Data Quality Threat Mitigation Strategies for Multi-Sourced Linked Data*

Ali Obaidi[a] and Adrienne Chen-Young[b]
Data Environment & Engineering Dept., The MITRE Corporation, 7515 Colshire Drive, McLean, VA, 22182, U.S.A.

Keywords: Data Quality, FCSM Framework, Data Integrity, Data Quality Mitigation, Data Driven Decision Making, Data Quality Strategies, Data Quality Threats, Data Quality Enhancement.

Abstract: Federal agencies link data from multiple sources to generate statistical data products essential to informing policy and decision making (National Academies, 2017). The ability to integrate and link data is accompanied by the challenge of harmonizing heterogenous data, disambiguating similar data, and ensuring that the quality of data from all sources can be reconciled at levels that provide value and utility commensurate with the integration effort. Given the significant resources and effort needed to consistently maintain high quality, multi-sourced, linked data in a government ecosystem, this paper proposes steps that can be taken to mitigate threats to data quality at the earliest stage of the statistical analysis data lifecycle: data collection. This paper examines the threats to data quality that are identified in the Federal Committee on Statistical Methodology's (FCSM) Data Quality Framework (Dworak-Fisher, 2020), utilizes the U.S. Geological Survey's (USGS) Science Data Lifecycle Model (SDLM) (Faundeen, 2013) to isolate data quality threats that occur before integration processing, and presents mitigation strategies that can be taken to safeguard the utility, objectivity, and integrity of multi-sourced statistical data products.

1 INTRODUCTION

The Federal Committee on Statistical Methodology's (FCSM) Data Quality Framework is a comprehensive structure developed through a rigorous collaborative process, by a cross-agency government team, to support federal agencies in identifying and reporting data quality. It is designed to apply to a wide range of data, including statistical data collected through surveys and censuses, nonstatistical data such as administrative records, and integrated data products. As illustrated in Figure 1, the framework is built on three broad components, or domains - utility, objectivity, and integrity - encompassing specific dimensions that represent areas for evaluating data quality. These dimensions include relevance, accessibility, timeliness, punctuality, granularity, accuracy and reliability, coherence, scientific integrity, credibility, computer and physical security, and confidentiality.

FCSM's Data Quality Framework offers a standard for statistical agencies to develop granular considerations and measures for establishing data quality expectations according to the characteristics defined (Dworak-Fisher, 2020, Parker, 2024), for each dimension of the framework.

FCSM compiled a set of use cases (Mirel, 2023) to illustrate the framework's application across different scenarios from a variety of agencies and demonstrate the use of the framework in addressing threats to quality. While several use cases explore how to apply the framework to assess the quality of data before it is used to make decisions, there are limited case studies proposing strategies specifically aimed at mitigating threats to data quality during the collection stage of the data lifecycle.

[a] https://orcid.org/0000-0001-5175-955X
[b] https://orcid.org/0009-0008-1270-0748

[1] Approved for Public Release; Distribution Unlimited. Public Release Case Number 25-0152

Obaidi, A., Chen-Young and A.
Data Quality Threat Mitigation Strategies for Multi-Sourced Linked Data.
DOI: 10.5220/0013462900003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 72-81
ISBN: 978-989-758-758-0; ISSN: 2184-285X

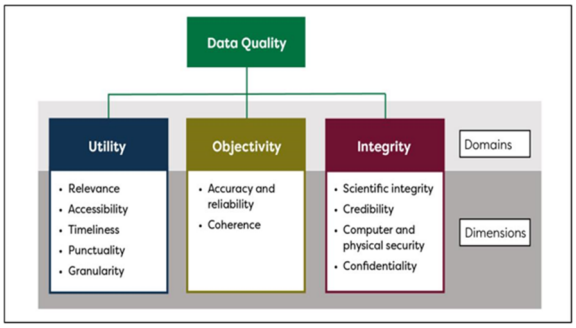

Figure 1: FCSM Data Quality Framework (Source: FCSM-20-04 A Framework for Data Quality).

2 THESIS

Data quality is not an absolute measure but is relative to the context in which the data is being used (Serra, 2022). Data is generally considered of high quality if it meets standards established for intended operational uses, decision making, product generation, and strategic planning (Wang, 1995). In a linked data ecosystem, each data source inherits the data quality standards established for the needs and requirements of its originating system (Radulovic, 2015). It is essential to develop an approach that establishes data quality expectations for a linked environment that at a minimum preserves the original integrity of the data from its source.

As a recognized standard for federal statistical agencies, the FCSM Data Quality Framework is well-suited for devising strategies to mitigate threats to the quality of integrated data used in government decision making.

This study proposes mitigation strategies to address data quality threats identified in dimensions of the framework that specifically pertain to the collection stage of the data lifecycle. By deliberately incorporating steps to safeguard data quality at the beginning of federal statistical data processing, agencies can establish an improved measure of confidence that their data curation investment will generate data products that are useful, credible, reliable, and able to support complex decision-making.

3 METHODS

The U.S. Geological Survey (USGS) publishes documentation describing their mature data management practices based on a Science Data Lifecycle Model (SDLM) (Faundeen,2013) that illustrates the flow of data "from conception through preservation and sharing". This model, illustrated in Table 1, helps researchers and data practitioners ensure that data products are well described, preserved, accessible, and suitable for reuse, while also serving as a framework to evaluate and enhance data management policies and practices, and to identify the need for new tools and standards.

Table 1: USGS SDLM Elements (Source: USGS).

Primary Model Element	Description
Plan	Assists scientists in considering all activities related to handling the project's data assets, from inception to publication and archiving. Involves evaluating, addressing, and documenting all elements of the model.
Acquire	Involves activities through which new or existing data are collected, generated, or evaluated for re-use. Emphasizes considering relevant USGS policies and best practices to maintain data provenance and integrity.
Process	Includes activities associated with preparing new or previously collected data inputs. Involves defining data elements, integrating datasets, and applying calibrations to prepare data for analysis.
Analyze	Covers activities related to exploring and interpreting processed data, where hypotheses are tested, discoveries are made, and conclusions are drawn. Includes summarization, graphing, statistical analysis, and modelling.
Preserve	Involves storing data for long-term use and accessibility. Emphasizes planning for the long-term preservation of data, metadata, ancillary products, and documentation to ensure availability and re-use.
Publish/ Share	Combines traditional peer-reviewed publication with data distribution through various platforms. Highlights the importance of publishing data and information as critical components of the USGS mission and Federal directives.

Given the maturity of USGS's data management practices, the public availability of the details of their data lifecycle, and the scientific applicability of model elements, USGS SDLM was selected as the basis for aligning identified FCSM data quality threats to the data collection stage of the statistical data product lifecycle.

The research team leveraged their data engineering expertise and lessons learned from architecting data integration solutions for several federal agencies to perform a semantic examination of each FCSM data

quality threat. Each threat was then mapped to the first stage in the USGS SDLM where the threat could potentially occur to isolate threats specific to the collection of data (SDLM element "Acquire"). A justification statement was crafted to articulate the rationale behind each mapping. Mitigation strategies were then formulated for each threat mapped to SDLM element "Acquire". Mitigation strategies were tailored to address nuances of each threat that could be effectively managed at the point where multi-sourced data are collected in preparation for integration and further analytic processing.

4 RESULTS

The methodology described in the previous section was used to conduct a comprehensive mapping of the FCSM data quality threats to elements in the USGS SDLM. The results of the mapping process are presented in Table 2. The columns in Table 2 are as follows:

1. **FCSM Data Quality Domain**: The data quality Domain name from the FCSM framework.
2. **FCSM Data Quality Dimension:** The data quality Dimension from the FCSM framework.
3. **FCSM Identified Threats**: Each threat associated each data quality Dimension as identified in the FCSM framework.
4. **Mapping to USGS SDLM Element**: The mapping of each FCSM threat to the first stage in the USGS SDLM where the threat could potentially occur. Threats mapped to the "Acquire" element are highlighted.
5. **Justification for Mapping to USGS SDLM Element**: For each mapping decision, the justification statement explains the rationale behind the association. These justifications are based on the authors' semantic understanding of the FCSM threat descriptions, and their data engineering expertise and insights gained from architecting data integration solutions.

The mapping process involved a semantic study of each data quality threat identified in the FCSM framework and associating each threat to a stage of the USGS SDLM to generate an alignment reflective of the potential impact. For each mapped threat, a justification is provided to explain the rationale behind the association. The justification highlights aspects of the data lifecycle that are vulnerable to the identified threat and underscores the importance of addressing those vulnerabilities for high data quality standards.

Table 2 serves as a comprehensive reference for understanding how each FCSM data quality threat is associated to the USGS SDLM, ensuring that potential issues are identified for mitigation at the data collection stage.

Table 2: Mapping of FCSM Data Quality Threats to USGS SDLM Elements.

FCSM Data Quality Domain	FCSM Data Quality Dimension	FCSM Identified Threat	Mapping to USGS SDLM Element	Justification for Mapping to USGS SDLM Element
Utility	Relevance	Difficulties in understanding and aligning user needs	Plan	User needs are typically captured during requirements gathering
Utility	Relevance	Availability of related data products	Plan	Due diligence regarding the development of data products typically performed during conceptualization
Utility	Relevance	Negative perceptions of users	Publish/Share	User feedback typically solicited either during testing or after publication
Utility	Accessibility	Costs to access data	Plan	Budgets typically established during conceptualization
Utility	Accessibility	Use of disclosure limitation methods	Publish/Share	Disclosure techniques typically employed at point of dissemination
Utility	Accessibility	Costs to create effective documentation	Plan	Budgets typically established during conceptualization
Utility	Accessibility	Confusion with other data products	Acquire	Assignment of discovery metadata typically applied during data collection
Utility	Timeliness	Significant lags of input data	Acquire	Arrival delays between data sets from multiple sources will impact intake[2] processing
Utility	Timeliness	Processing time needed for source data	Acquire	Appropriate curation typically applied before mission processing
Utility	Timeliness	Statistical and methodological rigor	Process	Data product creation subject to intentional processing rigor
Utility	Timeliness	Production of effective documentation	Publish/Share	Documentation for data product dissemination

[2] Intake processing is considered equivalent to data collection under this research effort

Table 2: Mapping of FCSM Data Quality Threats to USGS SDLM Elements (cont.).

FCSM Data Quality Domain	FCSM Data Quality Dimension	FCSM Identified Threat	Mapping to USGS SDLM Element	Justification for Mapping to USGS SDLM Element
Utility	Punctuality	Low response and participation rates	Acquire	Impact to rate of data collection
Utility	Punctuality	External events	Acquire	Impact to the availability of source data
Utility	Punctuality	Changes in secondary-use source data	Acquire	Impact to the availability of source data
Utility	Punctuality	hanges in agency priorities	Plan	Agency factors around data typically considered during planning
Utility	Granularity	Small sample size	Acquire	Source data typically evaluated for extent to which information is welltargeted for identified and anticipated needs before mission processing
Utility	Granularity	Unavailable data	Acquire	Granularity of data typically accessed before mission processing
Utility	Granularity	Confidentiality protections	Publish/Share	Disclosure techniques typically employed at point of dissemination
Objectivity	Accuracy and Reliability	Sampling error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Nonresponse error and missing data	Acquire	Identification of missing data at point of data collection preferred to identification during mission processing
Objectivity	Accuracy and Reliability	Coverage error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Measurement error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Linkage error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Harmonization error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Modelling error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Processing error	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Accuracy and Reliability	Additional threats involving geographic data	Process	Processing methodology typically designed to reveal characteristic discrepancies
Objectivity	Coherence	Multiple sources of data and definitions	Acquire	Data lineage typically established at point of data collection
Objectivity	Coherence	Changes in data over time	Process	Longitudinal considerations typically included in processing methodology
Objectivity	Coherence	Changes in statistical and processing methods	Process	Longitudinal considerations typically included in processing methodology
Objectivity	Coherence	Misalignment	Acquire	Assignment of meaningful use metadata typically applied during data collection
Integrity	Scientific Integrity	Political interference	Plan	External influences on data typically considered during conceptualization
Integrity	Scientific Integrity	Obsolescence	Process	Evolving and modern statistical methods typically accounted for during process management and maintenance
Integrity	Scientific Integrity	Computer generated data	Plan	Credibility of sources typically confirmed before data collection
Integrity	Credibility	Dissemination of inaccurate data products	Publish/Share	Processing methodology and errata typically provided on dissemination
Integrity	Credibility	Competing data sources and methods	Acquire	Consistent, attributed curation of appropriate data use applied during data collection
Integrity	Credibility	Political interference	Plan	External influences on data typically considered during conceptualization
Integrity	Credibility	Obsolescence	Process	Evolving and modern statistical methods typically accounted for during process management and maintenance
Integrity	Computer and Physical Security	Supply chain risk	Acquire	Bad actors or breeches in processing may corrupt collected data
Integrity	Computer and Physical Security	Human error	Plan	Minimization of threat typically addressed before data collection
Integrity	Computer and Physical Security	Insider threat	Plan	Minimization of threat typically addressed before data collection
Integrity	Computer and Physical Security	External threats	Plan	Vulnerabilities typically considered before data collection
Integrity	Confidentiality	Granularity	Publish/Share	Granularity of data typically accessed before mission processing
Integrity	Confidentiality	Large number of data elements in microdata	Publish/Share	Disclosure techniques typically employed at point of dissemination

Threats that were mapped to the USGS SDLM "Acquire" element were isolated and comprehensive mitigation strategies devised. Table 3 below presents the culmination of the research outlined in this paper: mitigation strategies for safeguarding data quality at the data collection stage of statistical analysis.

Table 3: Data Collection Threat Mitigation Strategies.

#	Data Collection Threat Mitigation Strategies (Content developed during research)	FCSM Identified Threat (Source: FCSM)
1	Source data must be accompanied with metadata describing the characteristics of the data to promote its correct discovery and interpretation. Metadata is typically provided in data dictionaries and/or file layouts to describe the structure, content, meaning and lineage of data elements including data types, allowable values, measurement units, constraints/rules for use, and any relationships between or among data elements. This helps users understand data values, allowable values, measurement units, constraints and reduction in confusion. Request that data providers include metadata specifications for data ingest. Applying a taxonomy to each data element at collection is an effective strategy to ensure that incoming data is classified and organized by consistent characteristics relevant to the mission thus avoiding confusion from external definitions and among datasets.	Confusion with other data products
2	It is crucial that intake processing be designed to accommodate varying arrival times of data. Message queuing systems, batch aggregation, and time-based triggered processing are methodologies that can be employed to manage incoming data with asynchronous data arrivals and ensure that data processing occurs at regular intervals.	Significant lags of input data
3	Resilient intake processing must be designed to handle flow delays ensuring the system functionality and efficiency even under unexpected delays. Data pipelines must be designed to adapt to varying input timing and volume with fail safe processing mechanisms that prioritize the confirmation of data quality at collection and preventing downstream issues and maintaining data integrity.	Processing time needed for source data
4	Intake processing must be designed to accommodate varying volumes of incoming data. Message queuing systems, batch aggregation, and time-based triggered processing are methodologies that can be employed to manage fluctuations in incoming data by providing flexibility in handling data as it arrives. Targeted outreach to data providers and incentives for participation may be offered to increase participation rates, fostering strong relationships and encourage consistent data submissions.	Low response and participation rates
5	Resilient intake processing must be designed to account for unforeseen events to ensure that the system can continue to function effectively even when unexpected external events occur. Data pipelines must be designed to adapt to varying input timing and volume with fail safe processing mechanisms that prioritize the confirmation of data quality at collection which is essential for handling the irregularities caused by external events. Contingency plans could be explored to diversify data sources and provide additional security and flexibility.	External events
6	Intake processing must be designed to accommodate varying volumes of incoming data. Message queuing systems, batch aggregation, and time-based triggered processing are methodologies that can be employed to manage fluctuations in incoming data. The metadata accompanying incoming data must be verified at every intake instance to ensure that an accounting of possible changes is considered for routine sources as well as new sources.	Changes in secondary-use source data

Table 3: Data Collection Threat Mitigation Strategies (cont.).

#	Data Collection Threat Mitigation Strategies (Content developed during research)	FCSM Identified Threat (Source: FCSM)
7	Determine the appropriate sample size needed to achieve the desired level of precision in data products which involves understanding level of confidence and acceptable margin of errors to ensure statistically significant results. Conduct statistical power analysis and consider the variability within the population to minimize the risk of type II errors. Leverage techniques such as oversampling for underrepresented groups to help ensure that the data collected is both broad and comprehensive making the data more equitable and representative of the entire population.	Small sample size
8	Design data collection processes with flexibility in mind, allowing for the capture of additional data points that may be required for future analyses which will help accommodating changes in data requirements without significant disruptions. Link data from multiple sources to enhance the granularity and richness of the overall dataset. Profile data to discover gaps and add steps to enrich the data from other sources during intake. Machine learning models may also be used to infer missing details and allowing for the estimation of missing data within a reasonable degree of accuracy. Establishing strong data governance practices can also help ensure that data collection and integration processes are well-managed and transparent.	Unavailable data
9	Increase response rates through follow-ups to data providers and offer incentives to participants. Apply mission imputation techniques to handle missing data to mitigate the impact of nonresponse. Conduct nonresponse bias analysis to remove data quality impacts due to bias. Profile data to discover gaps and add steps to enrich the data from other sources during intake. Machine learning models may also be used to infer missing details and allow for the estimation of missing data within a reasonable degree of accuracy.	Nonresponse error and missing data
10	Apply a taxonomy to each data element at collection to ensure that incoming data is classified and organized by consistent characteristics relevant to the mission thus avoiding confusion across definitions from multiple sources.	Multiple sources of data and definitions
11	Clearly define the intended use of data during the acquisition planning process. Conduct a needs assessment to ensure that data to be collected aligns with the intended analysis. When repurposing data, conduct a gap analysis to identify and address potential misalignments.	Misalignment
12	Source data must be supplied with metadata describing the characteristics and intended use of the data. Metadata is typically provided in data dictionaries and/or file layouts to describe the structure, content, meaning, processing and lineage of data elements including data types, allowable values, measurement units, constraints/rules for use, and any relationships between or among data elements. Applying a taxonomy to each data element at collection ensures that incoming data is classified and organized by consistent characteristics relevant to the mission thus avoiding confusion among competing datasets.	Competing data sources and methods
13	Data providers must be vetted against their ability to deliver reliable and trustworthy data, and data transmission processing must be protected against threat vectors that may seek to alter data and introduce malicious content. Data infrastructure must adhere to information security management mandates, policies, and procedures to protect against unauthorized access. Data transmissions must be encrypted, and service level agreements used to establish rate limits and transmission media types to help manage expectations and ensure defined responsibilities and accountability among all parties. Network protocols must be established to authenticate connections and enumerations such as checksums used to detect changes during transmission.	Supply chain risk

5 DISCUSSION

As agencies increasingly rely on data to drive decision-making and innovation, understanding and improving data quality at the earliest stage of the statistical analysis data lifecycle is critical. With the comprehensive set of threat mitigation strategies presented in Table 3, the next step involves organizing these strategies into cohesive cross-cutting themes. This approach not only streamlines the implementation process but also enhances the effectiveness of each strategy by highlighting their interconnectedness and collective impact. By clustering these strategies into key thematic areas, organizations can more effectively address the multifaceted challenges of data quality, ensuring that their data assets remain robust, reliable, and ready to support strategic objectives. The following crosscutting themes provide a structured framework for understanding and applying these threat mitigation strategies:

Metadata and Taxonomy (Strategy #1, 12): Metadata and taxonomies are essential components in ensuring data quality, as they provide structure, context, and consistency to data management processes. Metadata provides detailed descriptions of data elements, including data types, allowable values, measurement units, and constraints related to the nature and limitations of the data, thereby promoting accurate interpretation and use. Metadata enhances clarity, facilitating the effective linkage and analysis of data from multiple sources. Source data must be accompanied by comprehensive metadata, typically presented in data dictionaries, to describe data characteristics. Implementing a taxonomy at the collection stage of the data lifecycle ensures consistent classification and organization, thereby preventing confusion arising from varying definitions. Taxonomies align data classification with specific organizational goals or missions, ensuring that data is relevant and useful for intended analyses or decision-making processes.

Data Pipeline Design and Intake Processing (Strategy #2,3,4,5,6): The design of the data pipeline and intake processing is critical in understanding and improving data quality by establishing robust mechanisms for data collection and validation. Effective pipelines must incorporate mechanisms for detecting, logging, and handling errors enabling prompt identification and resolution of issues. To manage varying data arrival times and volumes, methodologies such as message queuing, batch aggregation, and time-based processing should be applied to intake activities. Pipelines must also include steps for transforming and cleaning data, such as normalizing formats and terminology, removing duplicates, matching entities, and enriching data with additional information. These processes enhance data quality by ensuring uniformity and completeness. Verifying metadata during intake helps to ensure that data is correctly described and categorized, a crucial step towards maintaining data quality and ensuring accurate understanding and processing. Resilient intake processes are designed to handle flow delays and unforeseen events, with fail-safe mechanisms that prioritize data quality.

Sample Size and Data Collection (Strategy #4, 7, 8): The sample size in data collection plays a significant role in determining the quality of the data and the reliability of the insights derived from it. A sufficiently large and well-chosen sample size ensures an accurate reflection of the characteristics of the entire population, allowing valid inferences and generalizations to be applied to the broader population. A larger sample size facilitates better detection and understanding of variability within the population, aiding in the identification of trends, patterns, and outliers. A well-sized sample can help mitigate biases that may arise from nonresponse or other sampling issues, thereby minimizing their impact on data quality. Larger sample sizes provide greater confidence in the results and conclusions drawn from the data. A well-determined sample size enhances representativeness, precision, and statistical power, ensuring that the data collected is robust, reliable, and suitable for accurate analysis and decision-making.

Data Enrichment (Strategy #8, 9): Data enrichment significantly enhances data quality by adding relevant information or context to existing datasets. The primary purpose of enrichment is to add value and depth, creating a more comprehensive view of the data, reducing gaps that could affect analysis. Strategies for effective data enrichment include integrating external datasets, performing data cleaning and normalization, and using data augmentation techniques. Enrichment processes often involve verifying and updating existing data with more current or corrected information. This process can make datasets more relevant to specific analyses or business needs, ensuring that data is aligned with mission objectives to enhance its utility

and applicability. Enriched data supports more sophisticated analyses by providing additional variables and dimensions to explore, potentially leading to the discovery of new patterns, trends, and relationships that enhance the overall quality of insights derived from the data. Considerations for enrichment include ensuring that these processes do not introduce errors or biases. The added depth and breadth of information in enriched data provide a stronger foundation for strategic planning and decision-making.

Data Use and Alignment (Strategy #11): Aligning data with its intended use is a crucial aspect of ensuring data quality. By focusing on the intended use, data collection efforts are tailored to gather only the most relevant data, thereby increasing its utility and effectiveness. Strategies to achieve this include establishing clear data usage policies, regularly reviewing data alignment with objectives, and involving stakeholders in data use planning. This alignment helps streamline data collection and processing efforts, reduces unnecessary data handling, minimizes resource expenditure, and focuses efforts on data that truly matters to the mission. Clearly defining the intended use of data aids in setting precise objectives and criteria for data quality. When data is aligned with its intended use, the most suitable data collection methods and tools can be selected. Considerations for maintaining alignment include monitoring for misuse or misinterpretation of data. Aligning data with its intended use allows potential sources of bias and error to be identified and mitigated early in the data lifecycle.

Security and Transmission (Strategy #13): Cybersecurity practices and transmission protocols at the point of data collection are crucial for ensuring data quality from the outset, as they protect data from unauthorized access and ensure secure data transfer. Implementing cybersecurity measures such as encryption and secure protocols at this stage protects data from tampering or unauthorized alterations.

Using secure transmission protocols helps safeguard data from interception or corruption during transit. Protocols like HTTPS and TLS encrypt data as it is transmitted from the collection point to storage or processing systems, ensuring that the data remains intact and unaltered. These secure transmission protocols often include error-checking mechanisms that detect and correct errors during data transmission. At the point of data collection, robust authentication and access control mechanisms ensure that only authorized personnel can input or access data. Cybersecurity practices often involve maintaining detailed logs and audit trails of data access and modifications. Overall, cybersecurity practices and transmission protocols at the point of data collection are essential for safeguarding data integrity, confidentiality, and reliability. They prevent unauthorized access and modifications, ensure secure data transmission, and support compliance with standards, all of which contribute to maintaining high data quality from the very beginning of the data lifecycle.

The mitigation strategies against threats to statistical data quality presented in this study aim to ensure data is securely acquired, accurately classified, and appropriately sourced to maintain its objectivity, integrity, and utility before statistical processing begins. Organizations can use this methodology to align their data management efforts, helping them to systematically identify and address potential vulnerabilities in their practices. By implementing these strategies, organizations can enhance their data governance processes, ensuring that data-driven decision-making is based on reliable and high-quality information (Segun-Falade, 2024). This proactive stance fosters a culture of continuous improvement in data collection practices.

The strategies presented in this study were applied to work performed for the U.S. Census Bureau's Frames Program: a transformational program to link multi-sourced datasets across shared characteristics to enable more innovative uses of enterprise data. By employing secure acquisition methods, the program ensures that data is protected from unauthorized access and tampering, maintaining its integrity from the point of collection. Accurate classification through the use of metadata and taxonomies allows for consistent organization and categorization of data, facilitating seamless integration and analysis. Appropriate sourcing practices verify the provenance of data and enhance its quality through enrichment techniques and preparing the data for its intended use. The mitigation strategies presented in this study contributed to the achievement of the benefits expected from a modernized, linked data ecosystem and provided a framework for identifying and addressing potential vulnerabilities in data management activities (U.S. Census Bureau, 2024).

6 CONCLUSIONS

To accomplish data driven decision making, it is paramount to safeguard data quality from the outset. The integrity, objectivity and utility of data significantly influence the outcome of analysis, and the decisions based on them. As illustrated in Figure 2 below, approximately 30% of the threats identified in the FCSM framework pertain to the "Acquire" stage of the USGS SDLM data lifecycle which corresponds to the data collection stage of statistical analysis.

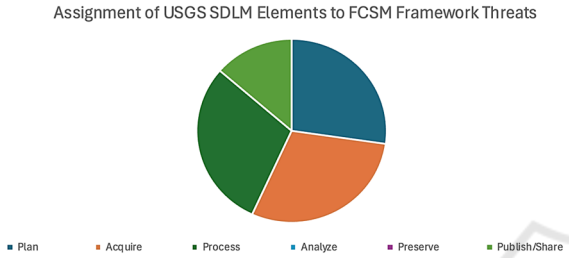

Figure 2: USGS SDLM Elements Assigned to FCSM Framework Threats.

This finding underscores the critical importance of addressing potential data quality threats during the data acquisition/collection phase. By taking deliberate steps to address and mitigate these threats, confidence in the statistical data used for decision making can be increased by 30%.

Addressing data quality during collection and before data processing offers several significant benefits that enhance the overall effectiveness and reliability of data-driven activities. Safeguarding data quality at the point of collection allows for the early identification of errors and inaccuracies, preventing them from propagating through subsequent stages of data processing and analysis. Detecting and addressing data quality threats early in the data lifecycle is generally more cost-effective than rectifying problems later, as early intervention reduces the need for extensive data cleaning and reprocessing, saving time and resources. High-quality data collected from the outset provides a solid foundation for analysis and decision-making, leading to more accurate insights and better-informed decisions, which enhance organizational outcomes.

In addition, maintaining data integrity, objectivity, and utility during collection is crucial for building trust in the data and the conclusions drawn from it. When data quality is assured at the collection stage, subsequent data processing becomes more efficient, as clean, accurate data reduces the complexity and time required for processing and analysis. Stakeholders can have greater confidence in the data and its analyses when data quality is prioritized from the beginning, fostering trust in data-driven strategies and initiatives. Addressing data quality early also helps ensure compliance with regulatory requirements and standards, reducing the risk of data breaches or misuse, which can have legal and reputational consequences.

High-quality data is easier to integrate with other datasets, enabling more comprehensive analyses and insights. Consistent and accurate data supports seamless data integration across systems and platforms. For systems that rely on user input, ensuring data quality at collection enhances the user experience by reducing the likelihood of errors and the need for repeated data entry. Quality data is also essential for advanced analytics, including machine learning and predictive modelling, providing a robust foundation for these sophisticated analyses.

Addressing data quality threats during collection and before processing establishes the utility, objectivity, and integrity of linked data in government ecosystems at the outset of statistical analysis. This proactive approach justifies the significant cost and effort of maintaining high-quality, multi-sourced integrated data; enhances trust in data-driven processes; and ultimately contributes to improved data-driven decision-making.

7 FUTURE RESEARCH

Next step for this work is the development of a vulnerability assessment tool designed to measure the level of exposure to data quality threats in linked data ecosystems.

A series of targeted questions will be formulated to evaluate the extent to which data quality levels may be compromised within the themes that emerged from this study. These questions will be directly aligned with the mitigation strategies pertinent to each identified data quality threat creating a structured framework to facilitate the systematic assessment of vulnerabilities within data collection processes.

Each question in the assessment will be accompanied by a range of potential responses, each assigned a score indicative of the associated vulnerability level. The scoring mechanism will enable a quantitative assessment of the data quality threats, providing a nuanced understanding of the robustness of data collection activities. The resultant vulnerability tool will serve as a diagnostic instrument, helpful in identifying data curation areas for improvement.

The versatility of this tool will be a key aspect of future research. While designed to be generic and applicable across various data integration scenarios, the tool will also offer customization options to cater to specific thematic focuses. This adaptability will ensure that the tool can be effective across diverse integration platforms.

ACKNOWLEDGMENTS & DISCLOSURE

This paper was funded by the MITRE Corporation Research Enablement and Augmentation Program (REAP). We thank Mike Fleckenstein, Mala Rajamani, and Sherri Walter for their perspectives and input to this work. The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

REFERENCES

Dworak-Fisher, K., Mirel, L., Parker, J., Popham, J., Prell, M., Schmitt, R., Seastrom, M., & Young, L. (2020). *A framework for data quality.* Federal Committee on Statistical Methodology.

Faundeen, J., Burley, T., Carlino, J., Govoni, D., Henkel, H., Holl, S., Hutchison, V., Martín, E., Montgomery, E., Ladino, C., Tessler, S., & Zolly, L. (2013). *U.S. Geological Survey Science Data Lifecycle Model.* U.S. Geological Survey.

Mirel, L., Singpurwalla, D., Hoppe, T., Liliedahl, E., Schmitt, R., & Weber, J. (2023). *A framework for data quality: Case studies.* Federal Committee on Statistical Methodology.

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in federal statistics: Combining data sources while protecting privacy.* The National Academies Press. https://doi.org/10.17226/24652

Parker, J., Mirel, L., Lee, P., Mintz, R., Tungate, A., & Vaidyanathan, A. (2024). Evaluating data quality for blended data using a data quality framework. *Statistical Journal of the IAOS.* 40(1), 125-136

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., & Gómez-Pérez, A. (2015). A comprehensive quality model for linked data. *Semantic Web, 9*(1), 3–24.

Segun-Falade, O., Osundare, O., Abioye, K., Adeleke, A., Pelumi, C., & Efunniyi, E. (2024). Operationalizing data governance: A workflow-based model for managing data quality and compliance. *International Journal of Engineering Invent, 13*(9), 142–150.

Serra, F., Peralta, V., Marotta, A., & Marcel, P. (2022). Use of context in data quality management: A systematic literature review. *Journal of Data Quality Management, 12*(3), 45–67.

U.S. Census Bureau. (2024). *Frames enabling transformation.* Census.gov. https://www.census.gov/library/video/2024/frames-enabling-transformation.html

Wang, R., Storey, V., & Firth, C. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering, 7*(4), 623–640. https://doi.org/10.1109/69.404034