USE CASE 1

Healthcare Insurance Data Procurement

Persona

Name

Bella Ramirez

Role

Procurement Team Lead

Background

Bella holds a BS in Accounting and has over 15 years of experience in the healthcare industry, specializing in procurement. Before her current role, she worked in various capacities within healthcare organizations, including as a contracting specialist and an accounting auditor. This gave her a deep understanding of the technical and procurement of healthcare data.

Responsibilities

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

Use case

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.

Goals

- Assess metadata coverage: Bella prioritizes the evaluation of the dataset's metadata to ensure it includes essential information like the dataset title, unique metadata identifier, metadata location, and details about data origin and collection methods. This step is crucial for establishing the dataset's lineage, context, and usage restrictions, aligning with the company's data provenance standards.
- Ensure regulatory compliance: By collaborating with the legal department, Bella ensures the dataset for adherence to healthcare data regulations, focusing on confidentiality classification, consent documentation, and data processing and storage geographies. This provides the dataset's use will not breach any legal or ethical boundaries.
- Operational efficiency and integration: Bella validates the dataset's potential impact on operational efficiency by working with the data team. She meets with the team that assesses how well the dataset integrates with existing systems and whether it can provide the expected enhancements to the analytical models without significant overhaul or disruption.
- Strategic use and innovation: Beyond compliance and integration, Bella explores how the dataset can be used to develop innovative marketing strategies and improve customer trust. This involves touching base with the marketing team, which is focused on analyzing the dataset's intent and proprietary data presence to identify new opportunities for personalized customer engagement and service delivery.

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.

How the standards are used

For Bella to ensure the dataset under evaluation meets the standards required for her healthcare insurance company's analytical models, she would assess the following values within the specified metadata:

Version used for metadata	Bella checks for the specific version number of the metadata schema or standard used, ensuring it's the latest or a widely recognized version to maintain consistency and future-proof the dataset. The version used for metadata is "1.0.0" which indicates the dataset uses the third revision of the 1.0.0 version of the metadata standard, suggesting that it is up to date with current standards.
Dataset title/name	She evaluates whether the dataset's title is descriptive and precise, clearly reflecting the dataset's content and intended use, facilitating easy identification and retrieval. The title of the dataset is "2024 Comprehensive Patient Care and Insurance Claims Dataset", a descriptive title indicating the dataset's content and scope, including the year for currency.
Unique metadata identifier	Bella verifies the presence of a unique identifier, like a UUID, ensuring it is correctly formatted and unique to prevent any confusion or overlap with other datasets. A unique identifier, "UUID-1234-5678-9012-3456" ensures this dataset's metadata is distinguishable.
Metadata unique URL	She examines the URL's accessibility, ensuring it leads directly to a comprehensive metadata page that provides detailed information about the dataset, and checks that the URL is secure and reliable. She uses the provided value of https://example.com/dataset/metadata/UUID-1234-5678-9012-3456, which is a direct link to the dataset's detailed metadata page, providing ease of access for further information.
Metadata location for datasets feeding the current dataset	Bella assesses this to understand the dataset's lineage and dependencies, ensuring that the metadata for source datasets is easily accessible and well-documented. She uses https://example.com/metadata/sources, a URL that points to metadata for source datasets, establishing the lineage and dependencies of the current dataset.
Date of previously published version of the dataset	She looks for the date to track the dataset's evolution, ensuring that any updates or revisions are noted and the dataset's version history is clear. "March 15, 2023" indicates when the last version of the dataset was released, helping track updates over time.
Creator	Bella verifies the creator's identity, ensuring they are reputable and have the necessary expertise. This provides accountability and a reliable point of contact for inquiries. The creator is noted as the "National Health Data Systems," which Identifies the organization responsible for creating the dataset, adding a layer of trust and accountability.

Source (if different from Creator)	She checks the source(s) of the data, ensuring transparency and the integrity of the data collection process, significantly if the source differs from the dataset creator. "Nationwide Hospitals Systems, Insurance Providers Ltd." Are specified in the source, which indicates the original data sources, providing context for the data's origin.
Data origin geography	Understanding the geographical context of the data's origin is crucial for Bella, especially for compliance with regional laws and regulations and for the dataset's applicability to her company's customer base. The geographic location where the data was collected, relevant for regulatory and contextual reasons, is listed as "United States" for the country, "California" for the state, and "Arcata, Eureka, San Francisco" for the cities.
Dataset creation date	She confirms the dataset's creation date to assess its freshness and relevance to current analytical needs. The metadata lists "January 10, 2024" as the creation date value and indicates when the dataset was compiled, providing context for its currency and relevance of business cases.
Range of dates for data generation	Bella evaluates the time frame during which the data was collected, ensuring it is relevant to the current analytical models and reflects recent trends or patterns. "January 1, 2023 - December 31, 2023" dates that are provided show Bella the time during which the data was collected, highlighting the dataset's recency.
Method	Understanding the data collection or generation methodology is critical for assessing the dataset's reliability and validity. Bella scrutinizes this aspect thoroughly and passes the information on to the data team. "Database feed" is the value in the method metadata field, and in the specification field, Bella notes, "Electronic Health Records Extraction and Insurance Claim Processing Logs" is added. This information describes how the data was collected, providing Bella with insight into its reliability.
Content type	She assesses whether the dataset's content type (numerical, textual, multimedia, etc.) is compatible with the company's analytical tools and is suitable for the intended analyses. "SQL" as the value and "Structured Data - Numerical and Categorical" specifies the nature of the dataset, which is crucial for understanding how it can be analyzed.

Privacy-enhancing technologies (PETs)	Bella confirms whether PETs were applied to the dataset to protect personal data, ensuring the dataset's compliance with privacy regulations and ethical standards. She notes that the metadata indicates a "No", which confirms that no measures have been taken to protect personal information within the dataset.
Organizational content classification	She ensures the dataset's classification aligns with the company's data handling policies, assessing whether its classification is appropriate and clear. "Restricted" indicates the data handling requirements and access restrictions, which is a flag to Bella that the dataset has protected health information (PHI) under the Health Insurance Portability and Accountability Act (HIPAA) and that the medical information must be carefully managed under provincial, state, or other healthcare privacy laws.
Confidentiality classification	Bella examines the level of sensitivity assigned to the dataset, ensuring it is adequately secured and that access is appropriately controlled based on its classification. "Private Health Information" reflects the dataset's sensitivity level and dictates security measures.
Consent location	Because the dataset involves confidential information, she verifies the location and adequacy of consent documentation, ensuring compliance with legal standards. https://example.com/dataset/UUID-1234-5678-9012-3456/consent1.html points to where consent documentation is kept, which Bella forwards to the legal department for review and determination of whether the consent meets the organization's legal obligations for processing.
Data processing geography	Bella assesses any geographical restrictions on where the data can be processed, ensuring compliance with legal and regulatory requirements regarding data processing locations. The metadata lists "United States" for the country and "California" for the state. Bella sees this as a flag since from experience, she knows that California's data privacy laws don't have geographic restrictions on data processing. She flags the information for legal department review.

Data storage geography	She checks for any geographical restrictions on where the data is stored, which is crucial for adherence to data sovereignty laws and protecting sensitive information. The metadata lists "United States" for the country and "California" for the state. Again, this is a flag for Bella as she knows that California's data privacy laws don't have geographic restrictions on data storage. However, she knows that additional protections require special handling of personal healthcare data, so she checks in with the legal department for a final determination.
License to use	Bella reviews the metadata in this category but passes along to the legal team information about the terms under which the dataset can be used, including any restrictions or obligations, to ensure the company's use of the dataset is legally sound and in line with the licensing terms. "License details available upon request from the Data Governance Department, National Health Data Repository, contactme@example.com" is listed and provides information on how to access terms of use. Bella sent an email asking for clarification about the license to use the data under the name "2024 Comprehensive Patient Care and Insurance Claims Dataset".
Intent	She evaluates the purpose of the dataset's creation, ensuring it aligns with her company's intended use and supports identified use cases without misalignment or misuse. The intent in the metadata indicates "AI" use and specifies "Pre-Training," which aligns with the use cases and requirements specified by the data team. Bella can gain insights from this information, along with the method metadata and dates generated, whether the data cost aligns with the organization's value assessment.
Proprietary data presence	Bella assesses whether the dataset contains proprietary information, ensuring its use does not infringe on intellectual property rights and is consistent with contractual agreements. All the values are empty, which indicates to Bella that the data is free to be used by her company without infringing on proprietary rights.

Outcome

Bella met with the legal department and the data team, to incorporate their assessment into the procurement analysis package. After investigating, the legal team determined that the data supplier mistakenly tagged the data processing and storage metadata incorrectly and had to correct the metadata associated with the dataset before the procurement process could proceed. This delayed the data procurement process by four business days.

However, by successfully evaluating and integrating the new dataset, Bella ensures that the organization is well-positioned to positively impact the company's business operations, including enhanced analytical capabilities, improved customer trust, and the development of responsible AI applications. This will align with the company's business considerations and set a benchmark for efficient and trustworthy data usage in the healthcare insurance industry.

USE CASE 2

Media Consumption Pattern Dataset for Consumer Behavior Insights

Persona

Name

Jordan Liu

Role

Data Strategy Director

Background

Jordan holds an MA in Data Science with a focus on market analytics. With a decade of experience in media analytics, Jordan has developed a deep understanding of the media consumption landscape. His career began in data analysis, evolving into strategic roles that leverage data insights to drive industry innovation, particularly in media consumption patterns.

Responsibilities

- Overseeing the development and distribution of comprehensive media consumption datasets.
- Ensuring datasets adhere to the latest data provenance standards for transparency and reliability.
- Collaborating with stakeholders across healthcare, consumer goods, and travel industries to tailor data offerings.

- Guiding the integration of datasets into client systems to optimize targeted content delivery and marketing strategies.
- Advocating for data-driven decision-making within the company and among clients to foster industry innovation.

Use case

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.

Goals

- Ensure comprehensive coverage of media consumption patterns to provide actionable insights for diverse industries.
- Maintain high standards of data transparency to build trust and encourage collaboration.
- Enhance clients' operational efficiency and compliance through strategic data integration.

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.

How the standards are used

For Jordan to ensure the dataset his company is offering to buyers meets the standards required for AI analytical models, he fills out the following metadata associated with the dataset he has curated:

Version used for metadata	Jordan ensures the dataset utilizes version 1.0.0 of the data provenance standards, thereby future proofing the metadata and making it backwards compatible, especially with systems that automate metadata ingestion.
Dataset title/name	He titles the dataset "March 2024 Global Media Consumption Trends," which reflects the data contents and production time frame.
Unique metadata identifier	Jordan generates a UUID and assigns "550e8400-e29b-41d4-a716-446655440000" as the unique identifier, to ensure identification within the data ecosystem.
Metadata unique URL	Jordan enters "example.com/ 550e8400-e29b-41d4-a716-446655440000/metadata.html as the URL to be used for direct access to the data provenance standards metadata, ensuring transparency and easy reference by downstream consumers assessing the data for consumption.
Metadata location for datasets feeding the current dataset	Jordan outlines the lineage and sources contributing to the dataset for comprehensive understanding and traceability by entering three values, "example.com/ 550e8400-e29b-41d4-a716-44665543902, example.com/ 550e8400-e29b-41d4-a716-44665544732, and example.com/ 550e8400-e29b-41d4-a716-446655465722" as the values.
Date of previously published version of the dataset	This metadata field documents the date, if applicable, when the dataset was previously published and allows downstream consumers to track dataset evolution and updates. Jordan leaves the field blank as this dataset is published for offer for the first time.
Creator	The market analysis company name, "AnalytiQuest Ventures" is listed by Jordan, attributing dataset ownership and responsibility for its integrity.
Source (if different from Creator)	Since AnalytiQuest Ventures is the generator of the data, Jordan leaves the field blank; the creator of the dataset already accurately depicts the source of the data as AnalytiQuest Ventures.
Data origin geography	This metadata field identifies the geographic data points that are included to contextualize the media consumption patterns, and so Jordan enters the following values for the metadata in order of requirements – country, state, city: United States, Florida, Miami; United States, Florida, Ft. Lauderdale; United States, Florida, Orlando; United States, Florida, Clearwater; United States, Florida, St. Petersburg; United States, Florida, Tampa; United States, Florida, Pensacola; United States, Florida, Augusta; United States, Florida, Jacksonville; United States, Florida, Cape Coral.

Dataset creation date	The dataset creation date indicates the dataset's compilation date, ensuring relevance. Jordan specifies "January 10, 2024" as the creation date, indicating when the dataset was compiled. This date provides a context for the dataset's freshness.
Range of dates for data generation	Jordan enters "January 1, 2023 - December 31, 2023" as the period during which the data was collected. This timeframe is crucial for ensuring the dataset reflects the latest media consumption patterns and is relevant for current analysis.
Method	The data collection methodology, described as "User Generated Content: Digital Interaction Tracking and Survey Responses" in the method metadata field, outlines how consumer interactions with various media platforms were recorded alongside targeted survey data. This method provides a comprehensive view of media consumption behaviors, enhancing the dataset's transparency for in-depth consumer insights.
Content type	Jordan enters .xls (30%), .doc (40%) and .sql (30%) to depict the structured and unstructured data means used to gauge multimedia engagement metrics and the textual Responses that characterize the dataset's content. This mix is essential for analyzing both quantitative media engagement metrics and qualitative consumer feedback, offering a multifaceted approach to media consumption analysis but also points to a potential level of cleanup that may be required on the dataset.
Privacy-enhancing technologies (PETs)	Jordan selects the metadata's indication of "Yes" to denote the use of PETs, and specified "Anonymization and Data Aggregation Techniques Applied" via "Adverity" tool, which confirms that steps have been taken to anonymize and aggregate personal data, ensuring the dataset's adherence to privacy standards and ethical considerations in media research.
Organizational content classification	Jordan identifies the dataset as being for "Internal Use", and this classification signals the dataset's designed purpose for in-house analytics and strategic planning, aligning with corporate data governance policies and ensuring appropriate handling within the organization.
Confidentiality classification	This is a common field that Jordan uses in his work, when there is personally identifiable information on consumers present. However, since PETs have been used to aggregate and anonymize the personal data in the dataset, this classification can be left blank and Jordan does so.

Consent location	Since PETs have been used to aggregate and anonymize consumer personal data in the dataset, no consent is required to process or share the consumers' data and Jordan leaves this field blank.
Data processing geography	Since PETs have been used to aggregate and anonymize consumer personal data in the dataset, there are no limitations on where the dataset contents can be processed and Jordan leaves this metadata field blank.
Data storage geography	Since PETs have been used to aggregate and anonymize consumer personal data in the dataset, there are no limitations on where the dataset contents can be stored and Jordan leaves this metadata field blank.
License to use	Jordan knows that the license terms are flexible for this dataset, so rather than specifying the terms, the AnalytiQuest Ventures's Office of General Counsel should be contacted for usage details. He enters "AnalytiQuest Ventures's Office of General Counsel, legalconsumptionlicense@example.com and (555) 123-4567" into the metadata field.
Intent	The purpose of this dataset, as Jordan labels it, is "AI", "Evaluation", and "Training", which aligns with strategic objectives in media planning and content development, as well as research endeavors in the media industry.
Proprietary data presence	Jordan leaves this metadata field blank. The absence of proprietary restrictions, as indicated by the metadata, confirms the dataset's availability for broad analysis within the stipulated legal frameworks, facilitating unrestricted exploration of media consumption trends.

Outcome

Under Jordan's guidance, the "March 2024 Global Media Consumption Trends" dataset emerges as a good resource for understanding intricate media consumption behaviors across various platforms. By curating and documenting the dataset's metadata, including the adoption of version 1.0.0 for data provenance standards, assigning a unique identifier, and providing a transparent metadata URL, Jordan ensures the dataset's integrity and usability for AI analytics. This attention to detail, coupled with the clear documentation of data origin, collection methodologies, and privacy-enhancing measures, positions the dataset as a trustworthy and comprehensive tool for media buyers and sellers. The dataset's rich insights into consumer behaviors, derived from diverse geographical regions and articulated through a mix of structured and unstructured data types, empower stakeholders across multiple industries to tailor personalized content strategies effectively. Jordan's emphasis on data transparency and legal

requirements fosters collaboration but also enhances operational efficiency and compliance for clients, setting a new benchmark for data-driven decision-making in media consumption.

USE CASE 3

Financial Services Customer Product Enablement

Persona

Name

Minh Quang Nguyen

Role

Data Architecture and Policy Analyst

Background

Minh is a seasoned Data Architect and Policy Analyst at ProForma Financial Services. With a Master's degree in Data Science and over a decade of experience in data management and policy development, Minh has become a pivotal figure in the company's strategic planning department. Minh's expertise extends to data governance, where he leads initiatives to improve data quality and accessibility across departments. He is a strong advocate for data-driven decision-making and often conducts workshops and training sessions for employees to enhance their data literacy.

Responsibilities

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

Use case

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.

Goals

- Improve the precision of AI models used in tailoring customer credit card products, leading to more personalized and effective offerings.
- Confirm that the dataset's use aligns with international regulations, including GDPR, safeguarding against legal and reputational risks.
- Maintain the highest standards of data privacy and security, particularly for personally identifiable information (PII) and sensitive personal information (SPI), through the application of Privacy Enhancing Technologies (PETs).
- Streamline data processing and storage practices to enhance efficiency while staying within the bounds of data processing and storage geography restrictions.
- Provide clear documentation of the dataset's origins, methodologies, and purposes to uphold transparency and accountability standards.
- Ensure the dataset's quality and integrity by verifying its collection methods, update history, and content type, thereby fostering trust in the AI-driven insights derived from it.

Challenges

- Understanding the dataset's lineage and metadata to ensure its origin, collection methods, and processing are well-documented and credible.
- Navigating complex and varied international data regulations, particularly concerning data privacy, AI deployment, and data sovereignty.
- Seamlessly integrating the new dataset with the company’s existing data architecture and AI systems without disrupting ongoing operations.
- Balancing the use of proprietary data within the dataset with the need to protect sensitive information and maintain competitive advantages.
- Verifying that the dataset's use is backed by proper consent and aligns with ethical standards, especially when dealing with sensitive customer data.

- Staying ahead of rapid technological advancements and evolving data standards to ensure the dataset remains relevant and effective for AI applications.

How the standards are used

Minh is assessing a new dataset's compliance with data provenance standards and detailed metadata to refine AI algorithms focused on customer credit card offerings. His review focuses on ensuring the dataset's integrity, transparency, and compliance with international regulatory requirements, especially considering the implications of GDPR, but also with an eye towards the new EU AI Act. Minh considers the following aspects of the metadata:

Version used for metadata	Minh checks that the metadata described is aligned with the most current schema version, "v1.0.0".
Dataset title/name	He reviews the dataset title "Consumer Spending Patterns 2020-2024", which clearly reflects the dataset's focus.
Unique metadata identifier	Minh verifies the dataset's unique identifier, "LFS-1234-5678", to avoid any confusion with other datasets.
Metadata unique URL	He accesses the metadata through its URL, "http://luminadataservices.com/metadata/1234-5678", providing a direct pathway to detailed dataset information.
Metadata location for datasets feeding the current dataset	Minh is able to review the metadata for source datasets, such as "Retail Transaction Records 2023-2024", as they are found at "http://luminadataservices.com/metadata/sources/retail-transactions-2023" and "http://luminadataservices.com/metadata/sources/retail-transactions-2024".
Date of previously published version of the dataset	Minh notes the last update was on "March 15, 2023", indicating recent revisions.
Creator	Minh reviews the creator metadata field. The dataset is credited to the "Lumina Financial Services", establishing accountability.
Source (if different from Creator)	He then confirms the original data sources, "Global Retail Partners Consortium" and "PreciTech Data Inc.", differentiating the sources from the dataset creator.

Data origin geography	The data originates from "Europe, France; Europe, Germany;Europe, Italy;; Europe, Poland" which is important for compliance considerations.
Dataset creation date	Minh reviews the dataset creation date range, which is reflected in the metadata as "March 14, 2024". The data is recent, which is a great sign for Minh's use case but he needs to understand the range of data generation to verify that it is not stale and that it has been collected recently reflective of legal and regulatory data privacy requirements.
Range of dates for data generation	Minh then reviews the dataset generation date range, which is reflected in the metadata as "January 5, 2023 through March 14, 2024", providing context for the data's recency and relevancy for establishing consumer trends.
Method	The methodology, "Feeds, Interval timed database info, Aggregated Consumer Transaction Analysis, PoS", is reviewed for data collection integrity. Minh notes that 100 percent of the data was received in this structured format, which is what he expected and knows to be relatively clean data requiring minimal pre-processing for his data ingestion needs.
Content type	Minh notes that the dataset contains "application/sql", which reflects the Oracle database source he is accustomed to seeing and knows to be suitable data for AI modeling.
Privacy-enhancing technologies (PETs)	Minh reviews the PETs metadata, noting the presence of a "Yes", indicating that privacy concerns have been addressed through data anonymization. "Google differential privacy library" is listed as the tool and "Differential privacy" is listed as the method used for applying PETs. This signals to Minh that confidential data is unlikely to be present and that consent for further consumer data processing is not required due to the PETs application.
Organizational content classification	Labeled as "Internal Use", this metadata value for content classification will be guiding Minh on access limitations.
Confidentiality classification	This metadata value was not presented to the data supplier for completion as PETs were identified as being applied in a previous metadata field.
Consent location	This metadata value was not presented to the data supplier for completion as PETs were identified as being applied in a previous metadata field.
Data processing geography	This metadata value was not presented to the data supplier for completion as PETs were identified as being applied in a previous metadata field. As a result, Minh knows that no data processing limitations are in place for protection of personal data.

Data storage geography	This metadata value was not presented to the data supplier for completion as PETs were identified as being applied in a previous metadata field. As a result, Minh knows that no data localization is in place for storage of personal data.
License to use	He checks the license terms at "http://luminadataservices.com/license/1234-5678", confirming usage rights. Since he is unfamiliar with some of the licensing clauses expressed in the documentation, Minh forward the license terms to the General Counsel's Office for review and analysis. There have been some issues with pricing of data coming from the EU recently, and he wants to ensure a complete legal picture is assessed before proceeding with a decision to acquire this dataset.
Intent	Minh reviews the dataset metadata intent, and notices that the data is intended for "AI, Other, Enhancing AI-driven Credit Card Offerings", aligning with his project goals.
Proprietary data presence	Minh notes that in the proprietary metadata field, the data supplier has indicated "No" to the presence of copyright, trademark or patent presence, indicating possible exclusivity to his company of whatever AI modeling outcomes may be obtained, which could offer competitive advantages.

Outcome

Minh's review of the metadata for the "Consumer Spending Patterns 2020-2024" dataset results in advancements in ProForma Financial Services' AI algorithms for customer credit card offerings. By not relying on high level descriptions of the dataset offered by the data supplier and instead reviewing the standards and metadata, Minh increased the chances for strategic success. He verified the dataset's compliance with the latest data provenance standards, including a review of its versioning, unique identifiers, and comprehensive metadata URLs. Minh thus ensured the dataset's integrity and alignment with international regulations. His attention to the dataset's lineage, original sources, and the application of Privacy Enhancing Technologies (PETs) helped meet the data privacy requirements set by his company and will mitigate potential legal and reputational risks associated with GDPR and the EU AI Act blunders.

The detailed metadata, including data origin geography, creation dates, and collection methodologies, provided Minh with the assurance of the dataset's relevance and quality. The absence of proprietary data restrictions, coupled with clear licensing terms, positions ProForma to leverage this dataset for creating more personalized and effective customer credit card products. Minh's approach to dataset integration will enhance operational efficiencies going forward, ensuring seamless compatibility with the company's existing data architecture and AI

systems.

Overall, Minh's review of the metadata to ensure alignment with requirements mean that ProForma Financial Services can harness AI-driven insights responsibly and innovatively, paving the way for data-driven product enablement and a competitive edge in the financial services sector.

USE CASE 4

Enhancing Global Logistics Efficiency Through AI-driven Tariff Harmonization

Persona

Name

Dr. Maya Hicks

Role

Lead Data Scientist

Background

Dr. Hicks holds a Ph.D. in Data Science with a specialization in artificial intelligence and machine learning. With over a decade of experience in the logistics industry, she has a keen interest in optimizing supply chain efficiency through innovative technologies. Maya is known for her analytical mindset and her ability to translate complex data insights into actionable business strategies.

Responsibilities

- Lead the AI research and development team in refining and enhancing the company's AI-driven tariff prediction models.
- Evaluate datasets for integrity and compliance with corporate policies and standards which reflect international regulations and privacy considerations.
- Collaborate with procurement and legal colleagues to ensure that the data and AI models are in line with global standards and regulations.
- Train and optimize AI models to accurately predict tariffs, involving sophisticated algorithms and machine learning techniques.
- Integrate the refined AI models into Navisphere Logistics' operational systems and conduct extensive testing to ensure accuracy and efficiency.
- Establish and maintain a feedback loop for continuous monitoring and improvement of the AI models based on real-world application insights.

- Ensure the responsible use of AI in accordance with Navisphere Logistics' standards and privacy laws, particularly in the handling of sensitive data.
- Communicate the progress and outcomes of the AI enhancements to stakeholders, including technical teams, management, and commercial clients.
- Stay updated with the latest developments in AI, machine learning, and international logistics practices to continually drive innovation within the company.

Use case

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.

Goals

- Harmonize global tariff schedules into a unified, AI-friendly format to enhance prediction accuracy.
- Refine and improve the AI-driven tariff prediction models to minimize cross-border delivery times and costs.
- Ensure that all collected tariff data meets stringent data provenance standards for integrity and compliance with international regulations.
- Achieve high accuracy in tariff predictions across different countries and product categories through sophisticated AI algorithms.
- Streamline customs clearance processes through more precise tariff assessments, benefiting the company's worldwide commercial clientele.

Challenges

- Navigating the complex web of international tariffs and customs regulations, each with its own classification system and rules.
- Meticulously evaluating the metadata for each dataset to ensure compliance with international standards, including data origin, collection methodology, and privacy considerations.
- Adapting to constant changes in international tariff regulations, requiring continuous updates to the AI models.

- Balancing the need for advanced AI capabilities with responsible use of AI, adhering to company standards and international privacy laws.
- Ensuring that the AI models can be seamlessly integrated into Navisphere Logistics' operational systems without disrupting existing workflows.

How the standards are used

Dr. Hicks leverages the metadata associated with global tariff schedule datasets to ensure the accuracy and reliability of AI-driven tariff prediction models, essential for optimizing logistics operations. The metadata, including the data's origin, collection methodology, and privacy considerations, enables her to assess the trustworthiness and relevance of the data for her analyses. This approach to metadata evaluation forms the foundation of Maya's ability to build and maintain robust, transparent, and compliant AI systems within Navisphere Logistics, Ltd.

Version used for metadata	Maya uses this attribute to ensure that the dataset conforms to the latest standards for metadata documentation, which is crucial for compatibility with Navisphere's AI systems. She evaluates the metadata schema value “1.4.5" and determines it is backwards compatible with the version Navisphere is using.
Dataset title/name	This metadata helps Maya quickly understand the dataset's focus and relevance to her needs. The title she evaluates is "2023 Global Tariff Schedules - Electronics".
Unique metadata identifier	Along with the dataset title, Maya uses the unique metadata identifier to uniquely identify and reference datasets without confusion, especially when dealing with multiple sources. She uses the "123e4567-e89b-12d3-a456-426614174000" value which she confirms in the company’s data procurement system has not previously been considered nor procured by Navisphere.
Metadata unique URL	This metadata component provides Maya direct access to detailed dataset information for deeper evaluation. She uses "https://globaltradedatahub.com/metadata/123e4567-e89b-12d3-a456-426614174000".

Metadata location for datasets feeding the current dataset	Maya relies on this metadata component to trace data lineage and verify the integrity of source data. She notes five values, denoting that five different datasets supplied the data contained in the set she is evaluating. The values she is presented with are: • "https://internationalcustomsdataconsortium.com/metadata/234f5678-f01c-23d4-b567-537625175111 • https://internationalcustomsdataconsortium.com/metadata/345g6789-g02d-34e5-c678-648736286222 • https://internationalcustomsdataconsortium.comm/metadata/456h7890-h03e-45f6-d789-759847397333 • https://internationalcustomsdataconsortium.com/metadata/567i8901-i04f-56g7-e890-860958408444 • https://internationalcustomsdataconsortium.com/678j9012-j05g-67h8-f901-971069519555"
Date of previously published version of the dataset	Maya can track via this field any dataset updates and revisions. The "N/A" value signifies that the dataset has not previously been published.
Creator	Knowing the creator provides accountability and a point of reference for Maya, since she can check what other datasets the company has previously procured from the same data supplier and satisfaction reviews of the supplier's historically provided data. "GlobalTradeDataHub" is reflected as the creator.
Source (if different from Creator)	Maya also reviews the source metadata for the dataset, noting that it is different from the creator. She notes the value of "International Customs Data Consortium" which again, helps Maya determine the reliability and historical satisfaction of supplied data. Together with the metadata location for datasets feeding the current set metadata, Maya can tell that all data in the current dataset originated with International Customs Data Consortium and that GlobalTradeDataHub was merely the curator or entity that put the datasets together to provide the current one on offer.

Data origin geography	Maya reviews the data origin geography, noting “Europe, Switzerland; Europe, United Kingdom; Europe, Netherlands” as the values and notes that if the dataset contains personal data, additional regulatory requirements will apply. If personal data is present, she will also need to perform additional pre-processing of data to anonymize the data, which will require legal review and will extend her project delivery date.
Dataset creation date	Maya seeks context on the data's recency and relevance and is satisfied to see the "February 1, 2024” creation date indicating that tariff schedules that took place with the first of the year are likely to be included in the set. However, the range of dates for data generation will confirm for her this fact, so she reviews that metadata next.
Range of dates for data generation	This helps Maya assess the dataset's timeliness. The range is "January 1, 2020 to January 31, 2024". Based on this information, Maya is satisfied that the date range covers new tariff schedules that were rolled out at the start of the current year. However, Maya already has data in house for the time period of January 1, 2020 through December 31, 2022, which means she will be paying to acquire data that already exists in the enterprise. Maya confirms with her procurement team and determines that negotiations should exclude the initial two years of data and the price of the dataset should reflect that adjustment.
Method	Understanding the collection method helps Maya judge the potential need for cleanup of data prior to its use. She notes “Feeds, Other, Automated Customs Entry Processing", which implies a highly structured format and the automation further points to patterns she can use to detect anomalies in the cleanliness of the data.
Content type	The content type metadata informs Maya about the kind of information the dataset contains, aiding in data parsing. She notes the value of “application/vnd.oasis.opendocument.database”. This media type is used for database files created with software that adheres to the OpenDocument standards, such as LibreOffice Base or Apache OpenOffice Base. Maya is confident around the reliability of the data for her needs.

Privacy-enhancing technologies (PETs)	Maya reviews the metadata reflecting the use of PETs to ensure data privacy compliance, and whether she will need to anonymize data prior to its use. She notes the "Yes" metadata, which confirms that personal data is not present in the dataset. “Clover DX” is listed as the toolset and “Anonymization” is listed as the method. Maya also notes in the outcome field “injected 3% random data into the mix” which signals to her an appropriate level of noise, as the company considers anything above 5% unacceptable.
Organizational content classification	This metadata value guides Maya on how the dataset can be used within Navisphere. "Internal Use Only" denotes there is no sensitive information contained in the dataset, but there are reasons for restrictions, most likely legal in nature. Maya will look at the license information and work with the legal team to determine what the limitations might be.
Confidentiality classification	Because PETs were used and declared in the previous metadata, this metadata field is grayed out for the data supplier when assigning metadata and Maya doesn’t see any values presented.
Consent location	Because PETs were used and declared in the previous metadata, this metadata field is grayed out for the data supplier when assigning metadata and Maya doesn’t see any values presented.
Data processing geography	Because PETs were used and declared in the previous metadata, there are no restrictions on the data processing geography. Maya notes “Included” and “Worldwide” as location, which means that the data can be used without additional protections or limitations on geographical processing.
Data storage geography	Because PETs were used and declared in the previous metadata, there are no restrictions on the data processing geography. Maya notes “Included” and “Worldwide” as location, which means that the data can be used without additional protections or limitations on geographical storage.
License to use	Maya notes that the license applicable to the dataset is available at “globaltradedatahub.com/license/123e4567-e89b-12d3-a456-426614174000”. While she scans the license information, she knows that she will need the legal department's eyes and sign off before proceeding with the purchase of the dataset. She routes the license, along with her question around specific data content classification to her colleagues in the legal department for consideration and approval.

Intent	Maya reviews the dataset intent, and notices that the data is intended for "AI, Training", which aligns with her project goals. The declaration also signals to Maya that the premium price placed on the dataset is appropriate because it is intended and expected to be of higher quality than data intended for other uses.
Proprietary data presence	Maya reviews the metadata around proprietary data presence and notes that there is a contact listed for the copyright. She flags the values "Jonathan Reeves, Esq., Email: jreeves@globaltradedatahublegal.com, Phone: +1-555-012-3456" for her legal colleagues to use in clarifying the copyright application and confirming whether the company should use the dataset, or whether the copyright will limit commercial applications once the AI is trained with the data input.

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 data 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.