Irving Fisher Committee on
Central Bank Statistics

BIS

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

18-20 February 2025

sdmx.io: where tools and learning resources come together to address official statistics use cases[1]

Douglas Araujo,
Central Bank of Brazil

Brian Buffett,
Bank for International Settlements

[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.

sdmx.io: where tools and learning resources come together to address official statistics use cases

Douglas Araujo (Banco Central do Brasil)[1] and Brian Buffett (Bank for International Settlements)

The increasing demands for data is creating pressure for more agile data organisations, processes, and software systems. The SDMX standard provides an information model and technical architecture that supports robust metadata-driven processes and systems. One challenge in realising this goal is to reduce barriers to entry and use throughout the data lifecycle for organisations of all size and stage of development. Sdmx.io helps to address this challenge. sdmx.io is not a single project but an ecosystem of open source tools, patterns, guidance, learning materials and other resources like pre-configured containerised environments that make the software quick and simple to deploy. The quantity of SDMX compliant data portals and data available via SDMX API's is becoming extensive and exploring ways to effectively leverage the richness of these data assets is essential. This paper discusses this topic from the perspective of data scientists and more specifically, various tools and methods to facilitate the discovery and use of SDMX data sources.

Keywords: Data science, SDMX, Data discovery

JEL classification: C81, C82, C89

[1] This paper reflects only the views of the authors and not necessarily of the Banco Central do Brasil.

1. Introduction

The demand for timely, granular, and reliable data is greater than ever before, driven by policymakers, businesses, and researchers' need for increasingly complex evidence bases. Data organizations face mounting pressure to become more agile, adapting their processes, infrastructures, and governance models to a world in which information flows are continuous, diverse, and rapidly changing. At the same time, the rise of artificial intelligence (AI) and machine learning amplifies expectations for structured, well-documented, and machine-readable data. Algorithms, no matter how sophisticated, depend on the quality and interpretability of the underlying datasets. For many data users, however, the reality is frustrating: valuable time and resources are consumed by data cleaning, format conversions, and metadata reconstruction before meaningful analysis can even begin.

The Statistical Data and Metadata Exchange (SDMX) standard was developed precisely to address these challenges (see Buffett and Napoles (2024)). SDMX offers a robust information model and technical architecture. It was initially designed to enable efficient and standardized exchange of official statistics between organizations. Over time, it has matured into a versatile framework that combines metadata-driven processes with interoperable formats and application programming interfaces (APIs), capable of supporting both one-off analysis and recurrent data pipelines. It is an ISO standard, backed by leading international organizations, namely the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the International Labour Organization (ILO), the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations, and the World Bank.

Despite its proven role in official statistics, SDMX remains little known in the broader data science community. Many economists, social scientists, and advanced users perceive it as a tool exclusively for national statistical offices and central banks. This perception is misleading and increasingly outdated. The core strengths of SDMX, ie rich metadata, standardized models, discoverability, and automation directly address the pain points faced by data scientists. This paper explores how SDMX is fit-for-purpose in the era of data science, outlines its value proposition for advanced users, and introduces practical initiatives such as sdmx.io and sdmx4ds that aim to make the standard more accessible.

2. SDMX: Beyond data producers

The traditional view of SDMX is tightly linked to its origins: a standard primarily for producers of official statistics. Under this framing, SDMX is seen as a tool for large institutions such as national statistical offices, international organizations, and central banks to harmonize reporting, reduce duplication, and improve data dissemination. It is true that these institutions make extensive use of SDMX (eg, Bogdanova et al (2025a)). But this impression overlooks how the same features that benefit producers can also empower data scientists and advanced users.

Rich metadata support

At the heart of SDMX lies metadata. Datasets are not just a collection of numbers; they are annotated with information about variables, units of measure, methodologies, and sources. SDMX describes these features, and more. Concretely, SDMX defines a set of metadata artefacts such as Data Structure Definitions (DSDs), code lists, concept schemes and dataflow that describe the dimensions, attributes, measures, and permissible values of a dataset. These structural metadata components specify, for example, what a "unit of measure" means, how a "reference period" is encoded, or which classification system is applied. In addition, SDMX supports reference metadata, which document methodological information such as data sources, compilation practices, quality assessments, and revision policies.

For a data scientist, this rich descriptive layer provides immediate clarity. Instead of manually reconstructing meaning from raw column headers or accompanying documentation (which is often not sufficiently clear), users can quickly assess relevance and understand structure. This reduces ambiguity and enables consistent treatment of time series, cross-sectional data, or panel datasets across domains. The standardized metadata in SDMX also facilitates reproducibility of analysis, and lowers the bar for the adaption of one analysis to other circumstances.

Improved Data Discovery

Because SDMX uses structured metadata, it enables far more effective search and filtering. Analysts can identify datasets by filters including geographic coverage, indicator type, or time period, all without downloading multiple files. For example, a researcher seeking long-term interest rate data across advanced economies can query SDMX APIs using metadata filters, retrieving exactly what is needed without manual scanning of spreadsheets.

Standardized Data Models

SDMX enforces consistency across datasets by applying common models for dimensions, measures, and code lists. This means data from the BIS, IMF, and ECB — though produced independently — can be integrated seamlessly.

For comparative analyses or multi-source dashboards, this is a decisive advantage. Data scientists can integrate multiple datasets using metadata described with a common language.

Automation and API Integration

SDMX's design aligns with the languages and platforms of data science. Whether using Python libraries for econometrics, R packages for visualization, or business intelligence software for reporting, SDMX's standardized formats reduce friction. For advanced users, the ability to automate analyses and integrate tools is critical. SDMX APIs allow programmatic retrieval of data directly into tools like Python, R, or Julia. Instead of downloading files manually, users can build reproducible pipelines that update automatically when new data is released. Analysts can integrate statistical data

into diverse toolchains without needing bespoke conversions. This functionality underpins modern workflows for monitoring systems, forecasting models, or interactive dashboards.

One illustration of how SDMX integrates with data science tools is *gingado*, the open-source tool for machine learning in economics and finance (Araujo (2023)). SDMX offers the capability to automatically augment the user's data to train machine learning models, by finding relevant data that might improve the models' performance.

Together, these features make clear that SDMX is not just a producer's tool. For advanced users, it represents an underutilized resource: a ready-made framework for discoverable, interoperable, and reproducible data use.

3. Fit-for-Purpose for Data Science

The true test of any data standard in the 2020s is whether it can meet the demands of advanced analytics. SDMX passes this test by design. Its metadata-driven architecture is not an optional add-on; it is the foundation. This makes SDMX particularly fit-for-purpose for data scientists, whose workflows depend on transparency, structure, and reproducibility.

Reducing data wrangling

A large share of data scientists' time is consumed by cleaning, formatting, and reconciling datasets. SDMX minimizes this burden. By embedding definitions, units, and code lists into the structure, it allows users to skip much of the pre-processing that is typically required. The result is faster time to analysis and more reliable outcomes.

Enabling reproducible workflows

Reproducibility is a core principle in both academic research and applied policy analysis. SDMX supports reproducibility through standardization. When datasets are pulled via SDMX APIs, they come with consistent identifiers and metadata. This ensures that the same script will yield the same dataset tomorrow, next month, or in another user's environment. Such reproducibility is critical in collaborative projects, peer review, and policy simulations.

Bridging Official Data and Advanced Analytics

Official statistics covering areas like macroeconomic indicators, labor markets, or trade are essential for economic and policy analysis. Yet these datasets often exist in fragmented formats, with limited programmatic access. SDMX bridges the gap by making official data more accessible to advanced analytics pipelines. It allows the trusted, authoritative data of statistical agencies and central banks to flow directly

into machine learning models, visualization dashboards, or econometric frameworks (see Bogdanova et al (2025b) for examples).

Supporting Multi-Source Integration

Complex problems rarely respect domain boundaries. Consider climate change, which requires integrating environmental, financial, and socio-economic data. SDMX facilitates such integration by enforcing consistent structures and metadata across domains. Analysts can combine datasets from multiple organizations without extensive re-engineering. This capacity is essential for interdisciplinary policy challenges and cross-border research.

Alignment with FAIR Principles

The principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data (Wilkinson et al (2016)) are widely recognized as best practice. SDMX embodies these principles: its metadata makes data findable, its APIs make it accessible, its standard models make it interoperable, and its structured formats make it reusable. For data scientists navigating a crowded data landscape, this alignment offers confidence that SDMX-based resources can be reliably integrated into workflows.

By addressing these core needs, SDMX establishes itself not just as a standard for producers, but as a fit-for-purpose enabler of modern data science.

4. The sdmx.io Ecosystem

Recognizing the dual needs of producers and users, the *sdmx.io*[2] initiative was launched as an ecosystem of resources designed to lower barriers and broaden adoption. Its purpose is straightforward: to make it easier for organizations of all sizes and stages of development to adopt and benefit from SDMX.

The ecosystem includes:

- Open-Source Tools: Ready-to-use libraries and utilities for working with SDMX data, compatible with common programming languages.
- Containerized Environments: Pre-configured setups that simplify deployment, allowing institutions to test and scale SDMX solutions without major IT investments.
- Learning Materials: Tutorials, webinars, and documentation that guide both technical and non-technical audiences through adoption.
- Community Resources: Shared repositories where users can exchange code, methods, and lessons learned.

[2] https://sdmx.io

By bringing together tools, environments, and knowledge in a single hub, sdmx.io reduces the perception of complexity that has sometimes discouraged uptake. It transforms SDMX from a niche technical standard into a practical solution accessible to both institutional producers and individual data scientists.

5. SDMX for Data Science (sdmx4ds)

In order to fully realize SDMX's potential for advanced users, the *sdmx.io* initiative has established a dedicated effort known as *sdmx4ds*[3] (SDMX for Data Science). This initiative builds a bridge between the world of official statistics and the growing community of data scientists.

At its core, *sdmx4ds* is a **repository of reproducible analyses**. Hosted openly, it invites data scientists to share Jupyter Notebooks, R Markdown files, or Quarto scripts that demonstrate the use of SDMX in practical applications. Submissions are encouraged to be short, focused, and replicable. Each submission is typically three to five pages in PDF, accompanied by code.

The objectives are clear:

- **Showcase Practical Use**: Demonstrate how SDMX data can be accessed, processed, and analyzed in real workflows.
- **Promote Reproducibility**: Encourage users to publish code alongside analysis, enabling others to replicate and adapt.
- **Foster Collaboration**: Create a community where methods and insights can be exchanged, accelerating innovation.
- **Encourage Citation**: Recognize contributions and embed SDMX more firmly into academic and applied literature.

Importantly, *sdmx4ds* is not confined to official statistics alone. Submissions may combine SDMX data with other sources, illustrating how structured standards can be integrated into diverse analytic contexts.

For advanced users, *sdmx4ds* offers both a learning resource and a platform for visibility. For institutions, it signals that SDMX is no longer just about compliance and reporting — it is about enabling innovation, collaboration, and knowledge exchange in the era of data science.

6. Case Examples and Pathways Forward

The value of SDMX for data science is perhaps best illustrated through concrete examples:

[3] https://github.com/bis-medit-sdmxio/sdmx4ds

- **Macroeconomic Forecasting**: A central bank develops a forecasting model that automatically retrieves interest rate, inflation, and exchange rate data via SDMX APIs. Instead of manual updates, the model refreshes itself with each new release, ensuring forecasts remain current.
- **Policy Dashboards**: A national statistical office builds an interactive dashboard combining SDMX data from domestic and international sources. Policymakers can visualize trends in near real-time, supported by metadata that ensures indicators are interpreted consistently.
- **Climate and Financial Risk Analysis**: A research team integrates climate statistics, financial exposures, and socio-economic data, all structured through SDMX. The resulting models support stress testing of financial institutions under climate scenarios, a critical emerging policy area.
- **Collaborative Repositories**: Through *sdmx4ds*, a data scientist publishes a Jupyter Notebook demonstrating trade data analysis using SDMX APIs. Others adapt it to their contexts, creating a virtuous cycle of learning and reuse.

Examples of analyses published on *sdmx4ds*:

- EU Energy Price Analysis using data from Eurostat and ECB
- Exploring the impact of high cocoa prices on Cameroon's productivity
- Exploring the novel BIS central bank total assets data set

Each of these pathways and examples illustrates the same principle: by embedding rich metadata and standardization at the core, SDMX transforms data from a raw input into a reusable, interoperable asset.

Conclusion

The same SDMX that is already widely adopted by data producers can provide immense value to data users. By combining standardized data structures with rich, machine-readable metadata and programmatic access, SDMX allows analysts to discover, interpret, and integrate official statistics with far less manual effort. For data scientists, this translates into faster data ingestion, more reliable automated pipelines, and greater reproducibility of analysis. As the volume and complexity of available data continue to grow, broader awareness and use of SDMX among the data user community can significantly enhance the efficiency, comparability, and impact of empirical research and policy analysis.

sdmx.io and *sdmx4ds* are efforts to further disseminate the usefulness of SDMX and increase its community. By focusing on open source tools and on open research, these platforms also contribute public goods to the data science community.

References

Araujo, DKG (2023): "gingado: a machine learning library focused on economics and finance", BIS Working Papers No 1122.

Bogdanova, B, B Buffett, B Ligani, I Mustafi, S Nikoloutsos, R Schmidt and B Tissot (2025a): "SDMX adoption and use of open source tools", IFC Report No 17.

Bogdanova, B, M Erdem, B Ligani and O Sirello (2025b): "Leveraging generative AI for official statistics: opportunities, challenges and use cases", 65th ISI World Statistics Congress.

Buffett B and J Napoles (2024): "Enhancing data management through the Statistical Data and Metadata Exchange Standard", Asian Development Bank.

Sirello, O (2024): "Editing metadata with generative AI", background paper prepared for the UNECE Expert Meeting on Statistical Data Editing, Vienna, 7 October.

Wilkinson, M et al (2016): "The FAIR Guiding Principles for scientific data management and stewardship", Scientific data, v.3, n.1, p.1-9.

sdmx.io – where tools and learning resources come together to address official statistics use cases

Douglas Araujo and Brian Buffett

Bank for International Settlements

Rome, 18-20 February 2025

Context

- **Increasing demands for data**: pressure for more agile data organisations, processes, and software
  - AI and advanced analytics only add to the demand

- **SDMX** (**S**tatistical **D**ata and **M**etadata e**X**change) standard:
  - information model and technical architecture that support robust metadata-driven processes…
  - … like **data analysis** for one-off use (eg, research) or for recurrent use (eg, an internal dashboard)

- **SDMX** is reliable and trustable:
  - an ISO standard with a **well-defined governance**
  - Sponsored by BIS, ECB, EUROSTAT, ILO, IMF, OECD, UN, WB
  - Used routinely by various authorities to produce and disseminate **official statistics**

- But for most data scientists in economics, **SDMX is still unknown and has high barriers to entry**

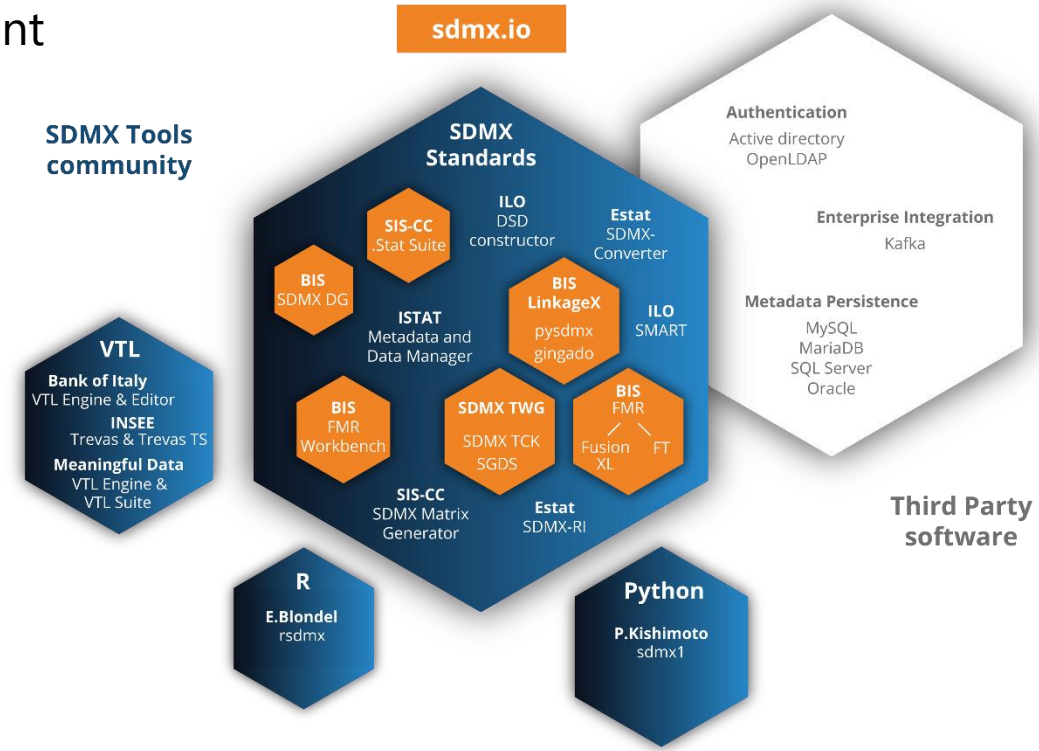
SDMX – Only for Data Producers?

- **Rich Metadata Support:** Metadata which describes datasets in detail, including variables, units, methodologies, and data sources. This helps data scientists:
  - Quickly assess data relevance.
  - Understand data structure without diving deep into raw files.
  - Ensure consistency across datasets for time series or comparative analyses.
- **Improved Data Discovery:** Metadata enables efficient searching and filtering of datasets based on specific criteria like time periods, geographic regions, or indicators.
- **Standardized Data Models:** Common formats for sharing statistical data, making it easier to integrate datasets from different sources (e.g., BIS, ECB, IMF, central banks).
- **Automation & API Integration:** APIs allow automated data retrieval and updates, which is crucial for dashboards or monitoring systems.
- **Interoperability:** Facilitates seamless data integration across various tools and platforms commonly used in data science (Python, R, BI tools).

Enter SDMX iO

- Goal is to facilitate the use of SDMX-based solutions
  - throughout the data lifecycle (eg, production -> dissemination -> use)
  - and for organisations of all sizes and stage of development

- To achieve this goal, **sdmx.io** is an ecosystem of resources:
  - open-source tools,
  - pre-configured containerised environments,
  - learning material (tutorials, webinars),
  - and more



What about data science?

- SDMX is typically used by authorities responsible for producing and disseminating statistics

- But it has ample potential for use in data science as well (Araujo, 2023):
  - Standard definitions of data and metadata modelling can help users find relevant data…
  - … from reliable sources …
  - … in a programmatic and reproducible way.

- **sdmx.io** will increasingly also cater to this target user group of end-users / data scientists

- Today, we are introducing a new series **sdmx4ds** (SDMX for Data Science)

sdmx4ds: a repository of data science work using SDMX

- Data scientists can share their work that uses SDMX (and potentially other data sources)
  - Citation is encouraged

- Goal is to publish Jupyter Notebooks in python, R, Julia showcasing:
  - an interesting analysis that uses data downloaded via SDMX
  - How to download data with SDMX? gingado sdmx1

- Submission process:
  - All public, via GitHub
  - Submissions should be short (circa 3-5 pages of PDF)
  - A template offers guidance on structure

Live demo of the submission process

Next steps

- Call for papers – coming out soon. Keep an eye on **sdmx.io**
- Publication
  - Lightly-edited Jupyter notebooks / Rmarkdown / quarto file in GitHub repo
- Easy for other users to fork and adapt to their own cases
  - 

 button allows for quick replication and customisation by readers

Thank You!