Irving Fisher Committee on
Central Bank Statistics

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Metadata integration as a key enabling technique for AI and ML[1]

Thomas Gottron and Daniel Suranyi,
European Central Bank

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Metadata integration as a key enabling technique for AI and ML

Thomas Gottron, Daniel Suranyi[1]

Abstract

In this paper we explore the critical role of harmonized and integrated metadata in maximizing the potential of Artificial Intelligence (AI) and Machine learning (ML) by addressing the challenges of fragmented metadata standards and silos. We emphasise the necessity of well-described, machine-readable data structures and the incorporation of methodological and business knowledge into analytical processes, illustrated with examples from central banking and official statistics. Based on these needs, we propose to create an integrated metadata layer that facilitates efficient and effective AI and ML applications. We define key requirements for such a layer. The paper outlines the key features of SDMX that make it suitable for this role and presents a high-level architecture for the proposed component.

Keywords: metadata, integration, machine learning, artificial intelligence, harmonisation, SDMX

JEL classification: C8, C55

[1] Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Contents

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) promise novel insights and gains in efficiency and effectiveness when analysing data. However, the full potential of AI and ML can only be reached when certain pre-requisites are fulfilled: (1) data from various and heterogeneous sources can seamlessly be integrated and jointly analysed, (2) data structures and semantics are well described in a machine-readable format to nurture the analysis and (3) methodological and business knowledge can be incorporated in the analytical processes. Harmonised, interoperable, and well-integrated metadata addresses all those pre-requisites, rendering it a key enabler for streamlining advanced data analytics based on AI and ML. However, reaching integrating metadata is not a trivial task and is not easily achieved. Challenges are encountered in the form of a fragmented landscape of metadata standards, metadata silos dispersed in organisations, and a lack of standardised access to metadata repositories.

In this paper we elaborate on the challenges and benefits of integrating metadata, with a specific focus on AI and ML solutions. Furthermore, we propose a solution based on SDMX to provide harmonised metadata to advanced analytical systems. In particular, we make three main contributions:

(1) We illustrate the importance of metadata for AI and ML applications. To this end, we provide prototypical example use cases from the world of central banking and official statistics. We indicate how these use cases benefit from high-quality metadata, using them to motivate the need for integrated metadata. Our analysis includes benefits like streamlining data integration, semantic handling and pre-processing of variables, and the enrichment of analytics with business context information.

(2) We propose an integrated metadata layer as unified interface for AI and ML tasks and highlight how this layer can facilitate efficient and effective applications of AI and ML. The integrated layer will provide two key capabilities: (a) to serve as standardised interface, that analytical solutions can use to consume metadata, and (b) to function as a kind of Rosetta stone for translating between and connecting to different metadata sources. Driven by those two main capabilities we postulate key requirements for an integrated metadata layer component.

(3) We demonstrate that SDMX satisfies all key requirements for such an integrated metadata layer component. We point to the key features that render SDMX a suitable candidate for an integrated metadata layer and devise the high-level architecture for the component.

Background

AI plays an important role for central banks. The topic predominantly appears in conferences and presentations. The IFC workshop series on data science in central banking (Araujo, Bruno, Marcucci, Schmidt, & Tissot, 2023) is an excellent example for the relevance of advanced analytics and its applications in central banking.

The range of use cases is wide, including network analysis to assess systematic risks (Aarab & Gottron, 2024), detecting novel plausibility checks by analysing supervisory data (Romano, Martinez-Heras, Natalini Raponi, Guidi, & Gottron, 2021) or automated data integration and measuring quality indicators like integrability and coverage (Gottron, Novello, Aarab, & Lauro, 2024). The recent popularity of large language models (LLMs) has also sparked interest and investigations in potential use cases in the domain of official statistics (Buono, Felecan, & Tessitore, 2024).

The relevance of metadata to support AI and ML use cases has been highlighted in the literature. In general, metadata can support trust in AI solutions by providing information about data pipelines and other artifacts involved in building and operating an AI model (Qin & Yu, 2023). In the context of LLMs there is evidence that a strong semantic metadata layer considerably reduces AI hallucinations and increases quality of large language models' replies to user queries (Farquhar, Kossen, Kuhn, & Gal, 2024). Likewise, metadata and provenance of ML experiments is deemed important and relevant in modern data ecosystems (Schelter, Böse, Kirschnick, Klein, & Seufert, 2017), underlining the need for an overall approach to consistently capture, manage and promote metadata management.

Integration of data and metadata is a topic which has been discussed extensively in the scientific literature. One example of interoperable and harmonised metadata deployed on a large and distributed scale is Linked Open Data (Heath & Bizer, 2011). Based on RDF (Cyganiak, Wood, & Lanthaler, 2014) and rich semantic descriptions the data comes with quasi-natural metadata. The degree of integration is visualised well in the LOD cloud diagram[2]. Various works have shown the benefits of this setup for automatic and advanced analytical applications: ranging from generic Machine Learning tasks (Mountantonakis & Tzitzikas, 2017) to schema inferencing (Konrath, Gottron, Staab, & Scherp, 2012).

Another common standard for metadata is provided by the DDI alliance (Ryssevik, 1999). The DDI standard is intended for describing social science data and other data based on observational methods. More recently, the Data Catalog Vocabulary (DCAT), published by the W3C as recommendation is designed to facilitate interoperability between data catalogues published on the Web (Albertoni, et al., 2024). SDMX, finally, is an ISO standard designed to describe statistical data and metadata. It also serves to exchange and share data across organisations (SDMX, 2021). We will provide further details on SDMX in the Section on Evaluation of SDMX as technology for an integrated metadata layer.

The relevance of metadata for AI applications

Data is used for training and operating the models underlying AI applications. Accordingly, the availability and accessibility of high-quality data is key for successful AI applications. Nowadays, applications frequently tap into more than one data source: they integrate data from multiple sources to facilitate a joint analysis and

2 The Linked Open Data Cloud: lod-cloud.net/# (last accessed: 17 April 2025)

generate novel insights. Different types of metadata play a crucial role in discovering, accessing, integrating, and analysing data from heterogenous sources.

- *Business metadata* provides context and methodological information on how data is processed and obtained. It can give insights on the scope of a dataset, how it is collected, and which considerations need to be accounted for during analytical activities. Business metadata is particularly important as it offers context and methodological information that enables modern AI and ML systems to understand the data in relation to specific business processes and objectives. This type of metadata helps in aligning data use with strategic goals and operational requirements, ensuring that generated insights are relevant and actionable.

- *Semantic metadata* provides relevant insights on the content aspect and which information is modelled in the data. This is also key for data integration tasks and the alignment of different and heterogeneous datasets. Metadata on semantics is another critical component, as it facilitates the integration, harmonization, and alignment of definitions across various data sources. It ensures that data from heterogeneous origins can be combined and compared in a meaningful way, which is essential for creating cohesive and comprehensive analytical models. By standardizing the language and definitions used within datasets, semantic metadata reduces ambiguity and enhances the interoperability of data.

- *Structural metadata* describes data from a syntactical point of view. As such, it provides the basic notion for which features are available in the data. Structural metadata outlines the data models, classifications, mappings, and transformations that govern how data is organized and accessed. This type of metadata provides the necessary framework for data analysis, allowing AI and ML models to leverage complex datasets effectively. It ensures consistency in data representation, which is crucial for maintaining the integrity and reliability of analytical outcomes.

- *Technical metadata* offers insights into the technical aspects of the data, such as encodings, provenance, and freshness. This information is vital for assessing the quality and trustworthiness of data, as well as for managing data lifecycle and compliance requirements. By providing detailed technical information, this type of metadata supports the operational efficiency and effectiveness of AI and ML systems, ensuring that they can deliver timely and accurate insights.

Use cases where metadata supports AI and ML include, but are not limited to, streamlining data integration, semantic handling and pre-processing of variables, and the enrichment of analytics with business context information. In the following sub-sections we motivate the importance of metadata and, in particular, of harmonised, interoperable, and well-integrated metadata, for AI applications using two illustrative use cases from the domain of central banking.

Analysis of bank exposures networks and systemic risks

Integrating data for performing advanced analytics on joint datasets is a common use case. Metadata plays a crucial role in the integration of datasets by providing essential context and details about the data, which makes combining and utilizing different datasets more effective. One specific use case is the integration of interbank exposures for the assessment of systemic risks based on network analytics (Aarab & Gottron, 2024).

In a first step, structural metadata enhanced data understanding by offering schema information, including the structure of the data such as tables, columns, data types, and relationships. Additionally, descriptive metadata explained data semantics, i.e. what each data element represents. Semantic descriptions ensured that datasets are compatible and can be meaningfully merged. This helped in comprehending how various datasets can be aligned and integrated. Metadata was also instrumental in mapping data across sources by providing details about how data elements in one dataset correspond to those in another, facilitating the alignment of different data models. This included the indication of attributes corresponding to unique identifiers as well as attributes following standards and conventions used in the datasets, aiding the harmonization of diverse datasets. Finally, business metadata helped to understand the scope and purpose of a dataset. Thereby, it helped to understand which datasets are relevant to a given analytical task and should be considered for integration and joint analysis.

Data quality assurance

Another strong analytical use case for AI in central banking and official statistics is quality assurance. In this case, metadata can be used to support the formulation of validation and plausibility checks, to identify patterns for quality assurance or to ensure alignment of data with business specifications.

On the lowest level of quality assurance mechanisms, technical and structural metadata can help to formulate simple rule-based validation checks to ensure formats are correct or that values comply with code lists. Descriptive metadata and business metadata provide semantic and practical context, which can be used to choose the most suitable AI model for anomaly detection or for performing the detection of more complex patterns in data. Such anomalies and complex patterns provide the basis for verifying non-trivial correlations and connections in the data. Changes in provenance metadata can indicate disruptions or modifications in upstream data production processes. Detecting such changes and assessing their impact on further processes can help in preventing data quality issues. Such monitoring approaches could, for instance, also be used to indicate when an AI model needs to be retrained and updated.

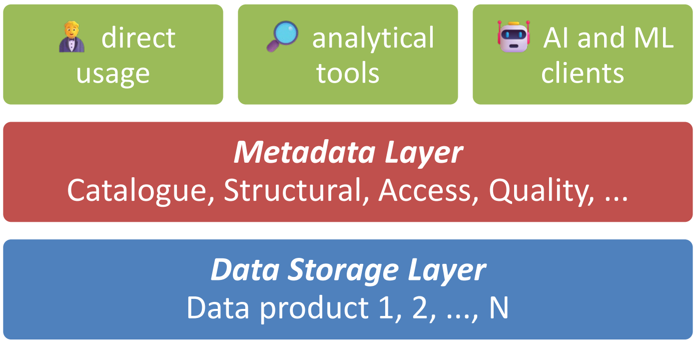
The above points motivate the positive impact of metadata as a support function for quality assurance. As soon as data processes interact with different data products and draw upon multiple sources, also the metadata of these sources needs to be included in order to support the corresponding data quality monitoring and assurance activities. Finally, the findings of DQ assessments and quality KPIs can feed into metadata as well and provide a notion of reliance for AI models. Additionally, integrated metadata can be an enabler for automating the step of data integration and measuring the fitness of a datasets for a joint analysis as demonstrated in (Gottron, Novello, Aarab, & Lauro, 2024). In this way it can save manual work, speed up the downstream analytical process and ensure a harmonised and consistent data integration approach.

A metadata layer for AI and ML

In a metadata driven system, data is always accessed through a metadata layer on top of the data stored in (various) storage locations. The metadata layer ensures a technology independent decoupling of the data storage infrastructure from the technologies used to access data and provides an abstraction layer between business functions and data storage technologies (Figure 1).

Data Management Layers

Figure 1



A metadata repository typically provides key functionality to search, access, explore and interact with metadata. Key services provided by a metadata layer or repository are:

- Serving as *single access* point for metadata
- Providing metadata in a *well-defined format*, utilising an established format (e.g. DCAT, SDMX, RDF).
- Providing programmatical access via *standardised interfaces* (e.g. APIs, SQL, linked data)

As depicted in Figure 1, also AI and ML clients are foreseen to be build upon metadata layers. Rational is, that metadata is a pivotal element for streamlining AI and ML solutions, as it provides essential context and structure to the data used in these solutions. The technical nature and interfaces of a metadata repositories are particularly relevant and important for AI and ML clients.

However, within many organizations, metadata silos present significant challenges to effective data utilization by AI and ML solutions. These silos occur when metadata is stored in isolated systems or by independent departments, preventing seamless access and integration across the organization. As a result, valuable metadata remains underutilized, hindering the ability to leverage data for comprehensive analysis and decision-making across a variety of data sources.

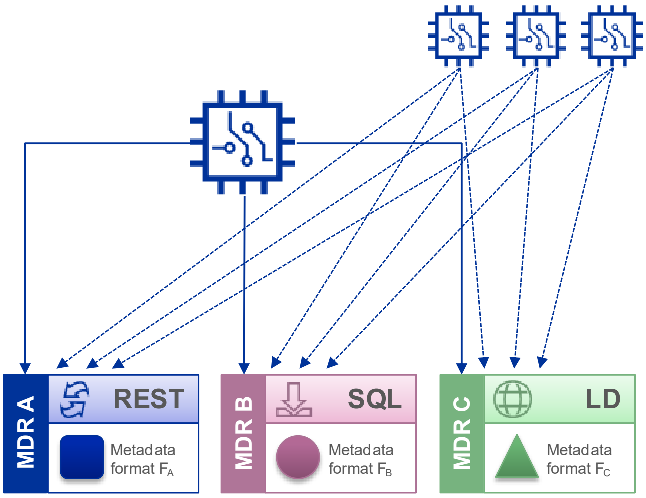
Additionally, the lack of semantic integration and metadata governance amplifies these challenges by potentially leading to inconsistencies and misalignments in how data is defined and used. Without a unified approach to metadata management, different parts of an organization may use varying definitions and standards, leading to confusion and inefficiencies.

Thereby, a fragmented landscape of metadata standards further complicates the situation, as it creates barriers to interoperability and standardization. With multiple standards, organizations face difficulties in aligning their metadata practices both internally and with external partners. This fragmentation makes it challenging to develop comprehensive and cohesive analytical solutions.

Typically, such a fragmented landscape is not a deliberate choice, but the result of different projects, incorporation of external sources and services or the extension of services to legacy systems. Nevertheless, it poses the challenge that AI and ML systems need to be capable to interact with multiple metadata access points, speak to potentially different metadata standards and interact with different types and natures of interfaces. Furthermore, the alignment and integration of metadata resides with the AI and ML solution, that has to ensure a semantically correct and seamless integration of metadata and data across the repositories. Such tasks are amplified and become a burden if more than one application needs to address, implement and solve them as depicted in Figure 2.

Scenario of several ML and AI applications accessing a fragmented metadata landscape consisting of heterogeneous metadata repositories

Figure 2



An integrated metadata layer as mediator

An integrated metadata layer, that unifies across a fragmented metadata landscape can bring benefits and remedy some of the issues described above. The main tasks of an integrated metadata layer are threefold:

(1) Provide a *unique and single access point* for all applications, in particular ML and AI solutions. This unique access point remains stable over time and is independent of underlying data storage systems and their connected metadata repositories.

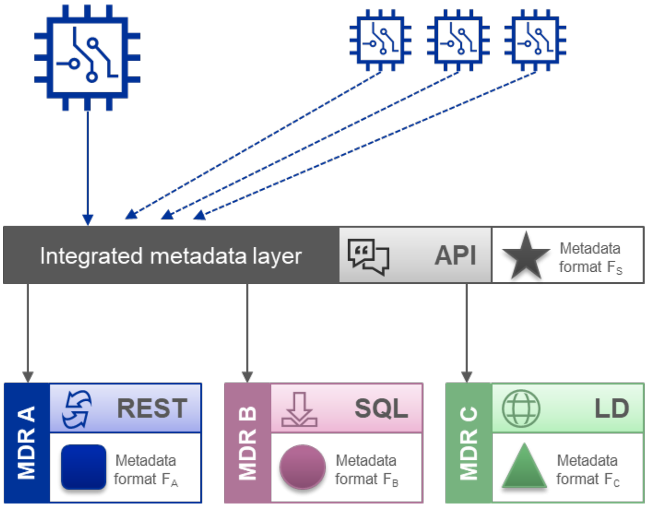
(2) At the same time, the integrated metadata layer provides a *standardised format and API for accessing information*. This includes transparent translation between different metadata standards, APIs and end points.

(3) An additional functionality to enhance the service is to provide a *semantic mapping and translation* between the semantics in the metadata repositories. Thus, it serves as a kind of Rosetta stone[3] to define mappings of code lists, concepts, taxonomies, ontologies, etc.

The result is a scenario as depicted in Figure 3. The integrated metadata layer would serve as mediator between AI and ML applications and a heterogeneous and fragmented metadata landscape, fulfilling the above three tasks.

An integrated metadata layer to mediate between AI applications and a heterogeneous metadata landscape

Figure 3



Requirements for an integrated metadata layer

This role implies certain requirements for the integrated metadata layer. In the following, we formulate and motivate these requirements based on the key functionality an integrated metadata layer has to implement to satisfy the needs of AI and ML applications. It may also serve further purposes and usage as indicated in Figure 1. We group the requirements in four distinct groups:

1. Discovery, access, and retrieval

a. Provision of a *unified and standardized information model* to ensure consistency and interoperability across different systems and applications. This model should facilitate the efficient management, discovery, and retrieval of metadata.

b. Implementation of a *modern, standardized API* that supports robust searching and retrieving of metadata. This API should be capable to handle complex queries and provide efficient and effective access to

[3] The Rosetta Stone, a granodiorite stele discovered in 1799 near Rosetta, Egypt, features the same text in Greek, Demotic, and hieroglyphic scripts. It played a crucial role in deciphering Egyptian hieroglyphs, as the Greek text enabled to unlock the meaning of the other scripts in the 1820s.

metadata, enabling seamless interaction and integration with various systems and applications.

2. Support for different types of metadata

a. Ensure the metadata layer can *handle business-related metadata*, which includes definitions, policies, and methodological information supporting and underlying the data. This should facilitate AI applications' understanding and intended use of data assets.

b. *Support semantic metadata* to capture the meaning and context of data elements. This includes relationships between data, ontologies, and vocabularies that enrich data understanding and interoperability.

c. Capability to model, *represent and serve quality metadata* to assess and ensure the reliability, accuracy, and timeliness of data. This should include information about metrics and indicators that help to assess and monitor data quality.

d. Provide robust *support for structural metadata*, which describes the structure and schema of data assets. This includes master metadata that defines critical data elements and hierarchies necessary for comprehensive data governance.

3. Data integration and transformation

a. Design the metadata layer to describe data at *various levels of granularity*, including detailed, aggregated, and master data. This ensures flexibility and precision in data management and utilization.

b. *Facilitate mappings between different metadata definitions* to allow for harmonization and alignment across various data sources and systems. This is crucial for seamless data integration and transformation processes.

c. Enable the metadata layer to *support and facilitate data transformation processes*. This includes capabilities to define transformation rules and processes that convert data into a desired format or structure.

4. Based on open standards

a. Ensure the metadata layer is *based on open standards* to foster integration with a diverse ecosystem of solutions. This promotes interoperability and collaboration across different technologies and platforms, primarily inside, but also outside the organisation.

b. Design the metadata layer to be *extensible and open to integration with external data sources* and systems. This includes support for plug-ins, APIs, and other mechanisms that allow for seamless expansion and integration, ensuring the system can adapt to evolving requirements and technologies.

Evaluation of SDMX as technology for an integrated metadata layer

SDMX lends itself as a solution. In fact, the framework for the SDMX technical standards states that "[...] It is felt that a clear, standard formatting of data and metadata for the purposes of exchange and dissemination can also facilitate internal processing by organizations and users [...]" (SDMX, 2021). This motivated us to investigate further, to which degree SDMX satisfies the requirements we elaborated above.

The SDMX standards package consists of various components, which standardise different aspects of data management and data exchange (SDMX, 2022; SDMX, 2023). The relevant parts of SDMX for this paper are listed below:

- **SDMX Technical Standards**
    - *Information Model*: definition of artefacts used to model data products, data exchange frameworks, and various other related aspects.
    - *Registry Specification*: description of logical interfaces used by a structural metadata repository for registration, storage, querying, and retrieval of SDMX artefacts.
    - *SDMX API*: standardised web service interfaces for data- and metadata machine-to-machine interaction.
    - *SDMX message formats*: specification of different formats used for structural metadata serialisation and file-based data exchange.
- **SDMX Content Oriented Guidelines**
    - *Globally agreed concepts, code lists, classifications* which may be used in SDMX data models.
    - Various *guidelines and governance approaches* detailing data management best practices in relation to SDMX.

Additionally, the SDMX governance structure also takes care of the *Validation and Transformation Language standard* (VTL)[4]. VTL is a standardised syntax for validation and transformation expressions, independent but fully compatible with the SDMX information model.

The different SDMX components are put in relation to the required capabilities analysed in the previous section and scored with a traffic light indicator (red/amber/green) for providing the required capability not sufficiently (red), partially (amber), or fully (green).

[4] Validation and Transformation Language: sdmx.org/validation-and-transformation-language-vtl/ (last accessed: 17 April 2025)

(1) Discovery, access, and retrieval

SDMX Score: **AMBER**

A structural metadata repository based on the SDMX standard (i.e. SDMX Registry) can be used to manage data domains and products. However, in most use cases a data catalogue will be a better choice for enterprise-level data discovery due to more advanced governance and data product management features available in data catalogue software packages available on the market. In the best case, a data catalogue is used for high level data product management activities including governance / ownership, discovery, business lineage, and other rather high-level tasks. The data catalogue is then integrated with an SDMX Registry to provide structural metadata supporting the data products and the rest of the data integration process. The standardised SMDX information model, API, and web service interfaces allow seamless machine-to-machine integration between the data catalogue, the SDMX Registry, and other platforms needing to consume metadata.

(2) Support for different types of metadata

SDMX Score: **GREEN**

An SDMX Registry is capable to provide structural metadata supporting data products and the relevant aspects for a data integration process.

The SDMX information model natively supports linking different metadata artefacts including a comprehensive support for versioning and referencing.

The SDMX Metadata Structure Definition and Metadata Set allows defining quality frameworks within structural metadata and exchanging quality information as structured messages. Ideally, the main KPIs stemming from such technically standardised quality metadata are integrated in the data product information in a data catalogue and may be used to feed data quality dashboards using an SDMX API.

(3) Data integration and transformation

SDMX Score: **GREEN**

SDMX provides at its core the standard information model, which can be used to express any structured data model and their relationships to each other, including microdata, transactional data, macro data, and master data. An SDMX Registry provides the repository required for management.

The SDMX cross-domain concepts and code lists provide a basis for master metadata management. Additionally, the content published by the SDMX agency in the Global SDMX Registry can be a starting point for enterprise-wide shared concepts and code lists. The standard allows extension of those artefacts by reference and thus an organisation framework can be placed on top of the global one. Various maintenance agencies may cooperate and share relevant structural metadata in a flexible and decentralised way.

SDMX Structure Maps provide the basic capabilities to express mappings between different structures (data models, concepts, code lists) for 1-to-1 as well as many-to-many relationships.

For more complex integration scenarios requiring metadata-driven transformations, SDMX natively integrates with VTL. Complex transformations can be stored in an SDMX Registry as VTL expressions and linked to the relevant structural metadata artefacts.

(4) Based on open standards

SDMX Score: **GREEN**

SDMX is an open source standard sponsored by 8 large international organisations (BIS, ECB, European Commission (Eurostat), ILO, IMF, OECD, United Nations (UNSD), World Bank). Many related software packages are available as open source or commercial software products. For custom use cases, SDMX allows custom extensions of the information model through a feature called controlled annotations (SDMX, 2021). The SDMX information model, API specifications, and web service standards allow seamless integration across systems requiring or managing metadata and allow for an enterprise-level single metadata management approach.

Summary and conclusions

Metadata is a key enabler for efficiently implementing and operating trustworthy AI and ML applications. It provides essential information on data that are needed for advanced analytics. A challenge for analytical applications remains an often fragmented and heterogeneous landscape of metadata repositories. We propose the solution of establishing an integrated metadata layer, that serves as single access point, provides a unified API and metadata format and harmonises and integrates metadata, thereby serving as a kind of Rosetta stone to translate between various metadata repositories. We postulated key requirements and demonstrated that SDMX seems a good fit as a basis for implementing an integrated metadata layer. As a way forward we consider implementing a first prototype with limited scope to test and validate our assessments in a real-world scenario.

References

Aarab, I., & Gottron, T. (2024). Network Topology of the Euro Area Interbank Market. *High-quality and Timely Statistics* (pp. 3-19). Cham: Springer Nature Switzerland.

Albertoni, R., Browning, D., Cox, S. J., Gonzalez Beltran, A., Perego, A., & Winstanley, P. (2024, August 22). *Data Catalog Vocabulary (DCAT) - Version 3*. Retrieved from W3C Recommendation: https://www.w3.org/TR/vocab-dcat-3/

Araujo, D., Bruno, G., Marcucci, J., Schmidt, R., & Tissot, B. (2023). Data science in central banking: applications and tools. *Proceedings of the Irving Fisher Committee on Central Bank Statistics (IFC)-Bank of Italy Workshop on "Data science in central banking: applications and tools"*. Basel: BIS.

Buono, D., Felecan, M., & Tessitore, C. (2024). An introduction to Large Language Models and their relevance for statistical offices. *Eurostat Statistical Working Papers*.

Cyganiak, R., Wood, D., & Lanthaler, M. (2014, February 25). *RDF 1.1 Concepts and Abstract Syntax*. Retrieved from W3C Recommendations: https://www.w3.org/TR/rdf11-concepts/

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024, June 19). Detecting hallucinations in large language models using semantic entropy. *Nature*, 625-630. doi:https://doi.org/10.1038/s41586-024-07421-0

Gottron, T., Novello, A., Aarab, I., & Lauro, B. (2024). Assessing Fitness for Integration – a Metadata-driven Approach. *11th European Conference on Quality in Official Statistics (Q2024).*

Heath, T., & Bizer, C. (2011). Linked data: evoluving the web into a global data space. *J. Web. Eng., 11*(2), 177-178.

Konrath, M., Gottron, T., Staab, S., & Scherp, A. (2012). Schemex—efficient construction of a data catalogue by stream-based indexing of linked data. *Journal of Web Semantics*(16), 52-58.

Mountantonakis, M., & Tzitzikas, Y. (2017). How linked data can aid machine learning-based tasks. *International Conference on Theory and Practice of Digital Libraries* (pp. 155-168). Springer.

Qin, J., & Yu, B. (2023). Metadata in Trustworthy AI: From Data Quality to ML Modeling. *International Conference on Dublin Core and Metadata Applications.*

Romano, S., Martinez-Heras, J., Natalini Raponi, F., Guidi, G., & Gottron, T. (2021). Discovering new plausibility checks for supervisory data. *ECB Statistics Paper Series (SPS), 41*, 1-38.

Ryan, L. (2016). Architecting for Discovery. In L. Ryan, *The Visual Imperative* (pp. 203-220). Burlington, MA, USA: Morgan Kaufmann.

Ryssevik, J. (1999). Providing Global Access to Distributed Data through Metadata Standardisation–the Parallel Stories of Nesstar and the DDI. *Conferenc of European Statisticians.* Geneva.

Schelter, S., Böse, J.-H., Kirschnick, J., Klein, T., & Seufert, S. (2017). Automatically tracking metadata and provenance of machine learning experiments. *NeurIPS 2017.*

SDMX. (2021, March). *Guidelines on using SDMX Annotations.* Retrieved from SDMX.org: https://sdmx.org/wp-content/uploads/Guidelines-on-the-use-of-SDMX-Annotations.pdf

SDMX. (2021). *SDMX STANDARDS: SECTION 1 FRAMEWORK FOR SDMX TECHNICAL STANDARDS VERSION 3.0* . SDMX. SDMX.org. Retrieved from https://sdmx.org/wp-content/uploads/SDMX_3-0-0_SECTION_1_FINAL-1_0.pdf

SDMX. (2021, October). *SDMX.org.* Retrieved from SDMX STANDARDS: SECTION 2 - INFORMATION MODEL: UML CONCEPTUAL DESIGN - Version 3.0: https://sdmx.org/wp-content/uploads/SDMX_3-0-0_SECTION_2_FINAL-1_0.pdf

SDMX. (2022, November 29). *SDMX Technical Specifications*. Retrieved from SDMX.org: https://sdmx.org/standards-2/

SDMX. (2023, May 24). *SDMX Content-Oriented Guidelines (COG)*. Retrieved from SDMX.org: https://sdmx.org/guidelines/

EUROPEAN CENTRAL BANK
EUROSYSTEM

Metadata integration as a key enabling technique for AI and ML

IFC Workshop 2025

The views expressed are those of the authors and do not necessarily reflect those of the ECB.

20/02/2025

Thomas Gottron, Daniel Suranyi
ECB Data Office

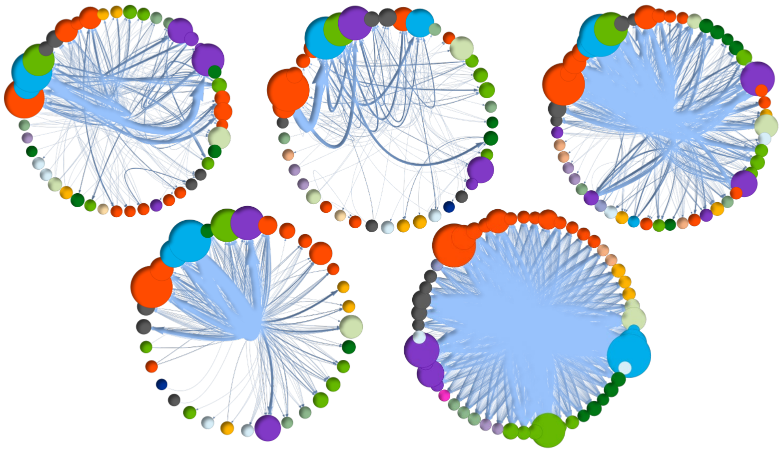
Metadata supports and enables AI applications

Metadata plays a crucial role for AI and ML:

- **Business metadata**: provision of context and methodological information
- **Semantic metadata**: integration, harmonisation and alignment of definitions
- **Structural metadata**: data models, classifications, mappings, transformations
- **Technical metadata**: insights on encodings, provenance, freshness, etc.

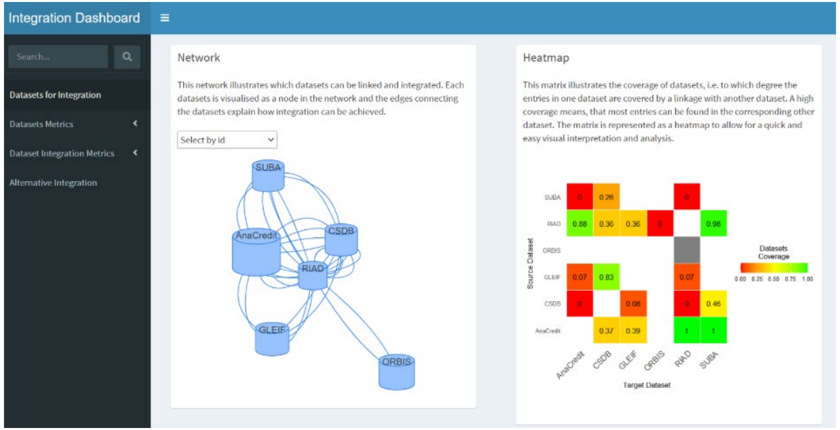
Metadata driven AI use cases in central banking

Analysis of bank exposures and systemic risks



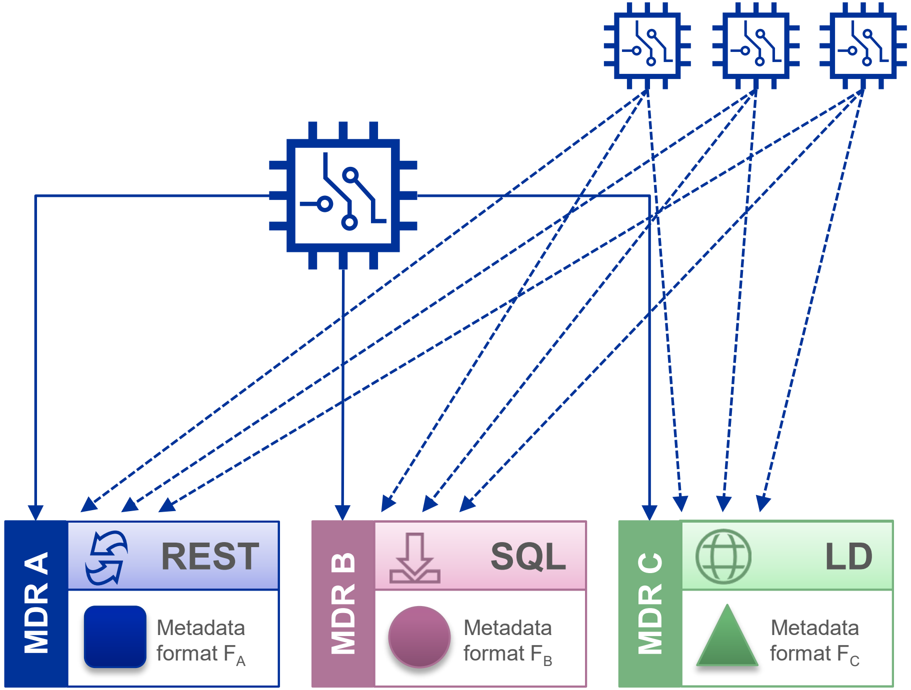
Aarab, I., & Gottron, T. (2024). Network Topology of the Euro Area Interbank Market. High-quality and Timely Statistics (pp. 3-19). Cham: Springer Nature Switzerland

Quality assurance for integrated datasets



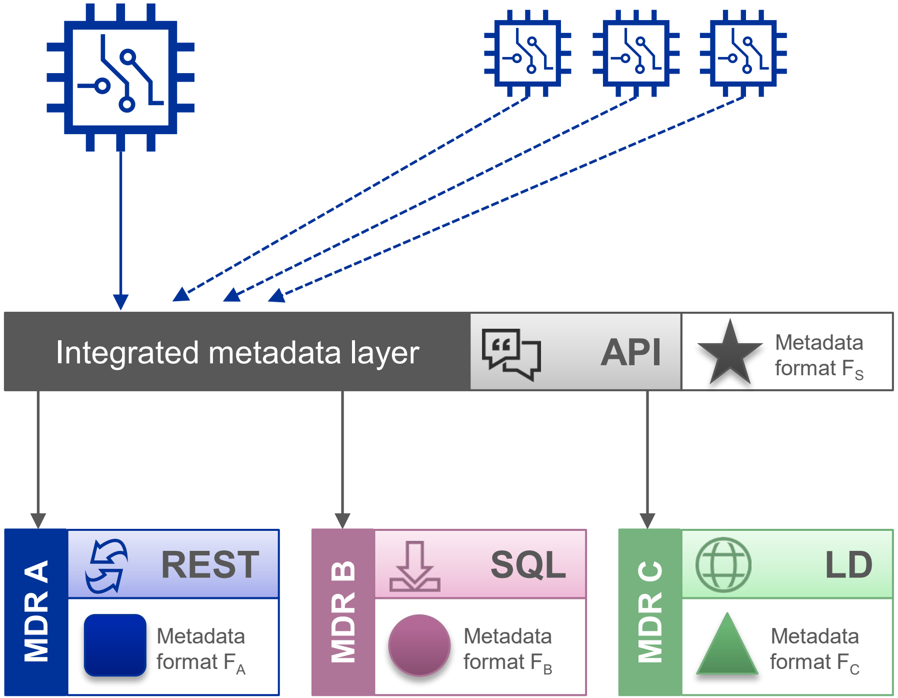
Gottron, T., Novello, A., Aarab, I., & Lauro, B. (2024). Assessing Fitness for Integration – a Metadata-driven Approach. 11th European Conference on Quality in Official Statistics (Q2024).

Challenges in metadata integration



- Metadata silos even within organisations
- Lack of semantic integration and metadata governance
- Fragmented landscape of metadata standards
- Lack of standardised access to metadata repositories

Idea and benefits of an integrated metadata layer



Single metadata access point for all AI and ML applications

Standardised interface (API & format)

„Rosetta stone“ for translating between and connecting to heterogeneous metadata sources

Key requirements for an integrated metadata layer

Discovery, access & retrieval

- Standardised information model
- Modern standardised API for searching and retrieving metadata

Support for different types of metadata

- Business metadata
- Semantic metadata
- Quality metadata
- Structural metadata, incl. master metadata

Data integration and transformation

- Capability to describe granular, aggregated and master data
- Mappings between metadata definitions
- Support for data transformation

Based on open standards

- Integrate with a wide ecosystem of solutions
- Extensibility and open to integration with external sources

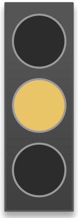
SDMX as backbone for an integrated metadata layer

"It is felt that a clear, standard formatting of data and metadata for the purposes of exchange and dissemination can also facilitate internal processing by organizations and users, but this is not the focus of the specification"

[SDMX Standards: Section 1]
Framework for SDMX technical standards, Version 3.0

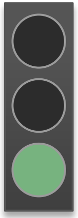
Assessment of SDMX and its features

Discovery, access & retrieval



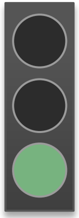
+ SDMX registry
+ Core standard information model
- Fully fledged data catalogue will be more suitable for discovery

Support for different types of metadata



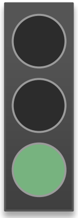
+ Support for reference data, structural information, description of data flows
+ SDMX Metadata Structure Definition and Metadata Set

Data integration and transformation



+ Cross domain concepts and code lists
+ Structure maps, representation maps, item scheme maps
+ Integration of VTL

Based on open standards



+ SDMX as open and extensible ISO standard
+ Wide range of existing third-party and open-source solutions and tools

Summary and Conclusions

Harmonised and integrated metadata is a key enabler for AI and ML application

An integrated metadata layer helps to overcome drawbacks in current metadata landscapes

SDMX is a suitable solution to serve as backbone technology for an integrated metadata layer