multiple sources play a role in your work?
  - What opportunities for uses of integrated data have you found?
  - What are common challenges with conducting data integration and/or with using integrated data?
  - Specifically, what are common challenges with standardizing different data sources for purposes of data integration?
  - How do you recommend mitigating such challenges regarding data standardization and integration?
- What kinds of data integration and standardization applications would you recommend for a National Secure Data Service to provide?
- Have you found opportunities for AI or related methods to enhance data quality, standardization, and integration for the data types you commonly use?
  - If so, what potential uses and benefits have you found from the use of AI or related methods for these purposes?
  - What challenges or concerns have you found from using AI for these purposes? Are the challenges related to privacy, ethics, and/or other concerns?
- Please tell us anything else about your experiences with data quality, standardization, and integration that you think is important for us to know for this project.