Data & Trust Alliance

Data Provenance Standards Use Case Scenarios

These use case scenarios showcase how the Data Provenance standards support diverse needs across the data ecosystem.

Visit the Data Provenance Standards Metadata Generator to create and download standardized metadata files in JSON, CSV, or CML format to meet the Data Provenance Standards and facilitate data sharing

Visit dataandtrustalliance.org/work/data-provenance-standards to learn more about the Data Provenance Standards.

Scenario 1

Healthcare insurance data procurement

Evaluating a new dataset that contains comprehensive patient and insurance payment information, for use in predictive analytics

Scenario 2

Media consumption pattern dataset for consumer behavior insights

Curating a high-quality dataset that tracks media consumption habits across diverse platforms for content personalization

Scenario 3

Financial services customer product enablement

Evaluating a new dataset for refining AI algorithms used in customer credit card offerings

Scenario 4

Enhancing global logistics efficiency through AI-driven tariff harmonization

Managing data to refine AI systems for accurately predicting tariff costs across countries and categories

Scenario 1 Healthcare insurance data procurement

Bella and her team are evaluating a new dataset that contains comprehensive patient and insurance payment information. This dataset is considered crucial for enhancing the company's predictive analytics models, which forecast healthcare trends, personalize insurance plans, and optimize claim processing.

Challenges

- Balancing the need for detailed, comprehensive data with privacy and confidentiality requirements.
- Ensuring the dataset's metadata is accurate, up-to-date, and compliant with evolving Data Provenance Standards.
- Integrating new datasets with existing systems and models without compromising data integrity or system performance.
- Navigating the complex landscape of healthcare regulations and ensuring all data usage is compliant.

Bella Ramirez
Procurement Team Lead

Responsibilities

- Leads the procurement team in evaluating and acquiring high-quality datasets to improve the company's analytical models.
- Facilitates vendor reviews and ensuring all datasets comply with Data Provenance Standards, including transparent AI data usage, metadata coverage, and regulatory requirements.
- Partners with the data team charged with integrating new datasets into existing systems ensures that procured data meets their operational needs.
- Collaborates with the legal and compliance departments to ensure data usage aligns with healthcare regulations and company policies.
- Contributes to the success of strategies to leverage data insights for innovative marketing and improved customer trust.

S Source

S1 Dataset title/name REQUIRED

This is the official name of the dataset, which should be descriptive and help easily identify the dataset's content and purpose. The data supplier should provide a concise but descriptive name. For example, '2023 Customer Satisfaction Survey Data'. There are no character limits, but best practice is to limit the assigned name to 31 characters.

2024 Comprehensive Patient Care and Insurance Claims Dataset

> The filename doesn't appear in the company's data governance tool, indicating the data is unlikely to have been acquired or ingested previously.

S2 Unique metadata identifier REQUIRED

This field must contain a distinct identifier (such as a Universally Unique Identifier or UUID) assigned to the dataset's metadata to uniquely distinguish it from others, ensuring no confusion or overlap. A UUID is represented as hexadecimal digits in five groups separated by hyphens, for example, c75f7c66-e858-47d6-bb82-7ea5547c800c. If your organization does not have a standardized tool for generating a UUID, there are plenty of free online UUID generators, and the probability of a duplicate being created is extremely small.

UUID-1234-5678-9012-3456

> The metadata unique ID doesn't appear in the company's procurement database, confirming that it has not previously been evaluated for acquisition.

S3 Metadata location

This field is designed to capture a web address where the dataset's metadata is published and can be accessed, providing a direct link to detailed information about the dataset. For example, 'https://data-provider.com/metadata/12345'.

https://example.com/dataset/metadata/UUID-1234-5678-9012-3456

> The metadata about the dataset under consideration is comprehensive and can be evaluated for trustworthiness.

S4 Data issuer REQUIRED

This field must contain the name and address of the legal entity responsible for providing the dataset, ensuring accountability and a point of contact for inquires. Enter in this field the name of the legal entity that created the dataset, for example, 'XYZ Corporation' and the legal entity's business address.

Legal entity name: National Health Data Systems

Legal entity address: 1234 Wellness Way, Healthville, HT 56789, United States

> The data in the dataset was provided by the National Health Data Systems, an entity listed in the organization's procurement database, meaning they have previously supplied data to the enterprise.

S5 Description of the dataset REQUIRED

This field must contain a detailed narrative that explains the contents, scope, and purpose of the dataset. It provides essential contextual information that helps users understand what the data represents, how it was collected, and any limitations or recommended uses. Enter a comprehensive description that will help the procurement team and data consumers understand the dataset's content, scope, and purpose. You should write a detailed description covering key aspects. For example, 'This dataset includes survey responses from over 10,000 customers, collected to gauge satisfaction levels across various service areas'.

This dataset contains detailed records of patient care and associated insurance claims across multiple healthcare facilities in California for the year 2024. It includes patient demographics, diagnoses, treatment procedures, medication records, and outcome metrics. Additionally, it covers comprehensive insurance claim data, detailing claim statuses, reimbursement rates, and insurance provider information.

> The dataset provides a detailed and localized snapshot of healthcare and insurance interactions in California, which can support targeted analyses and policy development specific to the state's healthcare system.

P Provenance

P1 Source metadata for dataset REQUIRED

Identifies where the metadata for any source datasets that contribute to the current dataset can be found, establishing lineage and dependencies. This field establishes lineage. In this field, provide the metadata location or reference for any source datasets. For example, 'Metadata available at 'https://data-provider.com/metadata/source123'

https://example.com/metadata/sources

> The single URL indicates there is only one source of data that contributed to the creation of the one being evaluated. This URL provides the same context for the feeder dataset as is available for the dataset under consideration for acquisition.

P2 Source

This field identifies whether the data originates from a different organization than the one who issued the dataset, and identifies that original source's legal name. In this field, enter the name of the original data source if different from the issuer. For example, 'ABC Data Services'.

[ ] Source is the same as Data Issuer

Legal entity name: Nationwide Hospitals Systems, Insurance Providers Ltd.

Legal entity address: 1234 Elm Street, Suite 567, Springfield, XY 78901 USA

> The data was provided by Nationwide Hospitals Systems and Insurance Providers Ltd., not National Health Data Systems. Although Nationwide is not in the procurement database, it has a favorable rating in the Dun & Bradstreet vendor database.

P3 Data origin geography REQUIRED

The geographical location where the data was originally collected, which can be important for compliance with regional laws and understanding the data's context. In this field, indicate the geographic location, by selecting the continent, followed by country, state and city from where the data originated. If there are multiple locations, indicate that by adding additional rows. If the data geography is highly diverse, enter worldwide as the value.

North America → United States of Ame → CA

> The data offered originated in California, which may mean it is subject to the California Consumer Privacy Act (CCPA) if the dataset contains personal data.

P4 Dataset issue date REQUIRED

This field must contain the date when the dataset was compiled or created, providing a temporal context for the data. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

2024-03-15:T01:00:00

> The dataset was created on January 10, 2024 and initially published on March 15, 2024, which indicates period updates to data that represents year-long data compilation.

P5 Date of previously issued version of the dataset

This field is intended to capture the release date of the last version of the dataset, if it has been updated or revised, to track changes and updates over time. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

[ ] Dataset has not been issued previously

2024-01-10:T11:00:57

> The current dataset was previously issued on March 15, 2024, almost two months after its creation date, indicating a relatively frequent update rate for periodic data.

P6 Range of dates for data generation

These fields represent the span of time during which the data within the dataset was collected or generated, offering insight into the dataset's timeliness and relevance. This field helps users understand the dataset's timeliness. You should indicate the date range, for example, '2024-01-01T00:00:00' to '2025-01-01T00:00:00'.

Oldest component of data contained in the dataset.: 2023-01-01:T00:00:01

Youngest component of data contained in the dataset.: 2023-12-31:T23:23:59

> The dataset was created ten days after the data collection period, thus the data is recent and appropriate for refinement of predictive analytics.

P7 Method REQUIRED

This field must account for the methodology, or procedures used to collect, generate, or compile the data, giving insight into its reliability and validity. Understanding the methodology is crucial for assessing data reliability. Describe the data collection method by selecting values from the drop-down options, for example, 'Survey conducted via online questionnaires'.

General use category: Feeds

Specific use: Real time database info

Value: Electronic Health Records Extraction and Insurance Claim Processing Logs

> The dataset was created from real time database feeds and thus the data is well structured and any errors or anomalies were likely addressed quickly since database feeds provide real time error detection.

P8 Data format REQUIRED

This field describes the nature of the data within the dataset, such as numerical, textual, or multimedia, helping users understand what kind of information is contained within the file or dataset. Knowing the data format helps users prepare for how to handle the data. In this field you should specify the data format. For example, 'application/json'.

General format: Application

Specific format: SQL

> SQL format will enable the data analytics team to perform precise and efficient querying and manipulation of the data. This makes data retrieval and management more straightforward and effective.

U Use

U1 Confidentiality classification REQUIRED

This field must specify the level of sensitivity assigned to the dataset, such as personally identifiable information, which dictates how the dataset must be secured and who can access it. Proper classification ensures data is handled appropriately. Use the dropdown menu to indicate the confidentiality level, for example, 'Personally Identifiable Information, PII'.

Select which of the following are known to be present within the dataset:

- [ ] None
- [ ] Personal Information (PI)/Demographic
- [ ] Payment Card Industry (PCI)
- [ ] Personal Financial Information (PFI)
- [ ] Personally Identifiable Information (PII)
- [x] Personal Health Information (PHI)
- [ ] Sensitive Personal Information (SPI)
- [ ] Other

> Strict data protection requirements due to the sensitive nature of the information will be necessary, ensuring only authorized personnel who have a legitimate need for the data in the course of their duties have access.

U2 Consent documentation location REQUIRED

Specifies where the consent documentation or agreements related to the data can be found, ensuring legal compliance and regulatory use. Documenting consent ensures compliance with data protection regulations. Provide the location of consent documents. For example, 'https://data-provider.com/consent/12345'. If the data contained within the dataset is not personal in nature and does not require a consent for use, ensure the provided checkbox is checked.

[x] Consent has been collected

Consent location: https://example.com/dataset/UUID-1234-5678-9012-3456/consent1.html

> The consent outlines the scope, purpose, and conditions under which the data was collected and how it can be used. Bella's colleagues in the legal department confirm that the consent extends to data sharing with third parties and can be used for predictive analytics purposes.

U3 Privacy enhancing technologies REQUIRED

Were privacy enhancing technologies (PETs) or tools applied to the dataset to remove, mask, or modify PII/SPI in the data? Indicate whether techniques were used to protect personally identifiable information (PII) or sensitive personal information (SPI), highlighting the dataset's privacy considerations. This field ensures that privacy measures are clearly communicated. State if PETs were used and describe them. For example, 'ARX' could be the name of the PETs tool used to apply differential privacy techniques to the dataset. You should specify if noise addition was used to obfuscate individual data points, ensuring that the privacy of personally identifiable information (PII) and sensitive personal information (SPI) is maintained.

No privacy enhancing technologies (PETs) or tools have been applied to the dataset

> California's data privacy regulation (CCPA) requires the use of privacy-preserving measures when handling personal data. Not applying PETs could result in non-compliance with such regulations, leading to legal and financial risk.

U4 Data processing geography

This field defines the geographical boundaries within which the data can or cannot be processed, often for legal or regulatory reasons. Use this field to specify any geographic restrictions. For example, if the dataset must be processed within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

[x] Data processing geography is the same as data origin geography

IN EX | North America → United States of Amer → CA

> If the dataset is acquired, the data will have to be ingested into algorithms and worked on through the company's Santa Clara office location since it it cannot be processed on the East Coast or in Europe.

U5 Data storage geography

This field specifies where the data is stored and any geographical restrictions on storage locations, crucial for compliance with data sovereignty laws. For example, if the dataset must be stored within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

[x] Data storage geography is the same as data processing geography

IN EX | North America → United States of Amer → CA

> If this dataset is acquired, the data will need to be stored in database at the company's Santa Clara, CA location or in the cloud in a manner that limits the data to servers located in California.

U6 License to use REQUIRED

Details the terms under which the dataset can be used, including any restrictions or obligations, clarifying legal use and distribution rights. Clear licensing terms ensure legal use and distribution of the dataset.

Select types of license

- [ ] None
- [ ] Non-commercial
- [ ] Public license
- [x] Commercial/Negotiated License

Location or point of contact: License details available upon request from the Data Governance Department, National Health Data Repository, contactme@example.com

> The dataset comes with a commercial license, which means there are specific terms and conditions governing how the dataset can be used, especially in commercial settings. Bella will need to confer with the legal department about uses and restrictions once they contact the Data Governance Department at the National Health Data Repository and obtain a copy of the license.

U7 Intended data use REQUIRED

This field must describe the purpose for which the dataset was created, guiding users on its intended use and potential applications against identified use cases. Stating the intended use helps users understand the dataset's purpose. You should describe the intended use, for example, 'AI, pre-training'.

Select from the following uses

Non-AI

- [ ] Staging/testing
- [ ] Production
- [ ] Quality assurance
- [ ] Other

AI

- [x] Pre-training
- [ ] Alignment
- [ ] Evaluation
- [ ] Synthetic data generation
- [ ] Other

> The dataset is designed for AI "Pre-Training," aligning with her company's specific use cases. This alignment ensures the dataset's applicability and avoids misuse.

U8 Proprietary data presence REQUIRED

This field must indicate whether the dataset contains proprietary information that is owned by or exclusive to the organization, affecting how it can be shared or used. Knowing if data is proprietary affects how it can be shared. In these fields, indicate if the dataset contains proprietary data subject to copyright, trademark, or patent. For example, if a copyright were applicable, you should enter the contact information for the individual who can speak to the copyright requirements, such as 'Jonathan Reeves, Esq., Email: jreeves@example.com, Phone: +1-555-012-3456'. If no IP restrictions are in place, select 'None'.

Select types of proprietary data present

- [x] None
- [ ] Copyright
- [ ] Patent
- [ ] Trademark

> The dataset contains no proprietary information, meaning that the company will avoid intellectual property infringement and ensure compliance with contractual agreements.

Additional comments

This dataset is on offer to three other organizations and is a complementary dataset to "Healthcare Assessments and Insurance Payments 2023" issues on January 1, 2024.

Outcome

Through comprehensive metadata evaluation, the procurement team was able to perform a more in-depth assessment of a crucial dataset, decreasing procurement time and risk while enhancing the company's predictive analytics capabilities and maintaining adherence to legal and ethical standards. The benefits of having access to the dataset metadata include:

Better dataset evaluation

- Quicker insights into the dataset's provenance without having to read through many pages of descriptive text —due to having the title, unique identifier, and details about data origin and collection methods.
- Evidence (not attestation) of a dataset's compliance with healthcare regulations—due to having confidentiality, consent documentation, data processing and storage geographies.
- Better coordination between procurement and data teams to assess the dataset's impact on analytical models—allowing for seamless integration and operational improvement—due to having clarity from the metadata.

Increased balance of risk and opportunity

- The organization's use of innovative marketing strategies require transparency into (1) the dataset's generation method and (2) the permitted use—in order to be both valuable and trustworthy. The metadata was able to provide the necessary transparency into both before purchase.
- Assessment of risk was ultimately much faster as the metadata helped legal and compliance teams clarify, at the onset, who needed to be involved in which aspects of the dataset acquisition evaluation.

Improved data procurement and legal validation

- Suspect data processing and storage metadata resulted in increased legal scrutiny in advance of the data acquisition, leading to the correction of inaccuracies in the metadata.
- The metadata evaluation process provided essential legal and regulatory assurance, enabling the organization to confidently progress towards its business goals with strategic and compliant use of the dataset.

Scenario 2

Media consumption pattern dataset for consumer behavior insights

Jordan's current project involves curating a dataset that tracks media consumption habits across diverse platforms. This dataset aims to empower media buyers and sellers in accurately targeting their audience segments, facilitating personalized content strategies for industries ranging from consumer goods to tourism.

Challenges

- Balancing data comprehensiveness with privacy and ethical considerations.
- Keeping pace with rapid changes in media consumption behaviors and technology.
- Ensuring data standards provide necessary transparency to data buyers and that the metadata is compatible with automated data procurement systems.

Jordan Liu
Data Strategy Director

Responsibilities

- Oversees the development and distribution of comprehensive media consumption datasets.
- Ensures datasets adhere to the latest Data Provenance Standards for transparency and reliability.
- Collaborates with stakeholders across healthcare, consumer goods, and travel industries to tailor data offerings.
- Guides the integration of datasets into client systems to optimize targeted content delivery and marketing strategies.
- Advocates for data-driven decision-making within the company and among clients to foster industry innovation.

S Source

S1 Dataset title/name
REQUIRED

This is the official name of the dataset, which should be descriptive and help easily identify the dataset's content and purpose. The data supplier should provide a concise but descriptive name. For example, '2023 Customer Satisfaction Survey Data'. There are no character limits, but best practice is to limit the assigned name to 31 characters.

March 2024 Global Media Consumption Trends

> Communicates the dataset's scope, timing, and focus, facilitating recognition and understanding what the dataset is about.

S2 Unique metadata identifier
REQUIRED

This field must contain a distinct identifier (such as a Universally Unique Identifier or UUID) assigned to the dataset's metadata to uniquely distinguish it from others, ensuring no confusion or overlap. A UUID is represented as hexadecimal digits in five groups separated by hyphens, for example, c75f7c66-e858-47d6-bb82-7ea5547c800c. If your organization does not have a standardized tool for generating a UUID, there are plenty of free online UUID generators, and the probability of a duplicate being created is extremely small.

550e8400-e29b-41d4-a716-446655440000

> Ensures precise tracking and referencing of the dataset within the data ecosystem, including customer's acquisitions databases.

S3 Metadata location

This field is designed to capture a web address where the dataset's metadata is published and can be accessed, providing a direct link to detailed information about the dataset. For example, 'https://data-provider.com/metadata/12345'.

example.com/550e8400-e29b-41d4-a716-446655440000/metadata.html

> Offers immediate, transparent access to the dataset's detailed specifications, fostering trust and ease of use among data consumers.

S4 Data issuer
REQUIRED

This field must contain the name and address of the legal entity responsible for providing the dataset, ensuring accountability and a point of contact for inquires. Enter in this field the name of the legal entity that created the dataset, for example, 'XYZ Corporation' and the legal entity's business address.

Legal entity name: AnalytiQuest Ventures

Legal entity address: 345 Innovation Drive, Suite 800, Tech Valley, CA 94025, United States of America

> Establishes the company as the authoritative source and responsible entity for the dataset, providing a clear point of accountability.

S5 Description of the dataset
REQUIRED

This field must contain a detailed narrative that explains the contents, scope, and purpose of the dataset. It provides essential contextual information that helps users understand what the data represents, how it was collected, and any limitations or recommended uses. Enter a comprehensive description that will help the procurement team and data consumers understand the dataset's content, scope, and purpose. You should write a detailed description covering key aspects. For example, 'This dataset includes survey responses from over 10,000 customers, collected to gauge satisfaction levels across various service areas'.

The dataset represents media consumption habits across various platform, providing key insights for developing targeted, personalized content strategies for sectors like consumer goods and tourism.

> Offers insights into data patterns, which can help tailor and optimize content strategies across diverse industries by accurately targeting specific audience segments.

P Provenance

P1 Source metadata for dataset
REQUIRED

Identifies where the metadata for any source datasets that contribute to the current dataset can be found, establishing lineage and dependencies. This field establishes lineage. In this field, provide the metadata location or reference for any source datasets. For example, 'Metadata available at 'https://data-provider.com/metadata/source123'

example.com/ 550e8400-e29b-41d4-a716-44665543902

example.com/ 550e8400-e29b-41d4-a716-44665544732

example.com/ 550e8400-e29b-41d4-a716-446655465722

> Demonstrates the three sources of data that make up the existing dataset, providing traceability into the origins and lineage of the data.

P2 Source

This field identifies whether the data originates from a different organization than the one who issued the dataset, and identifies that original source's legal name. In this field, enter the name of the original data source if different from the issuer. For example, 'ABC Data Services'.

- [x] Source is the same as Data Issuer

Legal entity name: AnalytiQuest Ventures

Legal entity address: 345 Innovation Drive, Suite 800, Tech Valley, CA 94025, United States of America

> Indicates that the entity that is supplying the data is the same as the entity that created the dataset, not just a data reseller.

P3 Data origin geography
REQUIRED

The geographical location where the data was originally collected, which can be important for compliance with regional laws and understanding the data's context. In this field, indicate the geographic location, by selecting the continent, followed by country, state and city from where the data originated. If there are multiple locations, indicate that by adding additional rows. If the data geography is highly diverse, enter worldwide as the value.

North America → United States of Ame → FL

> Conveys the focused nature of the dataset on the Floridian market. This granularity not only aids in compliance with regional data laws but also enriches the dataset's contextual relevance for analyses tailored to the state as the geographic area.

P4 Dataset issue date
REQUIRED

This field must contain the date when the dataset was compiled or created, providing a temporal context for the data. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

2024-01-10:T01:00:00

> Provides a clear context for the dataset, indicating its recency and relevance for users seeking the most current insights into media consumption trends as of early 2024.

P5 Date of previously issued version of the dataset

This field is intended to capture the release date of the last version of the dataset, if it has been updated or revised, to track changes and updates over time. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

- [x] Dataset has not been issued previously

YYYY-MM-DD:THH:MM:SS

> Conveys the dataset's foundation on real consumer behaviors and feedback, providing direct insight into media consumption patterns, which enhances the dataset's reliability and validity for analyzing consumer engagement across various media platforms.

P6 Range of dates for data generation

These fields represent the span of time during which the data within the dataset was collected or generated, offering insight into the dataset's timeliness and relevance. This field helps users understand the dataset's timeliness. You should indicate the date range, for example, '2024-01-01T00:00:00' to '2025-01-01T00:00:00'.

Oldest component of data contained in the dataset: 2023-01-01:T00:00:01

Youngest component of data contained in the dataset: 2023-12-31:T23:23:59

> Indicates that the dataset encompasses a full year's worth of media consumption data, providing insights into trends and patterns over this period and ensuring the dataset's relevance for analyses focused on the year 2023.

P7 Method
REQUIRED

This field must account for the methodology, or procedures used to collect, generate, or compile the data, giving insight into its reliability and validity. Understanding the methodology is crucial for assessing data reliability. Describe the data collection method by selecting values from the drop-down options, for example, 'Survey conducted via online questionnaires'.

General use category: User generated content

Specific use: Other

Value: Digital Interaction Tracking and Survey Responses

> Signifies that users are accessing the inaugural edition of this dataset, setting expectations for its novelty and current relevance in understanding media consumption trends.

P8 Data format
REQUIRED

This field describes the nature of the data within the dataset, such as numerical, textual, or multimedia, helping users understand what kind of information is contained within the file or dataset. Knowing the data format helps users prepare for how to handle the data. In this field you should specify the data format. For example, 'application/json'.

General format: Application

Specific format: SQL

General format: Text

Specific format: plain

Specific format: .doc

Specific format: csv

Specific format: xls

> Indicates that the dataset contains a mix of structured data (like spreadsheets and databases) and unstructured data (like text docs), offering a diverse range of information that caters to both quantitative analysis and qualitative insights. It also provides insight into the potential need for data cleanup.

U Use

U1 Confidentiality classification
REQUIRED

This field must specify the level of sensitivity assigned to the dataset, such as personally identifiable information, which dictates how the dataset must be secured and who can access it. Proper classification ensures data is handled appropriately. Use the dropdown menu to indicate the confidentiality level, for example, 'Personally Identifiable Information, PII'.

Select which of the following are known to be present within the dataset:

- [x] None
- [ ] Personal Information (PI)/Demographic
- [ ] Payment Card Industry (PCI)
- [ ] Personal Financial Information (PFI)
- [ ] Personally Identifiable Information (PII)
- [ ] Personal Health Information (PHI)
- [ ] Sensitive Personal Information (SPI)
- [ ] Other

> The dataset does not contain any personal information, as verified by running the Talend Data Fabric tool.

U2 Consent documentation location
REQUIRED

Specifies where the consent documentation or agreements related to the data can be found, ensuring legal compliance and regulatory use. Documenting consent ensures compliance with data protection regulations. Provide the location of consent documents. For example, 'https://data-provider.com/consent/12345'. If the data contained within the dataset is not personal in nature and does not require a consent for use, ensure the provided checkbox is checked.

- [ ] Consent has been collected

U3 Privacy enhancing technologies
REQUIRED

Were privacy enhancing technologies (PETs) or tools applied to the dataset to remove, mask, or modify PII/SPI in the data? Indicate whether techniques were used to protect personally identifiable information (PII) or sensitive personal information (SPI), highlighting the dataset's privacy considerations. This field ensures that privacy measures are clearly communicated. State if PETs were used and describe them. For example, 'ARX' could be the name of the PETs tool used to apply differential privacy techniques to the dataset. You should specify if noise addition was used to obfuscate individual data points, ensuring that the privacy of personally identifiable information (PII) and sensitive personal information (SPI) is maintained.

Specify tool: AnalytiQuest Ventures

Parameter	Key value
data_category	user_identifiers
anonymization_technique	differential_privacy
privacy_budget	1.0
data_sensitivity	high
aggregation_level	zip_code
anonymization_strength	95%
data_utility_retention	85%

> The reason that no personal data is present is because data anonymization technology (Adverity) was used with 95% resulting strength.

U4 Data processing geography

This field defines the geographical boundaries within which the data can or cannot be processed, often for legal or regulatory reasons. Use this field to specify any geographic restrictions. For example, if the dataset must be processed within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

- [ ] Data processing geography is the same as data origin geography

U5 Data storage geography

This field specifies where the data is stored and any geographical restrictions on storage locations, crucial for compliance with data sovereignty laws. For example, if the dataset must be stored within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

- [ ] Data storage geography is the same as data processing geography

U6 License to use
REQUIRED

Details the terms under which the dataset can be used, including any restrictions or obligations, clarifying legal use and distribution rights. Clear licensing terms ensure legal use and distribution of the dataset.

Select types of license

- [ ] None
- [ ] Non-commercial
- [ ] Public license
- [x] Commercial/Negotiated License

Location or point of contact: AnalytiQuest Ventures's Office of General Counsel, legalconsumptionlicense@example.com and (555) 123-4567

> Provides a clear pathway for understanding the legal framework and usage rights associated with the dataset, ensuring that potential users are aware of their obligations, restrictions, and permissions before leveraging the data for their purposes.

U7 Intended data use
REQUIRED

This field must describe the purpose for which the dataset was created, guiding users on its intended use and potential applications against identified use cases. Stating the intended use helps users understand the dataset's purpose. You should describe the intended use, for example, 'AI, pre-training'.

Select from the following uses

Non-AI

- [ ] Staging/testing
- [ ] Production
- [ ] Quality assurance
- [ ] Other

AI

- [ ] Pre-training
- [ ] Alignment
- [x] Evaluation
- [ ] Synthetic data generation
- [ ] Other

> Signals that the dataset is primed for use in artificial intelligence applications, performance assessments, and educational or developmental programs, guiding consumers towards its applications in enhancing media planning, content development, and industry research.

U8 Proprietary data presence
REQUIRED

This field must indicate whether the dataset contains proprietary information that is owned by or exclusive to the organization, affecting how it can be shared or used. Knowing if data is proprietary affects how it can be shared. In these fields, indicate if the dataset contains proprietary data subject to copyright, trademark, or patent. For example, if a copyright were applicable, you should enter the contact information for the individual who can speak to the copyright requirements, such as 'Jonathan Reeves, Esq., Email: jreeves@example.com, Phone: +1-555-012-3456'. If no IP restrictions are in place, select 'None'.

Select types of proprietary data present

- [x] None
- [ ] Copyright
- [ ] Patent
- [ ] Trademark

> Relays that the dataset does not contain proprietary information exclusive to AnalytiQuest Ventures, suggesting broader flexibility in its use and sharing, which can facilitate unrestricted analysis and collaboration within the legal boundaries established for the dataset.

Additional comments

Enter comments and notes here

Outcome

Metadata associated with the "March 2024 Global Media Consumption Trends" dataset is a vital resource for procuring complex media consumption patterns, ensuring its integrity and applicability in AI analytics. This approach to describing data can facilitate effective personalization of content strategies across various industries and will set a new standard for transparent, efficient, and compliant data usage in media consumption analysis. The outcome includes:

Comprehensive coverage and actionable insights

- Better insight into the data's relevance and quality—allowing for more targeted content strategies. This was due to the detailed data origin geography and collection methodologies in the metadata.

Transparency and trust building

- Both media buyers and sellers now find it easier to assess the dataset's credibility due to the transparency provided by adhering to version 1.0.0 of Data Provenance Standards, and the clear metadata URL.
- The commitment to data privacy and security is clear due to the documentation of dataset lineage and the use of PETs-like anonymization—ultimately building trust among clients.

Increased efficiency and enhanced compliance

- The data acquisition process was streamlined as the metadata prompted a negotiation with the AnalytiQuest Ventures' Office of General Counsel—due to the visible lack of proprietary data restrictions and the provision of a clear license to use.
- Legal and reputational risks were significantly reduced as the data processing and storage geography metadata helps downstream consumers comply with legal standards and privacy regulations.

Scenario 3 — Financial services customer product enablement

Minh is tasked with evaluating a new dataset for refining AI algorithms for customer credit card offerings. The dataset under consideration has been documented in accordance with the latest Data Provenance Standards, ensuring transparency and compliance, especially under GDPR and the new EU AI Act. Minh's evaluation process focuses on the detailed metadata provided for the dataset.

Challenges

- Ensure dataset credibility through clear documentation of its lineage and metadata.
- Navigate diverse international regulations related to data privacy and AI.
- Integrate the new dataset with existing systems without operational disruptions.
- Balance proprietary data use with information protection and competitive edge.
- Confirm dataset use is ethical and consensual, particularly with sensitive data.
- Keep pace with technological and data standard advancements for AI relevance.

Minh Quang Nguyen
Data Architect and Policy Analyst

Responsibilities

- Designs and implements efficient data architectures that support ProForma's business goals.
- Works closely with IT teams to ensure that data structures are scalable, secure, and optimized for performance.
- Plays a crucial role in developing and enforcing data management policies, ensuring compliance with regulatory standards and protecting customer information.

S Source

S1 Dataset title/name — REQUIRED

This is the official name of the dataset, which should be descriptive and help easily identify the dataset's content and purpose. The data supplier should provide a concise but descriptive name. For example, '2023 Customer Satisfaction Survey Data'. There are no character limits, but best practice is to limit the assigned name to 31 characters.

Consumer Spending Patterns 2020-2024

> The dataset's focus on analyzing consumer behavior over a five-year period, aids in immediate recognition and relevance for financial trend analysis.

S2 Unique metadata identifier — REQUIRED

This field must contain a distinct identifier (such as a Universally Unique Identifier or UUID) assigned to the dataset's metadata to uniquely distinguish it from others, ensuring no confusion or overlap. A UUID is represented as hexadecimal digits in five groups separated by hyphens, for example, c75f7c66-e858-47d6-bb82-7ea5547c800c. If your organization does not have a standardized tool for generating a UUID, there are plenty of free online UUID generators, and the probability of a duplicate being created is extremely small.

LFS-1234-5678

> Provides unambiguous identification and retrieval for uses of the data across the enterprise and tracking internal workflow actions to the dataset, such as sign off for use by the privacy officer.

S3 Metadata location

This field is designed to capture a web address where the dataset's metadata is published and can be accessed, providing a direct link to detailed information about the dataset. For example, 'https://data-provider.com/metadata/12345'.

http://luminadataservices.com/metadata/1234-5678

> Where colleagues from other departments can go to review detailed information about the "Consumer Spending Patterns 2020-2024" dataset, for compliance and relevance reviews to their use cases.

S4 Data issuer — REQUIRED

This field must contain the name and address of the legal entity responsible for providing the dataset, ensuring accountability and a point of contact for inquires. Enter in this field the name of the legal entity that created the dataset, for example, 'XYZ Corporation' and the legal entity's business address.

Legal entity name: Lumina Financial Services

Legal entity address: 789 Beacon Street, Floor 12, Metropolis, NY 10022, United States of America

> Knowing that "Lumina Financial Services" is the dataset's creator, allows Minh to lookup the vendor in the procurement system and understand what other datasets are under consideration for acquisition from the same supplier. There may be an opportunity to negotiate a broader agreement and save money in the process.

S5 Description of the dataset — REQUIRED

This field must contain a detailed narrative that explains the contents, scope, and purpose of the dataset. It provides essential contextual information that helps users understand what the data represents, how it was collected, and any limitations or recommended uses. Enter a comprehensive description that will help the procurement team and data consumers understand the dataset's content, scope, and purpose. You should write a detailed description covering key aspects. For example, 'This dataset includes survey responses from over 10,000 customers, collected to gauge satisfaction levels across various service areas'.

This dataset compiles comprehensive consumer spending data from 2020 to 2024, capturing transactional behavior across various sectors. It includes anonymized user profiles, spending categories, transaction amounts, and timestamps. It is designed to refine AI algorithms for optimizing credit card offerings, enhancing predictive accuracy for financial institutions.

> Minh can determine whether the dataset generally aligns to his internal business cases and the intended purpose of the dataset. Knowing this information helps identify who else might need to be consulted within the organization before the dataset is acquired and used.

P Provenance

P1 Source metadata for dataset — REQUIRED

Identifies where the metadata for any source datasets that contribute to the current dataset can be found, establishing lineage and dependencies. This field establishes lineage. In this field, provide the metadata location or reference for any source datasets. For example, 'Metadata available at 'https://data-provider.com/metadata/source123'

http://luminadataservices.com/metadata/sources/retail-transactions-2023

http://luminadataservices.com/metadata/sources/retail-transactions-2024

> The foundation of the current dataset and its dependencies are relayed through the two URLs, increasing the transparency of organizations involved in producing data that will be used by Minh and his team.

P2 Source

This field identifies whether the data originates from a different organization than the one who issued the dataset, and identifies that original source's legal name. In this field, enter the name of the original data source if different from the issuer. For example, 'ABC Data Services'.

- [ ] Source is the same as Data Issuer

Legal entity name: PreciTech Data Inc

Legal entity address: Enter legal entity address

> The source of the data is not the same as the creator of the dataset, indicating that Minh may be dealing with a data broker and not the generator.

P3 Data origin geography — REQUIRED

The geographical location where the data was originally collected, which can be important for compliance with regional laws and understanding the data's context. In this field, indicate the geographic location, by selecting the continent, followed by country, state and city from where the data originated. If there are multiple locations, indicate that by adding additional rows. If the data geography is highly diverse, enter worldwide as the value.

- Europe → France
- Europe → Germany
- Europe → Italy
- Europe → Poland

> All locations reflect GDPR requirements so data transfer and processing requirements will be more stringent and complicate Minh's project from a compliance perspective.

P4 Dataset issue date — REQUIRED

This field must contain the date when the dataset was compiled or created, providing a temporal context for the data. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

2024-03-14:T10:23:09

> The dataset is quite recent,which is ideal for refining customer-inference algorithms.

P5 Date of previously issued version of the dataset

This field is intended to capture the release date of the last version of the dataset, if it has been updated or revised, to track changes and updates over time. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

- [ ] Dataset has not been issued previously

2023-03-15:T01:25:50

> The previously issued date is one one day after the dataset was created, indicating frequent data refresh rates.

P6 Range of dates for data generation

These fields represent the span of time during which the data within the dataset was collected or generated, offering insight into the dataset's timeliness and relevance. This field helps users understand the dataset's timeliness. You should indicate the date range, for example, '2024-01-01T00:00:00' to '2025-01-01T00:00:00'.

Oldest component of data contained in the dataset.: 2023-01-05:T01:01:59

Youngest component of data contained in the dataset.: 2024-03-14:T17:13:47

> There is over a full years worth of data and once collected, the data was quickly published and immediately updated. Thus the recency and vendor ability to refresh is good.

P7 Method — REQUIRED

This field must account for the methodology, or procedures used to collect, generate, or compile the data, giving insight into its reliability and validity. Understanding the methodology is crucial for assessing data reliability. Describe the data collection method by selecting values from the drop-down options, for example, 'Survey conducted via online questionnaires'.

General use category: Feeds

Specific use: Interval timed database info

Value: Aggregated Consumer Transaction Analysis, PoS

> The point of service (PoS) data will require special handling and protection as it is subject to data privacy regulations. The data is structured and ideal for the use cases.

P8 Data format — REQUIRED

This field describes the nature of the data within the dataset, such as numerical, textual, or multimedia, helping users understand what kind of information is contained within the file or dataset. Knowing the data format helps users prepare for how to handle the data. In this field you should specify the data format. For example, 'application/json'.

General format: Application

Specific format: SQL

> The SQL structure adheres to the Internet Assigned Number Authority (IANA) Media Types properties, ensuring reliable transaction processing and data integrity even in the event of system failures. It is a good fit for ongoing feeds and frequent data refreshes.

U Use

U1 Confidentiality classification — REQUIRED

This field must specify the level of sensitivity assigned to the dataset, such as personally identifiable information, which dictates how the dataset must be secured and who can access it. Proper classification ensures data is handled appropriately. Use the dropdown menu to indicate the confidentiality level, for example, 'Personally Identifiable Information, PII'.

Select which of the following are known to be present within the dataset:

- [x] None
- [ ] Personal Information (PI)/Demographic
- [ ] Payment Card Industry (PCI)
- [ ] Personal Financial Information (PFI)
- [ ] Personally Identifiable Information (PII)
- [ ] Personal Health Information (PHI)
- [ ] Sensitive Personal Information (SPI)
- [ ] Other

U2 Consent documentation location — REQUIRED

Specifies where the consent documentation or agreements related to the data can be found, ensuring legal compliance and regulatory use. Documenting consent ensures compliance with data protection regulations. Provide the location of consent documents. For example, 'https://data-provider.com/consent/12345'. If the data contained within the dataset is not personal in nature and does not require a consent for use, ensure the provided checkbox is checked.

- [ ] Consent has been collected

U3 Privacy enhancing technologies — REQUIRED

Were privacy enhancing technologies (PETs) or tools applied to the dataset to remove, mask, or modify PII/SPI in the data? Indicate whether techniques were used to protect personally identifiable information (PII) or sensitive personal information (SPI), highlighting the dataset's privacy considerations. This field ensures that privacy measures are clearly communicated. State if PETs were used and describe them. For example, 'ARX' could be the name of the PETs tool used to apply differential privacy techniques to the dataset. You should specify if noise addition was used to obfuscate individual data points, ensuring that the privacy of personally identifiable information (PII) and sensitive personal information (SPI) is maintained.

Specify tool: Google differential privacy library

Parameter	Key value
Epsilon	0.5
Delta	1e-5

> The dataset is anonymized and compliant with privacy regulations, balancing data utility with individual privacy.

U4 Data processing geography

This field defines the geographical boundaries within which the data can or cannot be processed, often for legal or regulatory reasons. Use this field to specify any geographic restrictions. For example, if the dataset must be processed within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

- [ ] Data processing geography is the same as data origin geography

U5 Data storage geography

This field specifies where the data is stored and any geographical restrictions on storage locations, crucial for compliance with data sovereignty laws. For example, if the dataset must be stored within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

- [ ] Data storage geography is the same as data processing geography

U6 License to use — REQUIRED

Details the terms under which the dataset can be used, including any restrictions or obligations, clarifying legal use and distribution rights. Clear licensing terms ensure legal use and distribution of the dataset.

Select types of license

- [ ] None
- [ ] Non-commercial
- [ ] Public license
- [x] Commercial/Negotiated License

Location or point of contact: http://luminadataservices.com/license/1234-5678

> Minh can refer his legal team to the provided license URL at "http://luminadataservices.com/license/1234-5678" to understand the specific terms, restrictions, and obligations for using the dataset, ensuring its legal and compliant application within his projects.

U7 Intended data use — REQUIRED

This field must describe the purpose for which the dataset was created, guiding users on its intended use and potential applications against identified use cases. Stating the intended use helps users understand the dataset's purpose. You should describe the intended use, for example, 'AI, pre-training'.

Select from the following uses

Non-AI

- [ ] Staging/testing
- [ ] Production
- [ ] Quality assurance
- [ ] Other

AI

- [ ] Pre-training
- [ ] Alignment
- [ ] Evaluation
- [ ] Synthetic data generation
- [x] Other

Specify other AI use type: Model monitoring

System fit: Baseline establishment

Business case: Automated training and retraining

> The use is a match for the requirements and use cases, and worth a premium data price.

U8 Proprietary data presence — REQUIRED

This field must indicate whether the dataset contains proprietary information that is owned by or exclusive to the organization, affecting how it can be shared or used. Knowing if data is proprietary affects how it can be shared. In these fields, indicate if the dataset contains proprietary data subject to copyright, trademark, or patent. For example, if a copyright were applicable, you should enter the contact information for the individual who can speak to the copyright requirements, such as 'Jonathan Reeves, Esq., Email: jreeves@example.com, Phone: +1-555-012-3456'. If no IP restrictions are in place, select 'None'.

Select types of proprietary data present

- [x] None
- [ ] Copyright
- [ ] Patent
- [ ] Trademark

> The dataset does not contain any information that is owned or controlled exclusively by the organization that created it. This means the data can likely be shared or used more freely, within the parameters of the license but without concern for violation of intellectual property rights.

Additional comments

Enter comments and notes here

Outcome

Minh's evaluation of the "Consumer Spending Patterns 2020-2024" dataset through the lens of the Data Provenance Standards significantly enhanced ProForma Financial Services' AI algorithms for customer credit card offerings, ensuring both heightened personalization and strict adherence to international data regulations. This approach improved the precision and effectiveness of the company's AI models and ensured compliance, data privacy, and seamless integration with existing systems, paving the way for responsible and innovative use of AI insights in the financial sector. The outcome includes:

Better business case alignment

- Instead of relying solely on high-level dataset descriptions, the Standards gave Minh the specificity—around origin, creation dates, and collection methodologies—to determine if the dataset was relevant to the needs of his business case.

Faster data acquisition timeline and speed to market

- The absence of proprietary data restrictions and clear licensing terms sped up ProForma's data acquisition process and ability to develop AI capabilities.

Increased compliance and integrity

- Use of the Data Provenance Standards meant that an initial assessment of fit-for-use could be performed by an automated system and then passed to a human for deeper review, significantly speeding up the compliance process.
- The legal team had a much faster gauge on the level of pre-processing needed to comply with data privacy requirements simply by looking at the dataset lineage, original sources, and PETs allowed.
- The data collection, processing and storage information further helped meet data privacy requirements and mitigate legal and reputational risks associated with GDPR and the EU AI Act.

Scenario 4

Enhancing global logistics efficiency through AI-driven tariff harmonization

The global nature of Navisphere Logistics, Ltd.'s operations means that the company must navigate a complex web of international tariffs and customs regulations. Efficiently managing these tariffs is critical to minimizing delivery times and costs. Dr. Hicks and her team are tasked with refining the company's AI systems to accurately predict tariff costs across different countries and product categories.

Challenges

- Navigate the intricate international tariff and customs landscape with diverse rules.
- Rigorously assess dataset metadata for compliance with global standards and privacy.
- Continually update AI models to adapt to changing international tariff regulations.
- Balance advanced AI development with responsible usage and adherence to privacy laws.
- Ensure smooth AI model integration into Navisphere Logistics' systems without workflow disruption.

Dr. Maya Hicks
Lead Data Scientist

Responsibilities

- Lead the AI team in improving tariff prediction models.
- Assess dataset integrity and compliance with corporate and international standards.
- Collaborate with procurement and legal to align data and AI with global regulations.
- Train and optimize AI for precise tariff predictions using advanced algorithms.
- Integrate and test refined AI in Navisphere Logistics' systems for accuracy and efficiency.
- Establish a feedback loop for AI model refinement based on real-world feedback.
- Ensure AI is used responsibly, in line with company standards and privacy laws.
- Communicate AI project updates to stakeholders and clients.
- Stay informed on AI and logistics innovations to foster company-wide advancements.

S Source

S1 Dataset title/name

REQUIRED

This is the official name of the dataset, which should be descriptive and help easily identify the dataset's content and purpose. The data supplier should provide a concise but descriptive name. For example, '2023 Customer Satisfaction Survey Data'. There are no character limits, but best practice is to limit the assigned name to 31 characters.

2023 Global Tariff Schedules - Electronics

The link between the metadata and dataset supports tracking both throughout the assessment and acquisition process.

S2 Unique metadata identifier

REQUIRED

This field must contain a distinct identifier (such as a Universally Unique Identifier or UUID) assigned to the dataset's metadata to uniquely distinguish it from others, ensuring no confusion or overlap. A UUID is represented as hexadecimal digits in five groups separated by hyphens, for example, c75f7c66-e858-47d6-bb82-7ea5547c800c. If your organization does not have a standardized tool for generating a UUID, there are plenty of free online UUID generators, and the probability of a duplicate being created is extremely small.

123e4567-e89b-12d3-a456-426614174000

Allows the Procurement Department and Maya to avoids redundancy as multiple sets of metadata are automated for scoring for trustworthiness and business value.

S3 Metadata location

This field is designed to capture a web address where the dataset's metadata is published and can be accessed, providing a direct link to detailed information about the dataset. For example, 'https://data-provider.com/metadata/12345'.

https://globaltradedatahub.com/metadata/123e4567-e89b-12d3-a456-426614174000

Provides a working location where the metadata, describing the dataset, can be obtained.

S4 Data issuer

REQUIRED

This field must contain the name and address of the legal entity responsible for providing the dataset, ensuring accountability and a point of contact for inquires. Enter in this field the name of the legal entity that created the dataset, for example, 'XYZ Corporation' and the legal entity's business address.

Legal entity name
GlobalTradeDataHub

Legal entity address
Bahnhofstrasse 45, 8001 Zürich, Switzerland

This is critical context for assessing the data's credibility, potential biases, and the appropriate point of contact for any technical or data-specific inquiries.

S5 Description of the dataset

REQUIRED

This field must contain a detailed narrative that explains the contents, scope, and purpose of the dataset. It provides essential contextual information that helps users understand what the data represents, how it was collected, and any limitations or recommended uses. Enter a comprehensive description that will help the procurement team and data consumers understand the dataset's content, scope, and purpose. You should write a detailed description covering key aspects. For example, 'This dataset includes survey responses from over 10,000 customers, collected to gauge satisfaction levels across various service areas'.

Detailed information in English, German, and French on international tariff rates and customs regulations for electronics for the 2023 calendar year.

The dataset offers comprehensive and specific tariff information necessary to enhance AI-driven predictions and management of shipping costs and regulations for electronics globally.

P Provenance

P1 Source metadata for dataset

REQUIRED

Identifies where the metadata for any source datasets that contribute to the current dataset can be found, establishing lineage and dependencies. This field establishes lineage. In this field, provide the metadata location or reference for any source datasets. For example, 'Metadata available at 'https://data-provider.com/metadata/source123'

https://internationalcustomsdataconsortium.com/metadata/234f5678-f01c-23d4-b567-537625175111

https://internationalcustomsdataconsortium.com/metadata/345g6789-g02d-34e5-c678-648736286222

https://internationalcustomsdataconsortium.com/metadata/456h7890-h03e-45f6-d789-759847397333

https://internationalcustomsdataconsortium.com/metadata/567i8901-i04f-56g7-e890-860958408444

https://internationalcustomsdataconsortium.com/678j9012-j05g-67h8-f901-971069519555

Demonstrates that the dataset under consideration contains data from four other entities, which requires additional scrutiny of the sources in determining whether to acquire and use the dataset on offer.

P2 Source

This field identifies whether the data originates from a different organization than the one who issued the dataset, and identifies that original source's legal name. In this field, enter the name of the original data source if different from the issuer. For example, 'ABC Data Services'.

- [ ] Source is the same as Data Issuer

Legal entity name
International Customs Data Consortium

Legal entity address
123 High Street, 4th Floor, London, W1B 3HH, United Kingdom

Demonstrates that the existing data did not originate with the supplier, but another entity. The single value listed and the previous four URLs for metadata of other datasets which fed the current dataset, indicate that all data on offer originated with the International Customs Data Consortium.

P3 Data origin geography

REQUIRED

The geographical location where the data was originally collected, which can be important for compliance with regional laws and understanding the data's context. In this field, indicate the geographic location, by selecting the continent, followed by country, state and city from where the data originated. If there are multiple locations, indicate that by adding additional rows. If the data geography is highly diverse, enter worldwide as the value.

Europe → Netherlands
Europe → Switzerland
Europe → UK

Referring to the metadata concerning the range of dates for data generation, it suggests that the dataset was compiled immediately after the conclusion of the data collection period, ensuring its timeliness and relevance.

P4 Dataset issue date

REQUIRED

This field must contain the date when the dataset was compiled or created, providing a temporal context for the data. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

2024-02-01:T11:15:10

This is the first time the dataset is made available and there is no sense of the frequency at which it may be updated.

P5 Date of previously issued version of the dataset

This field is intended to capture the release date of the last version of the dataset, if it has been updated or revised, to track changes and updates over time. Provide the date in the ISO8601 standard format, for example, '2024-01-01T00:00:00'.

- [ ] Dataset has not been issued previously

YYYY-MM-DD:THH:MM:SS

P6 Range of dates for data generation

These fields represent the span of time during which the data within the dataset was collected or generated, offering insight into the dataset's timeliness and relevance. This field helps users understand the dataset's timeliness. You should indicate the date range, for example, '2024-01-01T00:00:00' to '2025-01-01T00:00:00'.

Oldest component of data contained in the dataset.
2020-01-01:T00:00:01

Youngest component of data contained in the dataset.
2024-01-31:T23:23:59

Reinforced the freshness of the data and that it was made available within a business date of when the collection period ended. The exact timestamp indicates that a full year's data is represented in the dataset.

P7 Method

REQUIRED

This field must account for the methodology, or procedures used to collect, generate, or compile the data, giving insight into its reliability and validity. Understanding the methodology is crucial for assessing data reliability. Describe the data collection method by selecting values from the drop-down options, for example, 'Survey conducted via online questionnaires'.

General use category
Feeds

Specific use
Other

Value
Automated Customs Entry Processing

Supports assessment of the data's reliability and suitability for training AI models, supporting that the dataset's generation aligns with Navisphere's standards for data quality and integrity.

P8 Data format

REQUIRED

This field describes the nature of the data within the dataset, such as numerical, textual, or multimedia, helping users understand what kind of information is contained within the file or dataset. Knowing the data format helps users prepare for how to handle the data. In this field you should specify the data format. For example, 'application/json'.

General format
Application

Specific format
Other

Specify other
vnd.oasis.opendocument.database

Helps Maya anticipate the structure and format of the data, ensuring compatibility with Navisphere's systems and facilitating efficient data processing and integration into AI models.

U Use

U1 Confidentiality classification

REQUIRED

This field must specify the level of sensitivity assigned to the dataset, such as personally identifiable information, which dictates how the dataset must be secured and who can access it. Proper classification ensures data is handled appropriately. Use the dropdown menu to indicate the confidentiality level, for example, 'Personally Identifiable Information, PII'.

Select which of the following are known to be present within the dataset:

- [ ] None
- [x] Personal Information (PI)/Demographic
- [ ] Payment Card Industry (PCI)
- [ ] Personal Financial Information (PFI)
- [ ] Personally Identifiable Information (PII)
- [ ] Personal Health Information (PHI)
- [ ] Sensitive Personal Information (SPI)
- [ ] Other

U2 Consent documentation location

REQUIRED

Specifies where the consent documentation or agreements related to the data can be found, ensuring legal compliance and regulatory use. Documenting consent ensures compliance with data protection regulations. Provide the location of consent documents. For example, 'https://data-provider.com/consent/12345'. If the data contained within the dataset is not personal in nature and does not require a consent for use, ensure the provided checkbox is checked.

- [x] Consent has been collected

Consent location
https://example.com/dataset/UUID-1234-5678-9012-3456/consent1.html

The consent documentation specifies the scope, purpose, and conditions under which the tariff data was collected and its permitted uses. Dr. Hicks' legal team confirms that this consent includes provisions for sharing the data with third parties and utilizing it for refining AI systems to predict tariff costs.

U3 Privacy enhancing technologies

REQUIRED

Were privacy enhancing technologies (PETs) or tools applied to the dataset to remove, mask, or modify PII/SPI in the data? Indicate whether techniques were used to protect personally identifiable information (PII) or sensitive personal information (SPI), highlighting the dataset's privacy considerations. This field ensures that privacy measures are clearly communicated. State if PETs were used and describe them. For example, 'ARX' could be the name of the PETs tool used to apply differential privacy techniques to the dataset. You should specify if noise addition was used to obfuscate individual data points, ensuring that the privacy of personally identifiable information (PII) and sensitive personal information (SPI) is maintained.

Specify tool
Clover DX

Parameter	Key value
Data masking	Injected 3% random data into the mix

Assures that the dataset adheres to privacy standards and regulations, which is critical for maintaining company AI practices and data handling within Navisphere Logistics.

U4 Data processing geography

This field defines the geographical boundaries within which the data can or cannot be processed, often for legal or regulatory reasons. Use this field to specify any geographic restrictions. For example, if the dataset must be processed within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

- [ ] Data processing geography is the same as data origin geography

U5 Data storage geography

This field specifies where the data is stored and any geographical restrictions on storage locations, crucial for compliance with data sovereignty laws. For example, if the dataset must be stored within the EU, you should specify all relevant EU countries and select the 'inclusion' option.

- [ ] Data storage geography is the same as data processing geography

U6 License to use

REQUIRED

Details the terms under which the dataset can be used, including any restrictions or obligations, clarifying legal use and distribution rights. Clear licensing terms ensure legal use and distribution of the dataset.

Select types of license

- [ ] None
- [ ] Non-commercial
- [ ] Public license
- [x] Commercial/Negotiated License

Location or point of contact
globaltradedatahub.com/license/123e4567-e89b-12d3-a456-426614174000

Defines restrictions, obligations, and rights around the legal use and distribution of the data, which stipulate that the data cannot be shared with global federal government entities and the data may be used for evaluation and training in AI-driven projects.

U7 Intended data use

REQUIRED

This field must describe the purpose for which the dataset was created, guiding users on its intended use and potential applications against identified use cases. Stating the intended use helps users understand the dataset's purpose. You should describe the intended use, for example, 'AI, pre-training'.

Select from the following uses

Non-AI

- [ ] Staging/testing
- [ ] Production
- [ ] Quality assurance
- [ ] Other

AI

- [x] Pre-training
- [ ] Alignment
- [ ] Evaluation
- [ ] Synthetic data generation
- [x] Other

Specify other AI use type
Transfer learning

System fit
Feature reuse

Business case
Performance enhancement

The dataset is best used for AI pre-training and training purposes, which may command a premium fee.

U8 Proprietary data presence

REQUIRED

This field must indicate whether the dataset contains proprietary information that is owned by or exclusive to the organization, affecting how it can be shared or used. Knowing if data is proprietary affects how it can be shared. In these fields, indicate if the dataset contains proprietary data subject to copyright, trademark, or patent. For example, if a copyright were applicable, you should enter the contact information for the individual who can speak to the copyright requirements, such as 'Jonathan Reeves, Esq., Email: jreeves@example.com, Phone: +1-555-012-3456'. If no IP restrictions are in place, select 'None'.

Select types of proprietary data present

- [ ] None
- [x] Copyright

Location or point of contact
Jonathan Reeves, Esq.,
Email: jreeves@globaltradedatahublegal.com,
Phone: +1-555-012-3456

- [ ] Patent
- [ ] Trademark

The company has made a contact available for discussing intellectual property rights. There are copyright implications of the dataset, which may not make it appropriate for use in context of the harmonization tariff schedule use case should the capability be productized or offered for downstream consumer use.

Additional comments

Enter comments and notes here

Outcome

Through application of the Data Provenance Standards metadata for its global tariff schedule datasets, Navisphere Logistics, Ltd. has achieved a significant enhancement in the operational efficiency and accuracy of its AI-driven tariff prediction models. The outcome includes:

Improved data consistency and compatibility	By specifying the version used for the metadata, Navisphere ensured that all datasets adhered to a uniform standard, facilitating seamless integration and interpretation by the AI models, regardless of the data's origin or when it was collected.
Enhanced data identification and access	The establishment of a unique metadata identifier and a metadata unique URL for each dataset enabled easy identification, access, and reference, streamlining the data ingestion process for the AI systems, and reducing the time spent on data preprocessing.
Streamlined lineage and dependency tracking	The metadata location for datasets feeding the current dataset allowed Navisphere to efficiently manage data dependencies and lineage, ensuring that updates or corrections in source datasets could be rapidly propagated through the system, maintaining the accuracy and timeliness of tariff predictions.
Increased accountability and data integrity	Detailed metadata entries for the creator, source, and data origin geography provided clear accountability and context for the data, enhancing trust in the data's reliability and compliance with regional laws and international regulations.
Better data privacy and security measures	The application of privacy enhancing technologies (PETs) and the careful classification of data confidentiality ensured that personally identifiable information (PII) and sensitive personal information (SPI) were adequately protected, aligning with global privacy standards and ethical considerations in AI application.
Legal compliance	Detailed metadata on data processing and storage geographies, consent locations, and the license to use the data ensured that all AI operations remained within legal boundaries, respecting data sovereignty laws and consent agreements.