# HSBC Data Controls - AI Use Case Documentation

This document contains the detailed Confluence page templates for the six AI use cases being pursued within the Data Controls space.

# Confluence Page 1: AI Use Case: AI-Powered Policy Engine for Legislative Compliance

Page Title: AI Use Case: AI-Powered Policy Engine for Legislative Compliance

Parent Page: Data Controls AI Initiatives

Tags: data-privacy, ai, policy-engine, opa, odrl, governance-as-code

## 1. Executive Summary

This initiative details the development of an AI-Powered Policy Engine designed to fundamentally transform how HSBC interprets, operationalizes, and demonstrates compliance with complex data privacy legislation. The solution will leverage AI (specifically Natural Language Processing and generative models) to analyze new and existing legislation (e.g., GDPR, CCPA, and new jurisdictional rules) and automatically translate dense, ambiguous legal text into structured, machine-readable rules.

These rules will adhere to the W3C Open Data Rights Language (ODRL) ontology to define data rights in a standard way and will be compiled into executable policies for the Open Policy Agent (OPA) framework. This "governance-as-code" approach will enable real-time, preventative enforcement of data policies across the bank's critical applications and data flows, moving compliance from a manual, reactive process to an automated, proactive one.

## 2. Problem Statement

* **Manual Interpretation is Slow and High-Risk:** Translating dense legal and regulatory text into actionable IT controls is a burdensome, manual process. It requires extensive, high-cost collaboration between legal experts, compliance officers, and senior IT architects. This process is not only slow but also carries a high risk of misinterpretation or incomplete implementation.
* **Time-to-Compliance Lag:** The global regulatory landscape is volatile. There is a significant lag—often months—between a new regulation being published (or an old one updated) and our systems being fully configured to enforce it. This creates a persistent, measurable window of compliance risk.
* **Inconsistent Application:** Manual interpretation, often duplicated across different business lines and jurisdictions, inevitably leads to inconsistent application of rules. The same legal clause might be implemented differently in the UK versus Hong Kong, creating gaps, conflicts, and significant challenges for enterprise-wide risk reporting.
* **Auditability Challenges:** Demonstrating end-to-end compliance is a forensic, manually intensive effort. Answering a simple regulatory query like "Show me how you enforce Article 17 of GDPR for customer data in your wealth management platform" can take weeks, requiring teams to manually trace legal opinions, control documents, and system configurations.

## 3. Proposed AI Solution

The solution is a two-part, closed-loop system:

1. **AI Legislation Interpreter (The "What" and "Why"):** An AI model, likely a fine-tuned LLM trained on a corpus of global financial regulations and HSBC's internal policy framework.
   * **Ingestion:** It will ingest new legislation, legal opinions, and regulatory updates.
   * **Analysis:** It will use NLP to perform semantic analysis, identifying and extracting key elements: specific obligations (e.g., "must gain explicit consent"), data types involved (e.g., "biometric data"), actors (e.g., "data processor"), specific actions (e.g., "data deletion"), and penalties.
   * **Mapping:** It will map these obligations to HSBC's internal control framework and the Business Information Model (BIM), providing context.
2. **Rule Generator (The "How"):** This component will take the AI's structured interpretation and automatically draft policies in machine-readable formats.
   * **ODRL (Open Data Rights Language):** It will first generate ODRL-based policies to define the high-level data rights, permissions, and prohibitions in a standard, shareable ontology. This creates a clear, semantic layer.
   * **OPA (Open Policy Agent) - Rego:** It will then "compile" these ODRL concepts into executable policy rules written in Rego, the language for OPA. These are the "governance-as-code" artifacts.
   * **Deployment:** These Rego policies can be deployed in real-time to the central OPA engine, which enforces them at critical integration points like API gateways, microservices, and data platform access layers, effectively preventing non-compliant actions before they happen.

## 4. Strategic & Non-Financial Benefits

This table quantifies the *strategic* value this use case provides to the Data Controls function and HSBC.

| **Benefit Category** | **Strategic Benefit Description** | **How We Will Measure This (KPI)** | **Strategic Alignment** |
| --- | --- | --- | --- |
| **Regulatory Agility** | **Accelerated Time-to-Compliance:** Drastically reduces the time from a regulation's publication to its enforcement. This allows HSBC to be "first to market" with compliance, turning a regulatory burden into a demonstration of robust governance. | *Metric:* Reduction in average days/weeks for the "policy implementation lifecycle" (from legal analysis to-live policy). | Demonstrates proactive and agile governance to regulators, building trust. |
| **Risk Reduction** | **Enhanced Compliance Accuracy:** AI-driven interpretation ensures a more consistent, comprehensive, and accurate application of rules, reducing the risk of human error, interpretive gaps, and subsequent fines or reputational damage. | *Metric:* Reduction in policy-related audit findings. Reduction in the number of control gaps attributed to misinterpretation. | Strengthens all "Three Lines of Defense" by providing a single, consistent source of truth. |
| **Operational Resilience** | **"Governance-as-Code":** Moves HSBC from a reactive, audit-based posture ("find and fix") to proactive, preventative controls ("prevent and prove") embedded directly into our systems. This makes compliance the path of least resistance. | *Metric:* % of data privacy controls automated via OPA vs. manual checks or after-the-fact audits. | Creates a more robust, resilient, and self-healing control environment. |
| **Scalability** | **Scalable Policy Management:** Provides a framework to efficiently manage and deploy data policies across thousands of applications and dozens of jurisdictions without a linear increase in legal and compliance headcount. | *Metric:* Number of applications/systems integrated with the central OPA engine. Policy change deployment time (from days to minutes). | Future-proofs our governance framework for business growth and M&A activity. |
| **Auditability** | **Transparent, Immutable Compliance:** Creates a clear, auditable, and immutable digital link from a specific legal article to the exact OPA rule enforcing it. Audit requests become simple, automated queries. | *Metric:* Reduction in time/effort (person-hours) required for regulatory compliance reporting and internal/external audits. | Improves transparency and trust with regulators, auditors, and partners. |

# Confluence Page 2: AI Use Case: Embedded AI Helper Agents in Data Visa

Page Title: AI Use Case: Embedded AI Helper Agents in Data Visa

Parent Page: Data Controls AI Initiatives

Tags: data-privacy, ai, agentic-ai, data-visa, data-quality, user-experience

## 1. Executive Summary

This initiative involves embedding an AI-powered "Helper Agent" directly within Data Visa, HSBC's internal Data Privacy governance tool. The primary function of this agent is to act as an intelligent, real-time co-pilot for users (e.g., project managers, data owners, architects) who are completing critical assessments. It ensures the data they input (e.g., for Data Privacy Impact Assessments, data transfer requests, or new system registrations) is accurate, complete, consistent, and compliant *at the point of capture*, long before it becomes a downstream risk.

## 2. Problem Statement

* **"Garbage In, Garbage Out":** The Data Visa tool is the "source of truth" for our data privacy governance. Errors or incomplete information entered at this stage lead to flawed privacy assessments, incorrect risk ratings, and a cascade of wasted remediation efforts. This flawed data can also lead to incorrect reporting to senior management and regulators.
* **User Error and Expertise Gap:** Users of Data Visa are often Project Managers or Data Owners, not full-time data privacy experts. They may misinterpret complex questions about data lineage, PII classification, or cross-border transfer, leading to good-faith but inaccurate submissions.
* **High Remediation Cost:** Identifying and correcting these input errors is a manual, time-consuming "detect and correct" process for the specialist Data Controls team. This re-work occurs *after* the flawed submission, meaning the risk has already been incorrectly documented and time has been wasted by all parties.
* **Poor User Experience:** A complex, high-stakes form without proactive guidance can be a frustrating experience for users. This can discourage thoroughness, lead to users "giving up" and submitting incomplete data, and foster a negative perception of the data governance function.

## 3. Proposed AI Solution

The AI Helper Agent will be an interactive, non-intrusive component within the Data Visa user interface. Its functions will be deeply integrated into the user's workflow:

* **Real-Time Validation & Cross-Referencing:** As a user types, the agent will cross-reference inputs against other HSBC systems of record (e.g., CMDB, Business Information Model, Data Catalog).
  + *Example:* A user selects 'No' for PII. The agent flags: *"You selected 'No' for PII, but the connected data source (DS-1234) is flagged as containing 'Customer PII' in the Data Catalog. Please verify or update the data source information."*
* **Contextual Guidance & Policy-Bot:** The agent will provide dynamic help and definitions based on the user's current task and data.
  + *Example:* A user hovers over "Cross-Border Transfer." The agent provides a pop-up with the bank's official definition and links to the relevant policy, or a "Chat with Policy" feature to ask specific questions.
* **Smart Suggestions & Auto-Completion:** The agent will learn from previously approved, high-quality entries to provide smart suggestions.
  + *Example:* *"We see you are registering an application for the 'Wealth Management' division. Similar, approved entries have classified this data as 'High Business Impact' and subject to 'KYC-001' retention. Would you like to apply these?"*
* **Error & Contradiction Prevention:** The agent will alert users to high-risk entries or logical contradictions *before* they can submit the form.
  + *Example:* *"You have indicated a data transfer to a non-approved jurisdiction without completing the required 'Data Transfer Risk' section. Please complete Section 4 before proceeding."*

*(Placeholder for a UI mockup of Data Visa with an AI chat assistant sidebar or contextual pop-ups)*

## 4. Strategic & Non-Financial Benefits

| **Benefit Category** | **Strategic Benefit Description** | **How We Will Measure This (KPI)** | **Strategic Alignment** |
| --- | --- | --- | --- |
| **Data Quality** | **Improved "First-Time-Right" Submissions:** Enforces data accuracy and completeness at the point of capture. This dramatically improves the quality and reliability of our central privacy governance data, building a trustworthy foundation for all other controls. | *Metric:* Reduction in % of Data Visa submissions requiring "re-work" or correction by the Data Controls team. | Embeds "Data Quality by Design" and "Right First Time" principles into a critical business process. |
| **Risk Reduction** | **Proactive Risk Identification:** The agent can identify high-risk proposals, data entry patterns, or policy conflicts in real-time, long before a manual reviewer would see them. It flags risk *at inception*. | *Metric:* % of high-risk entries proactively flagged by the AI agent vs. later manual discovery. Reduction in "time to discover" risk. | Strengthens the First Line of Defense by embedding expert controls directly with the user. |
| **Operational Efficiency** | **Reduced Remediation Overhead:** Frees up the highly-skilled Data Controls team from manually correcting simple input errors, allowing them to focus on high-value, complex risk analysis and strategic advisory. | *Metric:* Reduction in average time (person-hours) spent by Data Controls team on data validation and correction per submission. | Improves the productivity, scalability, and job satisfaction of the central governance team. |
| **User Experience & Adoption** | **Improved User Journey & Data Culture:** Acts as a co-pilot, making a complex compliance tool easier, faster, and less frustrating to use. This improves user satisfaction and promotes a stronger, shared data culture. | *Metric:* Improved user satisfaction scores (NPS/CSAT) for the Data Visa tool. Reduction in time-to-complete for users. | Fosters a culture of "doing the right thing" by making compliance the easiest and most-supported path. |

# Confluence Page 3: AI Use Case: AI-Driven Data Mapping and Metadata Enrichment

Page Title: AI Use Case: AI-Driven Data Mapping and Metadata Enrichment

Parent Page: Data Controls AI Initiatives

Tags: data-usage, ai, metadata-management, iso-11179, data-catalog, bim

## 1. Executive Summary

This initiative leverages AI to address the systemic, bank-wide challenge of incomplete, inconsistent, and un-standardized technical metadata. The solution will use AI models to scan, analyze, and automatically enrich our technical metadata (e.g., cryptic column names, empty table descriptions) to comply with the **ISO 11179-2023** standard for metadata registries.

A key function will be to automatically map and classify these technical data assets (tables, columns, files) to the **HSBC Business Information Model (BIM) taxonomy**. This will, for the first time, create a clear, enterprise-wide link between our technical data and its business meaning, enabling true data lineage, usage governance, and value realization.

## 2. Problem Statement

* **"Dark Data" is a Liability:** We have vast, valuable data stores that are effectively "dark." Technical metadata is often cryptic (CUST\_NM\_V2), missing, or decades out of date. This makes data discovery by analysts a "tribal knowledge" exercise, prevents safe usage, and means we cannot govern or protect data we don't fully understand.
* **Non-Compliance with Data Standards:** Our current metadata state does not consistently comply with global standards like ISO 11179. This creates significant friction for data sharing with partners and regulators, hinders data integration projects, and signals a lack of data management maturity.
* **Siloed, Un-Linked View:** Technical assets (tables, columns) are disconnected from their business meaning (the BIM). This silo prevents us from answering fundamental questions like: "Where is all our 'Customer' data?" or "If we change this data element, what reports, and business units are impacted?" This lack of a unified view prevents enterprise-level risk modeling and a true 360-degree customer view.
* **Manual Effort is Unscalable:** Manually cataloging, describing, and mapping millions of data elements across thousands of systems to the BIM is not just costly; it is operationally impossible. Any manual effort is obsolete moments after it is completed due to constant change.

## 3. Proposed AI Solution

This solution will be an AI-powered service that integrates deeply with our data catalog and source data repositories:

1. **AI Metadata Profiler:** This engine scans not just the technical metadata but also samples the underlying data (using data-safe profiling techniques). It uses statistical analysis, pattern matching, and Named Entity Recognition (NER) to infer the "type" of data (e.t., "This column contains 95% UK-formatted postcodes").
2. **AI Enrichment Engine:** This engine uses LLMs and classification models, trained on HSBC's business glossary and the BIM, to:
   * **Generate Business Descriptions:** Suggests clear, human-readable descriptions for cryptic assets. (e.g., CUST\_NM\_V2 becomes "Customer Legal Last Name, Version 2. Sourced from Project Phoenix.").
   * **Standardize (ISO 11179):** Formats the metadata to align with ISO 11179 principles (e.g., clearly defined data element concepts, value domains, and stewardship).
   * **Classify & Map to BIM:** Proposes a mapping to the most relevant HSBC Business Information Model (BIM) entity with a confidence score. (e.g., "Maps to BIM Entity: 3.1.4 - Customer Legal Name with 98% confidence").
3. **Human-in-the-Loop (HITL) Workbench:** Data Stewards are the experts. The AI's role is to empower them, not replace them. A "Data Steward Workbench" will present the AI's suggestions and confidence scores, allowing stewards to quickly "accept," "reject," or "refine" the mappings. This feedback is then used to continuously retrain and improve the model.

## 4. Strategic & Non-Financial Benefits

| **Benefit Category** | **Strategic Benefit Description** | **How We Will Measure This (KPI)** | **Strategic Alignment** |
| --- | --- | --- | --- |
| **Data Governance Maturity** | **Creation of a "Single Source of Truth":** Enriches the data catalog to become the definitive, trusted, and "living" source for data meaning, context, and quality. This is the foundational pillar of any mature data governance program. | *Metric:* % increase in metadata "completeness" and "quality" scores in the data catalog. % of CDEs with "Gold Standard" metadata. | Directly supports the core mission of Data Controls and the Chief Data Officer. |
| **Data Discoverability** | **Accelerated Data Discovery & Trust:** Empowers business users and analysts to find, understand, and *trust* the data they need for analysis and reporting. This drastically reduces "time-to-insight" and data-wrangling effort. | *Metric:* Reduction in average time (days/hours) for data analysts to find, provision, and use required data. Increased usage of certified data assets. | Enables a data-driven culture, data democratization, and unlocks value from AI/ML initiatives. |
| **Risk Reduction** | **Clear Data Lineage & Impact Analysis:** By linking technical assets to the BIM, we create reliable, end-to-end data lineage. This is critical for regulatory reporting (e.g., BCBS 239), understanding change impact, and managing data-related risks. | *Metric:* % of Critical Data Elements (CDEs) with fully mapped, automated lineage to the BIM. Reduction in time for impact analysis. | Supports risk aggregation, change management, and regulatory compliance. |
| **Standardization** | **Global Data Interoperability:** Enforces a common business language (BIM) and technical standard (ISO 11179) across the enterprise, breaking down data silos and reducing data integration friction between old and new systems. | *Metric:* % of key applications with data elements mapped to the BIM. Reduction in data integration project costs/time. | Simplifies the global data architecture and supports application modernization. |

# Confluence Page 4: AI Use Case: AI-Assisted Mapping of Record Classes to Applications

Page Title: AI Use Case: AI-Assisted Mapping of Record Classes to Applications

Parent Page: Data Controls AI Initiatives

Tags: data-storage, records-management, ai, data-retention, bim, cmdb

## 1. Executive Summary

This use case focuses on applying AI to solve a core, high-risk records management challenge: identifying *what* data (Record Class) exists in *which* application. This is the foundation of data retention and deletion. The AI solution will systematically analyze applications in our inventory (CMDB) and propose mappings to their corresponding Record Classes from the official Records Management Schedule, enabling the consistent, automated, and defensible application of data retention policies.

## 2. Problem Statement

* **Data Sprawl and "Dark" Storage:** We have thousands of applications, each storing data. It is operationally unfeasible and cost-prohibitive to manually survey and certify the Record Classes for the data each one stores and processes. Many systems are "black boxes" where the data's true nature is lost to tribal knowledge.
* **Inconsistent Retention & Legal Risk:** Without a clear, systematic map, we cannot consistently apply data retention policies. This creates two significant risks:
  1. **Premature Deletion:** We delete data we are legally required to keep (business, legal, or regulatory risk).
  2. **Over-Retention:** We keep data *past* its legal retention period, which is a direct violation of data minimization principles (e.g., GDPR), increases storage costs, and massively expands our risk "attack surface" in a data breach.
* **"Defensible Deletion" is Impossible:** We cannot confidently and systematically delete data. Furthermore, in legal e-discovery, over-retention creates a "discoverable" liability. We cannot defend *why* we deleted (or kept) data if the process is not systematic and based on official policy. This also makes the "right to be forgotten" impossible to implement.

## 3. Proposed AI Solution

The AI model will use a multi-pronged, evidence-based approach to infer and propose mappings. It will act as a "forensic analyst" correlating data from multiple sources.

1. **Application Metadata Analysis:** The AI will analyze structured and unstructured data from our CMDB (e.g., application name, description, business owner, function, supporting documents) to find initial clues about its business purpose.
2. **BIM Taxonomy Linkage (Key Dependency):** This is the most critical link. By leveraging the BIM mappings (from Use Case 3: Metadata Enrichment), the AI can see *what business data* an application processes at a granular level (e.g., "This app processes 'Trade Confirmation' and 'Client Suitability' data").
3. **Inference Engine:** The model will correlate this evidence. It will look at the Records Management Schedule, which is also mapped to the BIM.
   * **Inference Example:**
     + **Evidence 1 (CMDB):** App "Tradify" is described as "Trade capture and reporting."
     + **Evidence 2 (BIM Map):** App "Tradify" processes data elements mapped to BIM entities "Trade Confirmation" and "Client Order."
     + **Evidence 3 (Records Schedule):** The Record Class "RC-Trade-001" (7-year retention) is mapped to these same BIM entities.
     + **AI Proposal:** "We are 95% confident that App 'Tradify' is subject to Record Class 'RC-Trade-001' (7-year retention)."

## 4. Strategic & Non-Financial Benefits

| **Benefit Category** | **Strategic Benefit Description** | **How We Will Measure This (KPI)** | **Strategic Alignment** |
| --- | --- | --- | --- |
| **Regulatory Compliance** | **Enabling Defensible Deletion & Retention:** This is a non-negotiable regulatory requirement. The AI creates the auditable, logical link needed to *systematically* enforce data retention and deletion policies across the enterprise, moving from "best effort" to "provable." | *Metric:* % of applications in the CMDB with an AI-verified Record Class mapping. | Ensures compliance with legal/regulatory data retention mandates (e.g., SEC, FINRA) and privacy laws (e.g., GDPR). |
| **Risk Reduction** | **Reduced Data "Attack Surface":** By enabling systematic, defensible deletion of over-retained data, we reduce our risk profile in the event of a data breach, litigation, or regulatory e-discovery request. Less data means less liability. | *Metric:* Reduction in total volume (Petabytes) of data held past its official retention period. | Directly supports the "data minimization" and "storage limitation" principles of global data privacy. |
| **IT & Data Strategy** | **Informed Application Rationalization:** You cannot safely decommission an application if you don't know the retention requirements of the data it holds. This mapping is a critical, often-missing input for decisions on application decommissioning and cloud migration. | *Metric:* Improved data accuracy in the CMDB (linking apps to data retention schedules). | Contributes to IT simplification, technical debt reduction, and storage cost management. |
| **Scalability** | **Sustainable Governance:** Provides a scalable, automated way to manage a "living" map of apps-to-records. This is impossible to do manually at HSBC's scale. The AI can re-assess mappings as apps are updated or new BIM links are made. | *Metric:* Reduction in time-to-certify a new application's records mapping (from weeks/months to days). | Creates a sustainable, low-friction, and evergreen governance process. |

# Confluence Page 5: AI Use Case: AI-Assisted Taxonomy Migration (CDM to GDC)

Page Title: AI Use Case: AI-Assisted Taxonomy Migration (CDM to GDC)

Parent Page: Data Controls AI Initiatives

Tags: data-movement, ai, taxonomy, data-migration, cdm, gdc

## 1. Executive Summary

This initiative details the use of AI to accelerate and de-risk the bank-wide, strategic migration of data categorization from the legacy **CDM (Conceptual Data Model)** taxonomy to the new target **GDC (Group Data Category)** taxonomy. This is a foundational modernization program. The AI will act as an intelligent "translation engine," analyzing technical data sets currently tagged with CDM and proposing the correct GDC classification with a high degree of accuracy and explainability.

## 2. Problem Statement

* **Massive, Unsustainable Manual Effort:** Manually re-classifying millions of data sets, data flows, interfaces, and reports from CDM to GDC is a monumental, multi-year effort. It would require thousands of hours from our most valuable Subject Matter Experts (SMEs), diverting them from high-value work and making the project cost-prohibitive and prone to human error and fatigue.
* **Transformation Blocker:** This migration is a critical dependency for many of the bank's highest-priority transformation programs, including cloud data lake initiatives, building a single customer platform, and data-driven risk modeling. Delaying this migration blocks bank-wide strategy and delays the realization of business value.
* **Inconsistent Mapping and "Polluted" Target State:** If done manually, different teams or SMEs will inevitably interpret the new GDC taxonomy differently. This will lead to inconsistent application and a "polluted" target state, defeating the very purpose of migrating to a single, unified standard and damaging trust in the new taxonomy from day one.

## 3. Proposed AI Solution

The solution is an AI classification model, augmented by a human-in-the-loop workbench, trained to understand the complex relationship between the two taxonomies.

1. **Model Training (The "Rosetta Stone"):** The AI model (e.g., a classifier or fine-tuned LLM) is trained on a "Rosetta Stone" set of high-quality examples where data has already been manually and correctly mapped from CDM to GDC by expert data architects. This seed data is critical.
2. **AI Analysis Engine:** The model will ingest data assets (e.g., tables, files, data flows) and analyze their context from multiple angles:
   * The existing (legacy) CDM tag.
   * Enriched Metadata (from Use Case 3): Column names, business descriptions.
   * Data Lineage: Where the data comes from and where it goes.
   * BIM Linkage: The business concept the data represents.
3. **Suggestion & Confidence Scoring:** The AI will propose the most likely GDC mapping (e.g., "CDM tag 78.B ('Trade Leg') maps to GDC tag 4.5.1 'Client Transactional Data'") and provide a **confidence score** (e.g., "95% confident"). This allows us to prioritize the review.
4. **Human-in-the-Loop (HITL) Migration Workbench:** A dedicated UI will be created for data stewards. This workbench will:
   * Show all AI-proposed mappings.
   * Allow stewards to focus *only* on low-confidence mappings (e.g., <80%).
   * Enable "batch approval" for all high-confidence mappings (e.g., >95%).
   * Capture all steward feedback to continuously retrain and improve the model.

## 4. Strategic & Non-Financial Benefits

| **Benefit Category** | **Strategic Benefit Description** | **How We Will Measure This (KPI)** | **Strategic Alignment** |
| --- | --- | --- | --- |
| **Transformation Velocity** | **Accelerated Adoption of Target Data Architecture:** Radically speeds up the migration to the GDC standard, unblocking critical data modernization, analytics, and cloud programs. This is about accelerating *business value realization*. | *Metric:* Reduction in person-months required for the complete migration. % of data assets migrated, tracked weekly. | Directly enables and accelerates core bank-wide data strategy and IT simplification. |
| **Mapping Quality** | **Improved Classification Consistency:** AI applies mapping rules and semantic logic consistently at scale, 24/7. This removes the human subjectivity and error that plagues manual mapping projects, ensuring a clean, trusted target state. | *Metric:* Reduction in mapping errors discovered post-migration. % of mappings approved with "high-confidence" score. | Ensures the new GDC taxonomy is adopted cleanly and provides value from day one. |
| **SME Productivity** | **Reduced Business Disruption & Burnout:** Frees up thousands of hours for business and data SMEs. Their role shifts from low-value, repetitive "manual classification" to high-value "expert validation." | *Metric:* Reduction in total SME hours required for the migration program. Increased SME job satisfaction. | Allows our best people to focus on high-value analysis, not migration grunt work. |
| **Scalability** | **A Reusable Strategic Capability:** Creates a permanent, reusable AI capability for mapping between *any* two taxonomies, not just CDM-to-GDC. This becomes a strategic asset for future M&A integrations or system modernizations. | *Metric:* Creation of a "Taxonomy Mapping Service" (as a platform capability) for future use. | Builds a long-term, strategic asset for the Data Controls function and the enterprise. |

# Confluence Page 6: AI Use Case: AI for Data Quality Root Cause Analysis in IT Incidents

Page Title: AI Use Case: AI for Data Quality Root Cause Analysis in IT Incidents

Parent Page: Data Controls AI Initiatives

Tags: data-quality, ai, it-operations, aipops, incident-management

## 1. Executive Summary

This initiative proposes using AI (specifically Natural Language Processing) to analyze IT incident logs (e.g., from ServiceNow) to automatically identify incidents that were *symptomatic* of an underlying data quality (DQ) issue. The solution will read unstructured incident text ("Nightly batch failed," "Report is blank"), assign a probable DQ Dimension (e.g., Accuracy, Completeness, Timeliness), and provide clear reasoning. This creates a new, invaluable, and proactive feed of data for the Data Quality and Data Controls teams, allowing them to fix the *root cause* of operational failures, not just the symptoms.

## 2. Problem Statement

* **Hidden Data Issues & Repeat Incidents:** Many IT incidents are symptoms of an underlying data quality problem (e.g., "Nightly batch failed" because of an unexpected NULL value). However, they are often resolved as a "code" or "app" issue ("added error handling for NULLs") without the DQ root cause ever being logged or fixed. The *same* DQ issue then causes a *different* incident in a *different* system, leading to a costly cycle of "whack-a-mole."
* **Reactive DQ Management:** The Data Quality team largely relies on users manually reporting issues *after* they have been impacted. We have no proactive, systematic way to find DQ issues based on their real-world impact on IT systems and operational stability.
* **Ineffective Prioritization:** Without a clear link between DQ issues and operational pain, we cannot effectively prioritize remediation. We cannot quantify which data quality problems are causing the most *actual business disruption* (i.e., the most P1/P2 IT incidents), making it difficult to allocate limited Data Controls resources effectively.

## 3. Proposed AI Solution

An NLP-based AI model will be developed to read, interpret, and classify closed IT incident tickets:

1. **Secure Data Ingestion:** The AI will securely read the unstructured text fields (short description, long description, resolution notes) from closed IT incident tickets in ServiceNow.
2. **AI Classification Engine:** The model will perform a two-stage classification using semantic understanding (not just keyword matching):
   * **Binary Classification:** Was this incident *likely* (e.g., >70% probability) caused by a data issue? (Yes/No).
   * **Multi-Class Classification:** If Yes, what is the most likely **DQ Dimension** at fault? (e.g., Accuracy, Completeness, Timeliness, Uniqueness, Validity, Consistency).
3. **Explainable AI (Reasoning):** This is key for trust. The model will extract and highlight the snippet of text that led to its conclusion.
   * *Example:* **Classification:** Completeness. **Reasoning:** "Incident text mentions 'null values in PII column' and 'downstream system failed to process'."
4. **DQ Incident Dashboard:** The output will feed a new dashboard for the Data Controls team. This dashboard will show which systems, data sets (mapped via BIM), and DQ dimensions are the top drivers of IT incidents, allowing for targeted, data-driven intervention.

## 4. Strategic & Non-Financial Benefits

| **Benefit Category** | **Strategic Benefit Description** | **How We Will Measure This (KPI)** | **Strategic Alignment** |
| --- | --- | --- | --- |
| **Proactive Governance** | **Proactive DQ Issue Discovery:** Moves Data Quality from a passive, reactive function (waiting for user reports) to a proactive, detective one (hunting for issues in IT logs). It acts as a "nervous system" for the Data Controls team. | *Metric:* % increase in DQ issues identified via AI incident analysis vs. manual user reporting. | Strengthens the Data Controls "detective" control framework, finding issues before they escalate. |
| **Operational Resilience** | **Improved Root Cause Analysis (RCA):** Provides a "first pass" analysis for incident managers, speeding up true root cause analysis for data-related incidents and helping to prevent repeat failures by fixing the *data*, not just the *code*. | *Metric:* Reduction in "repeat incidents." Reduction in Mean-Time-To-Resolution (MTTR) for incidents correctly identified as data-related. | Improves operational stability, service-level performance, and application reliability. |
| **Risk-Based Prioritization** | **Data-Driven Remediation:** Creates a new, quantitative dataset that links specific DQ issues to tangible business disruption (P1/P2 IT incidents). This is the business case for DQ. | *Metric:* A "Top 10" list of data sets/applications driving IT incidents, enabling fact-based prioritization of DQ remediation efforts. | Allows Data Controls to focus its limited resources on the DQ issues that *matter most* to the business. |
| **Cross-Functional Synergy** | **Breaks Down Silos (ITOps & Data):** Creates a formal, automated, and data-driven feedback loop between IT Operations and Data Governance. It aligns both teams on the shared goal of data-driven stability. | *Metric:* Creation of a shared dashboard used by both IT Incident Management and Data Controls. Joint remediation action plans. | Fosters a more integrated, collaborative, and effective control culture across the bank. |