Statistical Data and Metadata eXchange

Use of SDMX **to Enable AI-ready Data**

10th SDMX Global Conference
Smarter Data for Better Insights
Rome, Italy
29 September – 3 October 2025

Presented by: Lizette Traconis Lugo

INEGI

inegi.org.mx

WHY SDMX IS KEY FOR AI-READY DATA

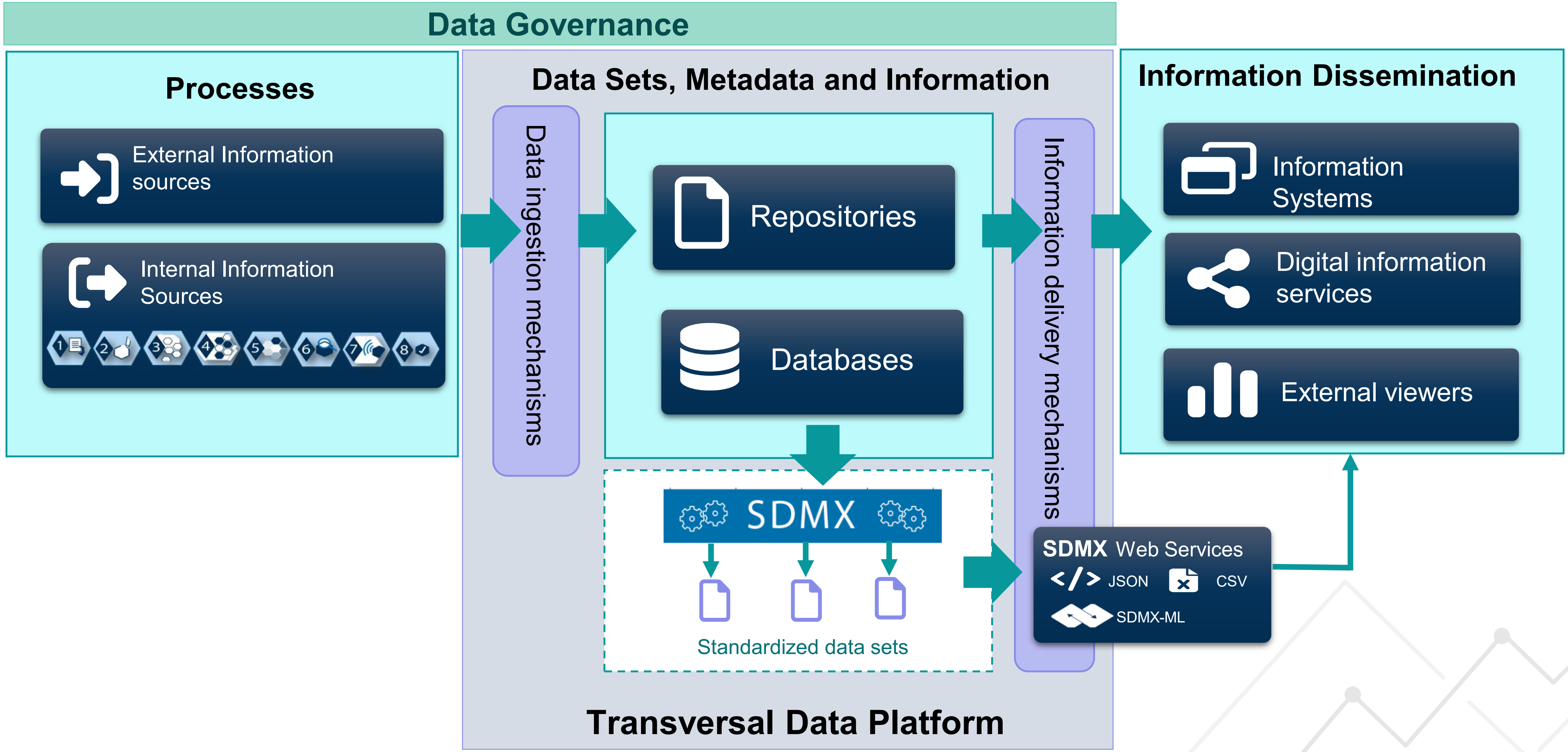
THE CHALLENGE

Artificial Intelligence (AI) systems, especially Large Language Models (LLM), require data that is **structured**, **contextualized** and **machine readable**. But many data sources still lack standardization, metadata and interoperability limiting AI's effectiveness.

SDMX is an international standard designed to improve exchange in official statistics. It provides a foundation form AI-ready data through:

- Structured data models.
- Rich metadata.
- Machine readable formats.
- Semantic clarity with standard concepts.
- Interoperability across systems.
- Scalability and automation for efficient workflows.

TRANSVERSAL **DATA PLATFORM**



Data Governance

Processes

External Information sources

Internal Information Sources

Data Sets, Metadata and Information

Data ingestion mechanisms

Repositories

Databases

SDMX

Standardized data sets

Information delivery mechanisms

SDMX Web Services

JSON

CSV

SDMX-ML

Transversal Data Platform

Information Dissemination

Information Systems

Digital information services

External viewers

USE CASES: **APPLYING AI TO SDG DATA WITH SDMX**

Demonstrate how SDMX enables AI-driven exploration of SDG data through two concrete use cases using two distinct AI architectures, each solving a different data interaction challenge.

1 **Information retrieval**

Use AI to answer natural language questions by retrieving relevant information from SDMX datasets and related documents.

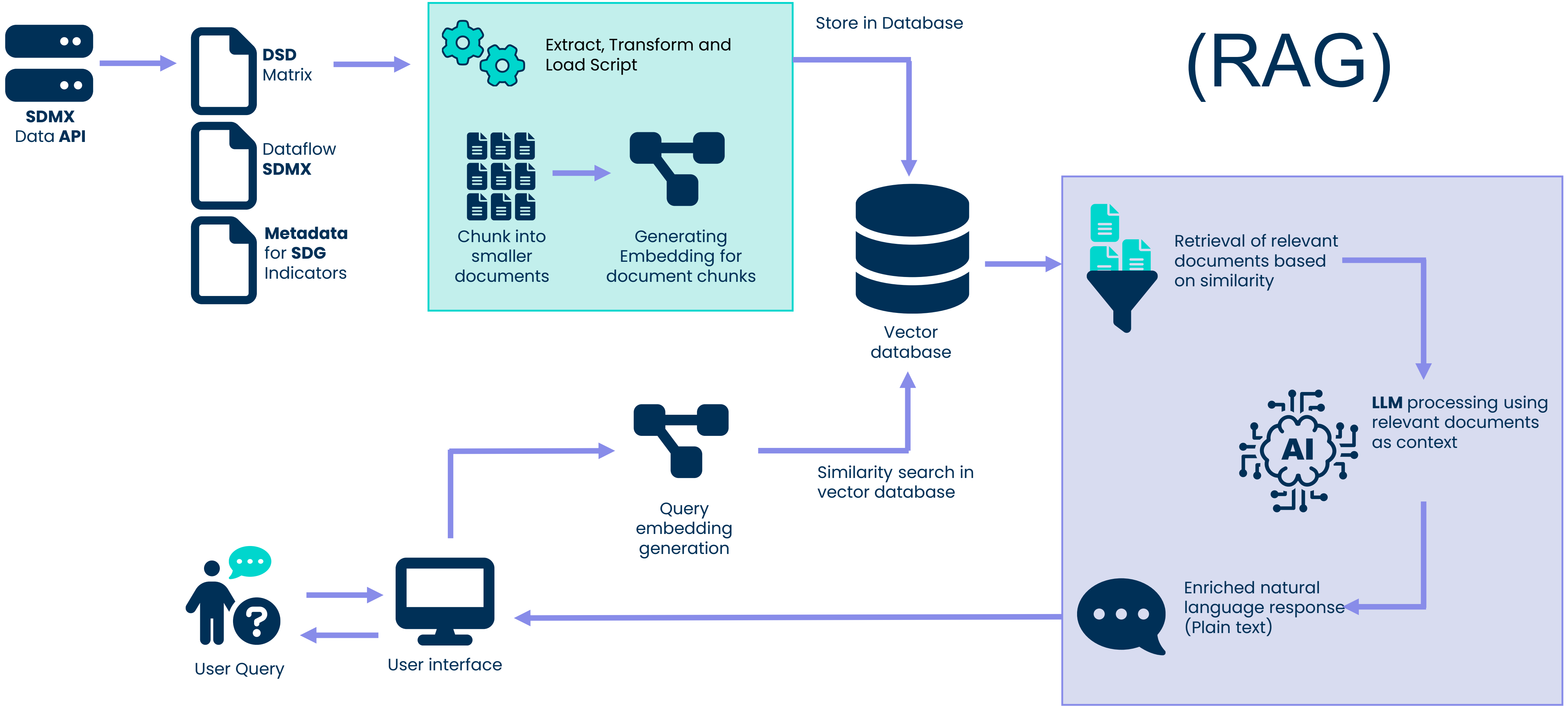
2 **AI Agent for dynamic queries**

Understand a user's natural language question, identify the correct parameters based on SDMX Data Structure Definition (DSDs), and call the SDMX REST API to return structured results, such tables or charts.

Knowledge Base:	**Data Sources:**
Metadata list	Metadata was collected via web scraping from: https://agenda2030.mx/
DSD Matrix (Data Structure Definitions)	DSD Matrix was obtained from the UN SDMX Working Group: https://unstats.un.org/sdgs/iaeg-sdgs/sdmx-working-group/
Data Flow from SDG datasets	SDG Data Flows were retrieved from the SDMX service of the SNIEG-INEGI: https://sdmx.snieg.mx/dataflows

1 INFORMATION RETRIEVAL **ARCHITECTURE**

(RAG)



SDMX
Data **API**

DSD
Matrix

Dataflow
SDMX

Metadata
for **SDG**
Indicators

Extract, Transform and Load Script

Chunk into smaller documents

Generating Embedding for document chunks

Store in Database

Vector database

Retrieval of relevant documents based on similarity

LLM processing using relevant documents as context

Enriched natural language response (Plain text)

User Query

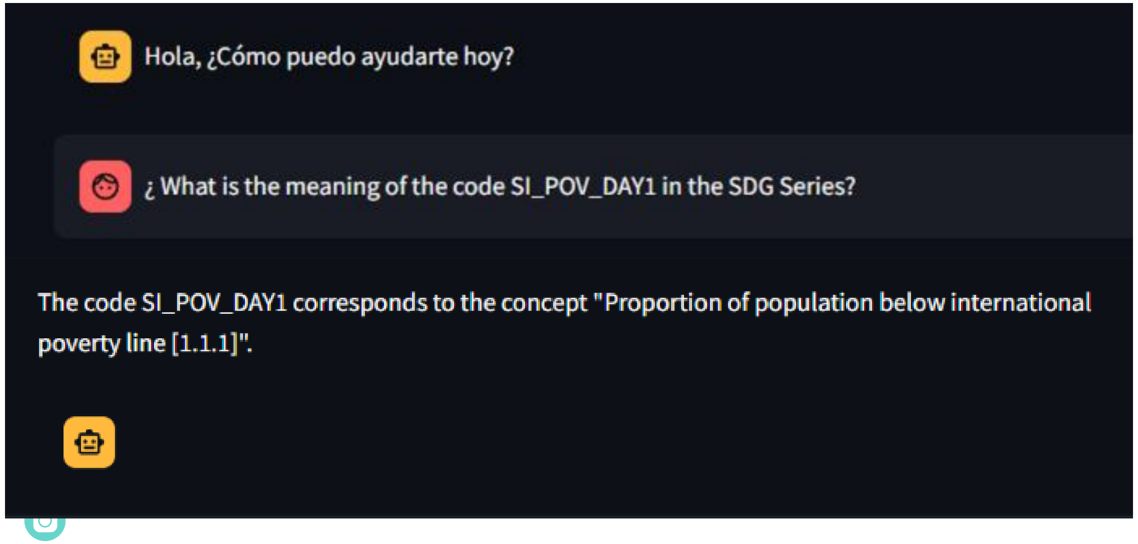
User interface

Query embedding generation

Similarity search in vector database

USE CASE: RESULTS

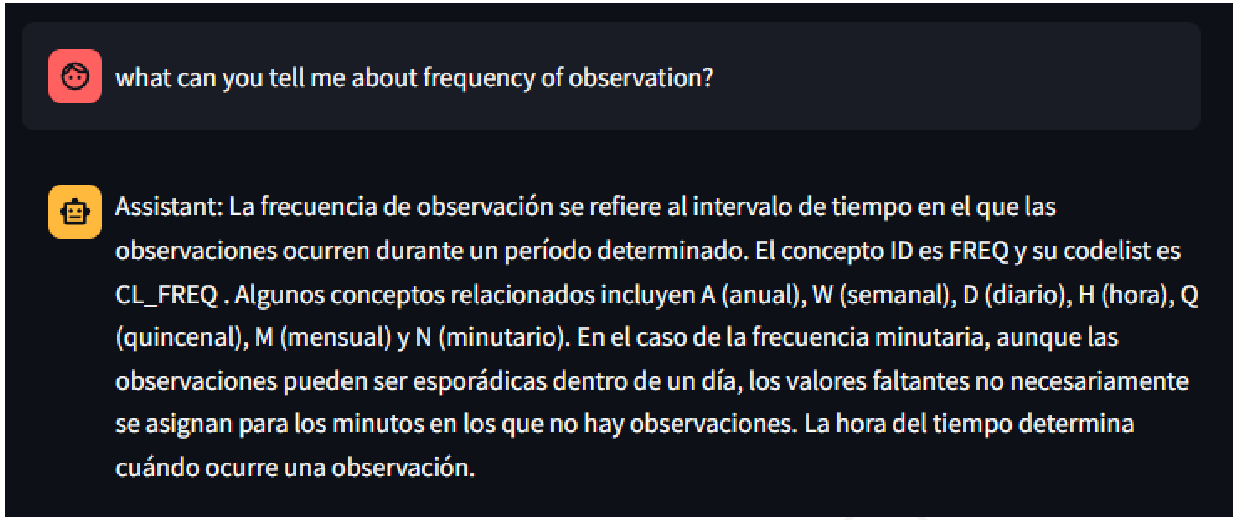
Knowledge base **DSD** Matrix



Hola, ¿Cómo puedo ayudarte hoy?

¿ What is the meaning of the code SI_POV_DAY1 in the SDG Series?

The code SI_POV_DAY1 corresponds to the concept "Proportion of population below international poverty line [1.1.1]".



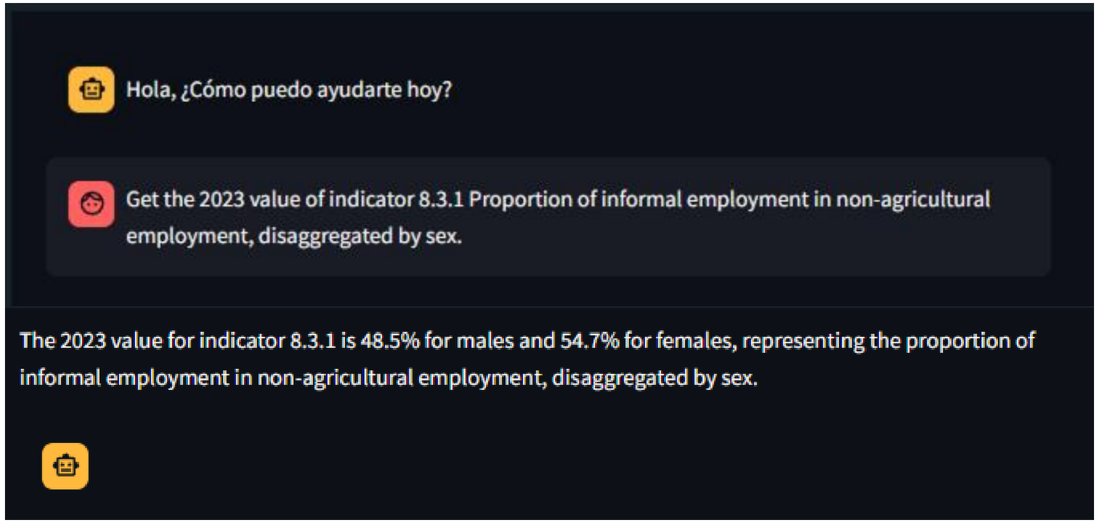
what can you tell me about frequency of observation?

Assistant: La frecuencia de observación se refiere al intervalo de tiempo en el que las observaciones ocurren durante un período determinado. El concepto ID es FREQ y su codelist es CL_FREQ . Algunos conceptos relacionados incluyen A (anual), W (semanal), D (diario), H (hora), Q (quincenal), M (mensual) y N (minutario). En el caso de la frecuencia minutaria, aunque las observaciones pueden ser esporádicas dentro de un día, los valores faltantes no necesariamente se asignan para los minutos en los que no hay observaciones. La hora del tiempo determina cuándo ocurre una observación.

USE CASE: RESULTS

Knowledge base

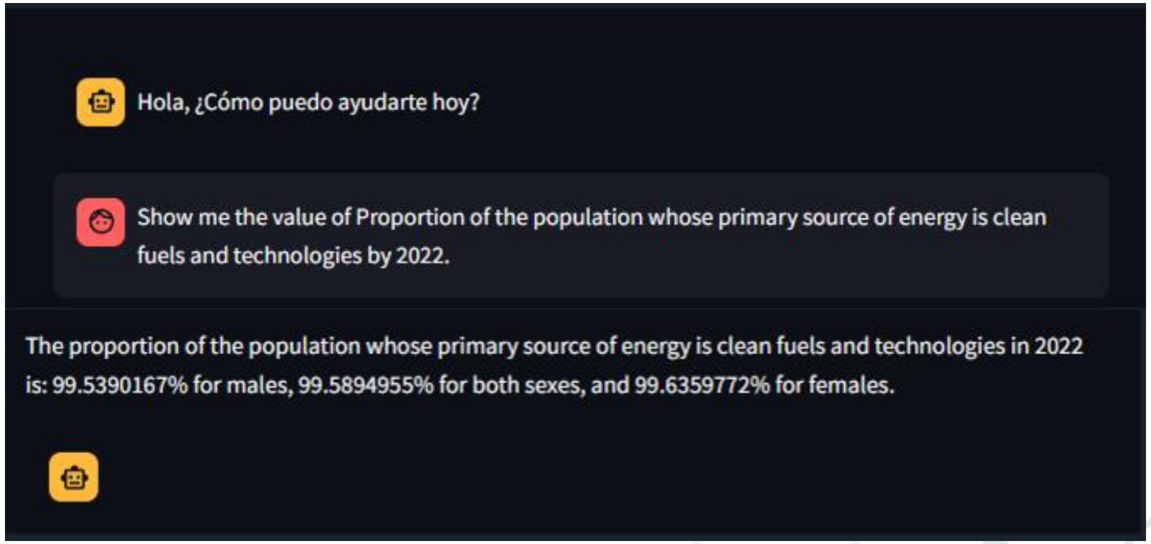
Dataflow
SDMX



Hola, ¿Cómo puedo ayudarte hoy?

Get the 2023 value of indicator 8.3.1 Proportion of informal employment in non-agricultural employment, disaggregated by sex.

The 2023 value for indicator 8.3.1 is 48.5% for males and 54.7% for females, representing the proportion of informal employment in non-agricultural employment, disaggregated by sex.

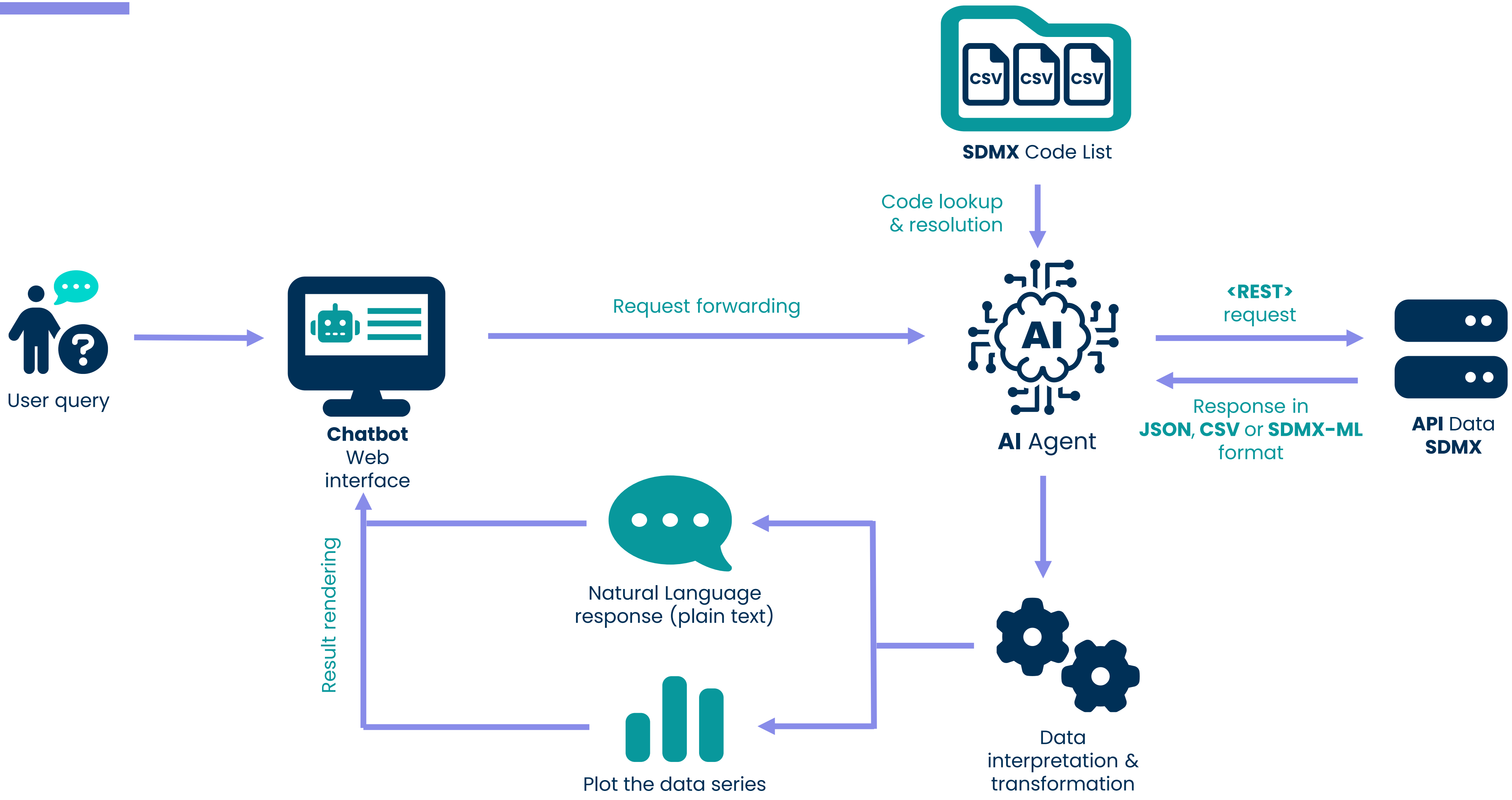


Hola, ¿Cómo puedo ayudarte hoy?

Show me the value of Proportion of the population whose primary source of energy is clean fuels and technologies by 2022.

The proportion of the population whose primary source of energy is clean fuels and technologies in 2022 is: 99.5390167% for males, 99.5894955% for both sexes, and 99.6359772% for females.

2 AI AGENT ARCHITECTURE **FOR DYNAMIC SDMX QUERIES**



SDMX Code List

Code lookup & resolution

User query

Request forwarding

Chatbot Web interface

AI

AI Agent

‹REST› request

Response in **JSON**, **CSV** or **SDMX-ML** format

API Data **SDMX**

Natural Language response (plain text)

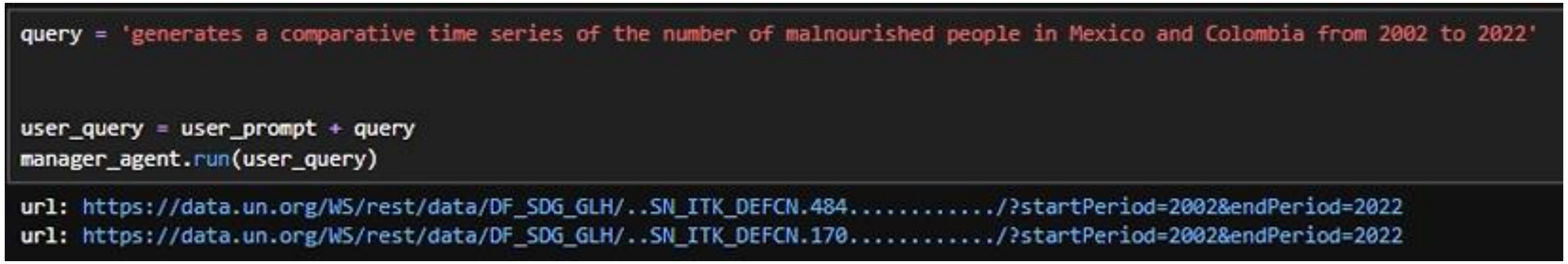
Result rendering

Plot the data series

Data interpretation & transformation

inegi.org.mx

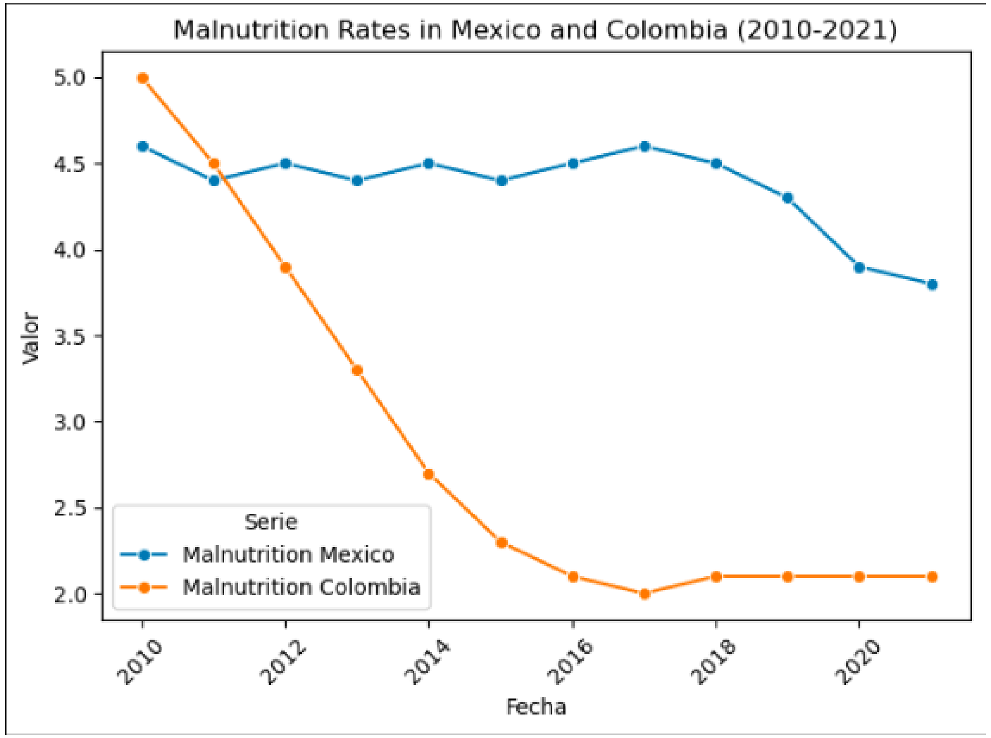
USE CASE: RESULTS



```
query = 'generates a comparative time series of the number of malnourished people in Mexico and Colombia from 2002 to 2022'


user_query = user_prompt + query
manager_agent.run(user_query)
```

```
url: https://data.un.org/WS/rest/data/DF_SDG_GLH/..SN_ITK_DEFCN.484............/?startPeriod=2002&endPeriod=2022
url: https://data.un.org/WS/rest/data/DF_SDG_GLH/..SN_ITK_DEFCN.170............/?startPeriod=2002&endPeriod=2022
```



The graph shows that malnutrition rates in Mexico have decreased from 4.6% in 2010 to 3.8% in 2021, while Colombia's rates have also decreased from 5.0% to 2.1% over the same period.

The generated graph illustrates the trends in malnutrition rates for both Mexico and Colombia from the year 2010 to 2021. Mexico's malnutrition rate started at 4.6% in 2010 and showed a slight decrease over the years, reaching 3.8% in 2021. Colombia's malnutrition rate was higher at 5.0% in 2010 and consistently declined to 2.1% by 2021. Both countries have made progress in reducing malnutrition rates, with Colombia showing a more

significant decline compared to Mexico.

HOW **SDMX** ENABLES **EFECTIVE RETRIEVAL**

Leveraging standardized metadata and structured data to improve document relevance, context preservation, an AI generated responses.

DATA PREPARATION

- Facilitates automated extraction through standard API
- Simplifies cleaning and transformation
- Standardized inputs allow using a single ETL process for all dataflows and make easier to automate updates.

DATA RETRIEVAL

- Metadata enhances semantic understanding during similarity search.
- Supports standardized query
- Structured data leads to accurate and meaningful AI responses.
- Facilitates alignment between user queries and the data content

FINAL **REMARKS**

"Getting data ready for AI is not just a technical upgrade, it´s a transformation in how we manage and share information"

- SDMX and well-governed data architectures are strategic enablers for AI-ready data ecosystems.
- AI-ready data requires structured, contextualized and interoperable information, which SDMX helps to provide.
- Data governance is essential for trust and machine usability because ensures quality, consistency, traceability.
- NSOs are in a unique position. As trusted data producers can lead the way in creating AI-ready ecosystems that are ethical, transparent and reliable.
- By following global standards, improving governance and building skills, we're taking important steps to better work with data an AI.

¡Thank you!

Acknowledgment:

Use cases were developed by the IT team at INEGI, lead by Juan Muñoz López.

- Juan Pablo Navarro Romo (juan.romo@inegi.org.mx)
- Luis Alejandro Arellano López (alejandro.arellano@inegi.org.mx)
- Mayra Alejandra Ávalos Alba (mayra.avalos@inegi.org.mx)
- Luis Daniel Cuéllar Garrido (daniel.cuellar@inegi.org.mx
- Abel Soto Valdez (abel.soto@inegi.org.mx)
- Adrián Alejandro García Richarte (adrian.garcia@inegi.org.mx)

CONOCIENDO MÉXICO

800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGIINFORMA