

Is SDMX data AI-ready? Is AI SDMX-ready?

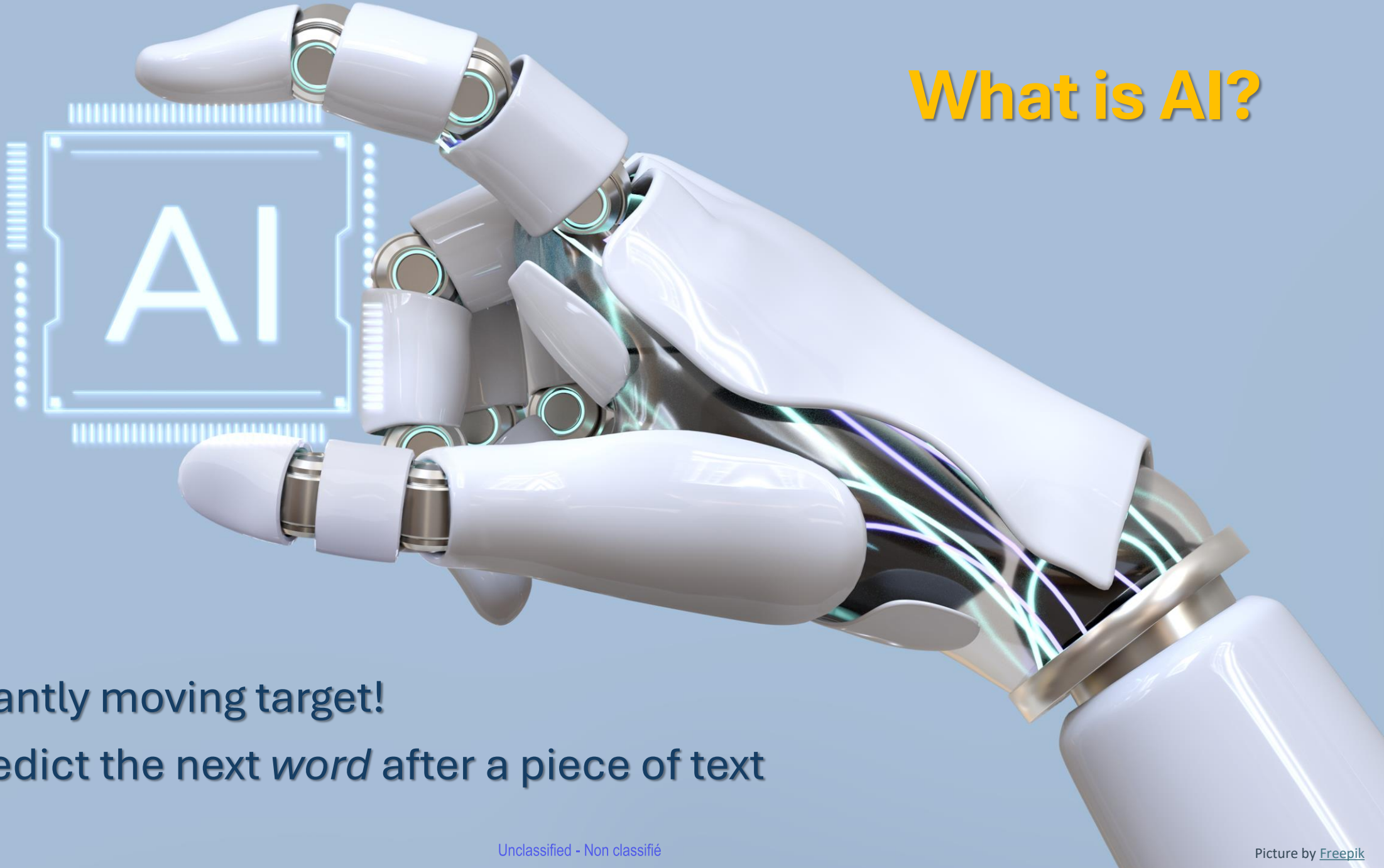
SDMX Global Conference 2025

Jens Dossé (SDMX-TWG, OECD)

Outline

1. What is AI?
 2. SDMX – AI use cases
 3. Criteria for AI-readiness of statistical data
 4. AI-readiness of SDMX
 5. TWG closing gaps in AI-readiness of SDMX
 6. SDMX-readiness of AI
- Annex - Capacity building for AI-readiness
- Annex - Overcoming challenges with AI

What is AI?



A constantly moving target!

LLM: predict the next *word* after a piece of text

Common SDMX – AI use cases

Making statistical
data more
findable/accessible

Making AI tech
leaders use our
data appropriately

Today

What would be the criteria for statistical data to be AI-ready?

Copilot ⓘ

To make statistical data **AI-ready**, it must meet several criteria that ensure it can be effectively used for training, validating, or deploying AI models. These criteria span **data quality**, **structure**, **context**, and **compliance**. Here's a breakdown:

1. Data Quality

- **Completeness:** Minimal missing values; if missing, they should be handled or imputed appropriately.
- **Accuracy:** Data should reflect real-world values or measurements reliably.
- **Consistency:** Uniform formats, units, and naming conventions across datasets.
- **Timeliness:** Up-to-date data, especially for time-sensitive applications.

2. Data Structure & Format

- **Structured Format:** Preferably in tabular form (e.g., CSV, Excel, SQL) with clear column headers and data types.
- **Standardized Encoding:** Categorical variables should be encoded (e.g., one-hot, label encoding) for model compatibility.
- **Normalized/Scaled Values:** Numerical data often needs normalization or standardization for model performance.

Provide checklist as PDF.

Help prepare dataset for ChatGPT.

Explain semantic clarity importance.



Message Copilot



Criteria for AI-readiness of statistical data (according to AI 🤖)

Copilot

To make statistical data AI-ready, it must meet several criteria that ensure it can be effectively used for training, validating, or deploying AI models. These criteria span data quality, structure, context, and compliance. Here's a breakdown:

1. Data Quality

- **Completeness:** Minimal missing values; if missing, they should be handled or imputed appropriately.
- **Accuracy:** Data should reflect real-world values or measurements reliably.
- **Consistency:** Uniform formats, units, and naming conventions across datasets.
- **Timeliness:** Up-to-date data, especially for time-sensitive applications.

2. Data Structure & Format

- **Structured Format:** Preferably in clear text-readable tabular form (e.g., CSV, Excel, JSON) or embedded in well-structured tables within documents with clear column headers and data types, consistent rows, and no nested tables, merged cells or ambiguous formatting.
- **Standardized Encoding:** Categorical variables should be encoded (e.g., one-hot, label encoding) for model compatibility.
- **Multilingual clarity:** If multilingual, clearly separate or label languages. Use UTF-8 encoding to ensure the file uses standard character encoding to avoid misinterpretation of accents or symbols.
- **Normalized/Scaled Values:** Numerical data often needs normalization or standardization for model performance.
- **Consistent formatting:** Dates, numbers, and categories should follow a consistent format.

3. Semantic Clarity through Metadata & Documentation

- **Clear Definitions:** Each column name or variable should be meaningful (or have a documented meaning using understandable categories) and include unit and scales.
- **Natural language compatibility:** Avoid jargon-heavy or overly technical formatting unless explained.
- **Data Provenance:** Information about how and where the data was collected, processed and published.
- **Versioning:** Track changes to datasets over time for reproducibility.
- **Narrative context:** If possible, embed the data within a short explanatory paragraph.
- **Example queries:** Optionally include sample questions or use cases to guide interpretation.

4. Contextual Relevance

- **Domain-Specific Suitability:** Data should be relevant to the problem the AI is solving.
- **Balanced Representation:** Avoid bias by ensuring diverse and representative samples.
- **Temporal Context:** For time-series or evolving data, timestamps and sequence integrity are crucial.

5. Legal & Ethical Compliance

- **Privacy & Anonymization:** Personally identifiable information (PII) must be removed or anonymized.
- **Consent & Licensing:** Ensure data usage complies with legal agreements and ethical standards.
- **Bias Auditing:** Check for and mitigate harmful biases in the dataset or note any known limitations or biases in the dataset.

6. Technical Compatibility

- **Machine-Readable Format:** Easily ingestible by AI pipelines (e.g., JSON, Parquet, CSV).
- **Interoperability:** Compatible with common AI frameworks (TensorFlow, PyTorch, Scikit-learn).
- **Scalability:** Suitable for processing at scale, especially for big data applications.

AI readiness...

**= AI optimisation
(extending SEO)**

**~ human
readiness**

So, is SDMX data AI-ready?



1) Standardised container:



SDMX IM + API + formats for data transmission and storage &



Model Context Protocol (MCP) as middleware to AI agents
→ discoverable, machine-accessible and
up to date data under public domain/open license
→ timeliness, accessibility, system interoperability

So, is SDMX data AI-ready?



2) Standardised content:



SDMX guidelines for high-quality data modelling
→ well-structured, well-identified, semantically
versioned data → coherence



Statistical quality framework
→ relevance, accuracy, credibility → trust



Data harmonisation within the organisation and at international level
→ maximal data consistency & comparability

So, is SDMX data AI-ready?



3) Augmented content:

- ⚠ Descriptive metadata: clear definitions and context at all levels, explanatory documentation (incl. for legitimate anomalies/biases)
- ⚠ Standardised semantic relationships between the data beyond harmonisation: knowledge graphs (“meaning graph”) as linked open data
- ⚠ Ontology-based international standardization of meaning
 - well understandable inter-related data
 - interpretability, transparency → trust



Persistent resolvable
SDMX identifiers
(URN) thanks to the
brand-new SGDS
→ SDMX fulfills
FAIR principles

Fallback: SDMX URN
registration with IANA



LOD readiness:
Allow for flexible
semantic linking
mechanism
between identifiable
artefacts and
external contexts



Expressing the
SDMX Information
Model via an SDMX
ontology linked with
other existing
ontologies



Statistical Data and Metadata Exchange

TWG closing gaps in AI-readiness of SDMX

Unclassified - Non classifié

And is AI also SDMX-ready?

AI agents can connect to:

What else is needed?

 Google announced that this September Data Commons will allow importing data in SDMX!

SDMX-API

SDMX-
powered
**dataset
catalogues**
on the web

SDMX-
powered
**Model-
Context-
Protocol
(MCP)**
services

Collaboration
& governance
Compliance
with agreed
rules

Returned
services:
visibility,
feedback &
support

Annex - Capacity building for AI-readiness: institutions, systems, individuals

Institutions (governance & legal frameworks)

- Strategic vision, data governance, innovation, data quality frameworks
- Community of practice approach
- International collaboration, governance at global level
- Agility

Systems (infrastructure & investment)

- SDMX-based, user-centric data lifecycle management platforms, sharing, co-operation
- Workflows, pipelines, automation

Individuals (workforce skills)

- In-depth knowledge of high-quality data modelling (SDMX) and awareness of required metadata content for AI readiness
- Online SDMX training courses

Annex - Overcoming common challenges in implementing AI-based systems

- Find expertise for the implementation → Partner with advanced public or private organisations
- High costs (depending on model and usage), high environmental impact → Cost-benefit analysis
- Keeping up with quickly advancing technology → Keep solutions simply and cost-effective, maintain and develop expertise, be ready for quick changes
- Probabilistic nature of genAI (LLMs) lead to hallucinations (making things up!)
 - → Avoid misinformation to maintain credibility (zero hallucinations!)
 - → Put responsibility on the user: Let the user validate/decide; choose the document, dataset, indicator, ...
- Findability use case:
 - Data availability issues (sparseness/gaps) → Combine AI with checks using SDMX data availability
 - Knowing what data users really need! → Test prototypes with real users, collect feedback

JOINT ST. PETERSBURG / LONDON 2025 0905/2900 St. Petersburg

In one word, what do NSOs need most to become AI-ready?

standardize the method

culture of experimenting

strategic target

just enough governance

sdmx

rule

partnership

money

ai lab

collaboration

open source

hybrid

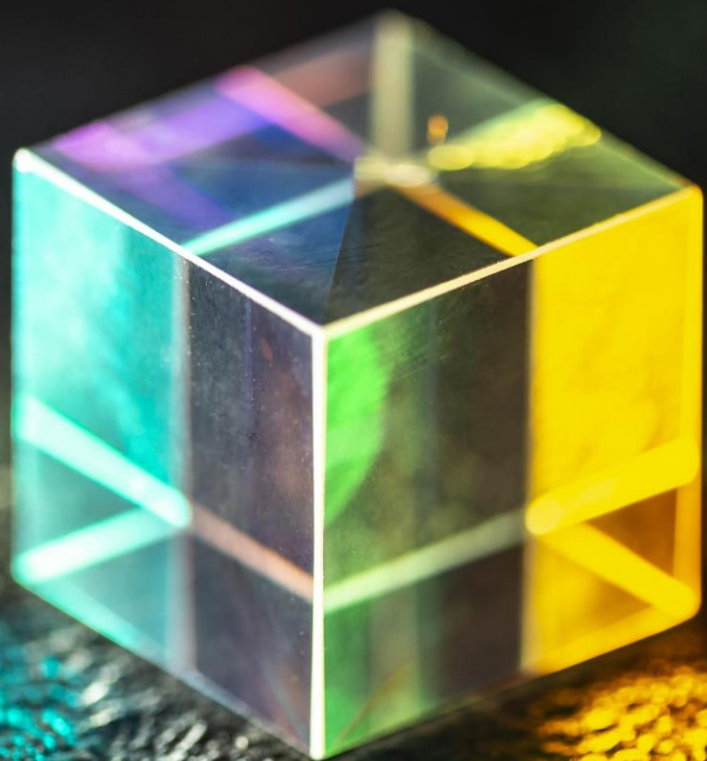
buy-in

energy

guideline



OECD-KOSTAT Global Conference 2025



Thank you!

Jens Dossé, OECD & SDMX-TWG