Ministry of Data and Statistics (MoDS)

Data for AI : Making Official Statistics AI-ready

Vision and streategy of MoDS Korea

2026. 2. 2.

Sunhi Yi
Director of AI Innovation Division
Ministry of Data and Statistics(formerly Statistics Korea)

Agenda

Overview of Today's Presentation

01 **Background - Our Experiences**

LLM-based Statistics Chatbot PoC & Insights

02 **Background – AI Era Challenge**

Vision, Strategic Pillars, and Key Strategic Plans for the AI Era

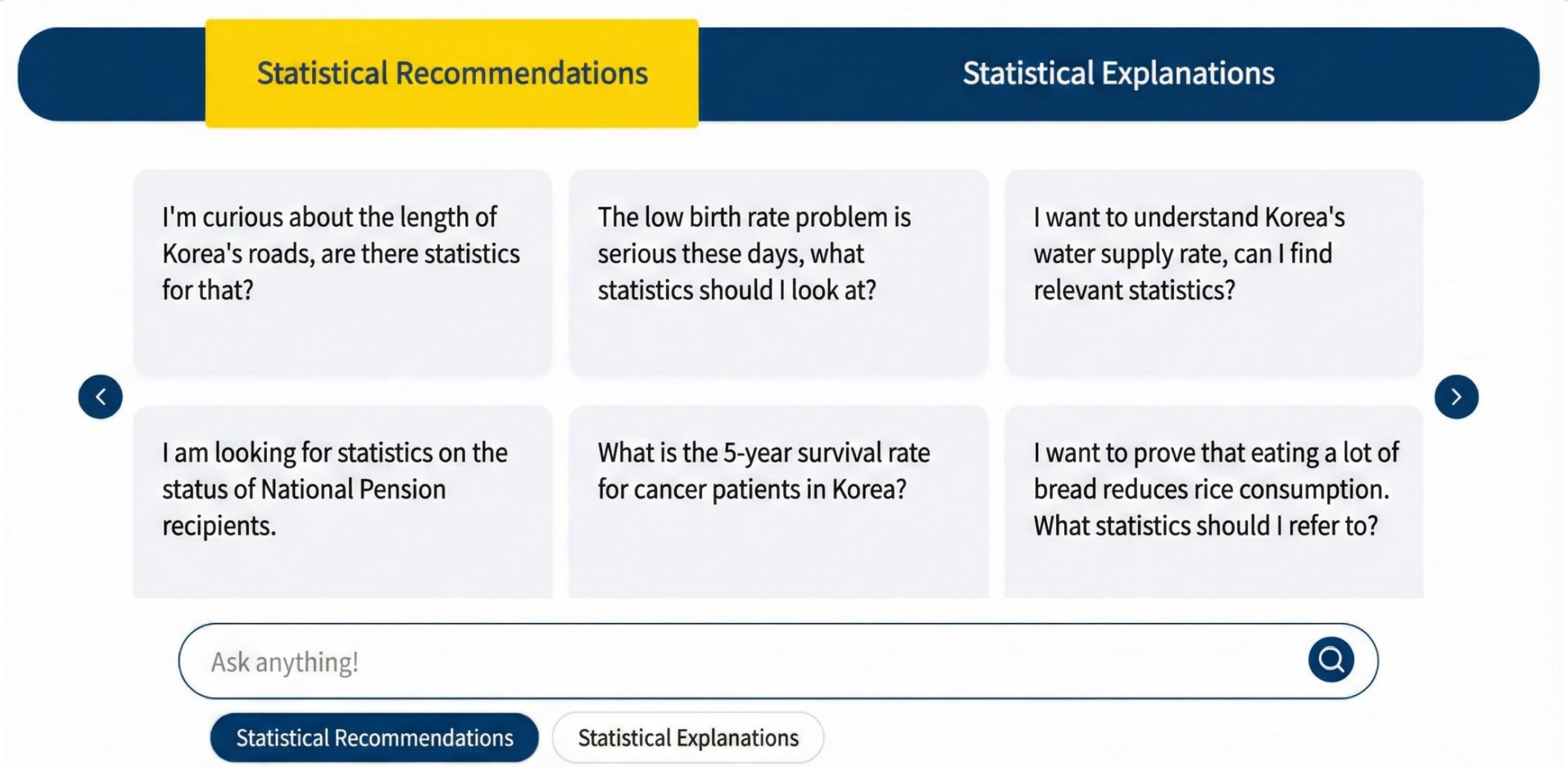
03 **AI-Friendly Metadata Construction**

Detailed walkthrough of our plan

04 **Closing & Q&A**

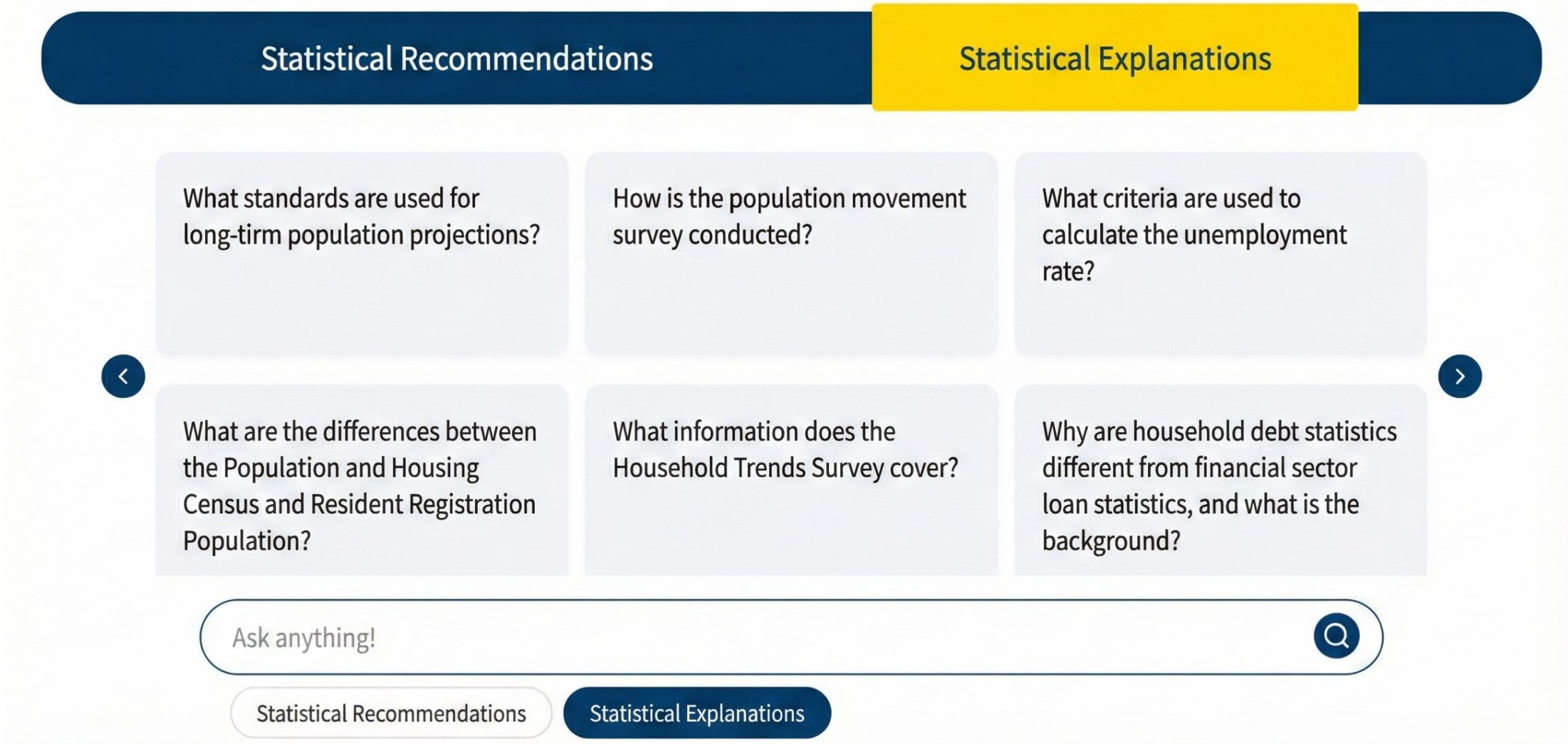
Discussion on international cooperation

LLM-based Statistics Chatbot



*** Currently available in Korean only. This image is a translated version using NanoBanana**

LLM-based Statistics Chatbot



*** Currently available in Korean only. This image is a translated version using NanoBanana**

Our Experiences

LLM Chatbot Implementation Insights

Primary Strength

Natural Language Access: Successfully lowered barriers by enabling users to query statistics using everyday language without knowing specific technical terms.

Root Cause

Insufficient Semantic Metadata: Currently, AI must infer meaning solely from table names or variables. Without deep semantic context, it struggles to "swim" through complex statistical concepts.

Key Takeaway

This experience highlighted the critical importance of AI-Friendly AI-Friendly Metadata, leading to the core initiatives detailed in detailed in Section III.

Challenges & The Paradox

Semantic Search Difficulty

Without exact names, finding relevant statistics by "meaning" alone remains difficult. The AI lacks the contextual map to link vague intent to precise data.

Frequent Hallucinations

The model frequently recommends fake or mismatched statistics ("hallucinations") when it cannot confidently bridge the gap between user query and database schema.

The Reasoning Paradox

To curb hallucinations, we restricted the LLM's role. However, this created a vicious cycle where we could not leverage its reasoning capabilities, reducing it to a simple keyword matcher.

AI ERA CHALLENGES

Official Statistics in the Age of AI

Shift in Touchpoints

Users no longer search for tables on portals. They ask questions to Chatbots and expect immediate answers. The "First Contact" has changed from Official Portals to AI Interfaces.

Emerging Risks

Hallucination & Misinformation AI often relies on secondary sources (blogs, news) rather than official DBs, leading to the distribution of incorrect data.

Loss of Trust Even if the AI is immature, the circulation of wrong statistics undermines the credibility of the official agency.

Why API is Not Enough

API Access Limitations

Providing API access to raw data doesn't solve the problem. AI is fundamentally an inference engine based on data, but it lacks context.

Context Blindness

Without understanding "Context of Numbers" (e.g., denominator definitions, population scope), AI cannot correctly interpret or compare variables.

The Need for "Meaning"

We need AI-Friendly Metadata that explicitly describes semantic relationships, enabling AI to "read" the rules of comparability and conversion.

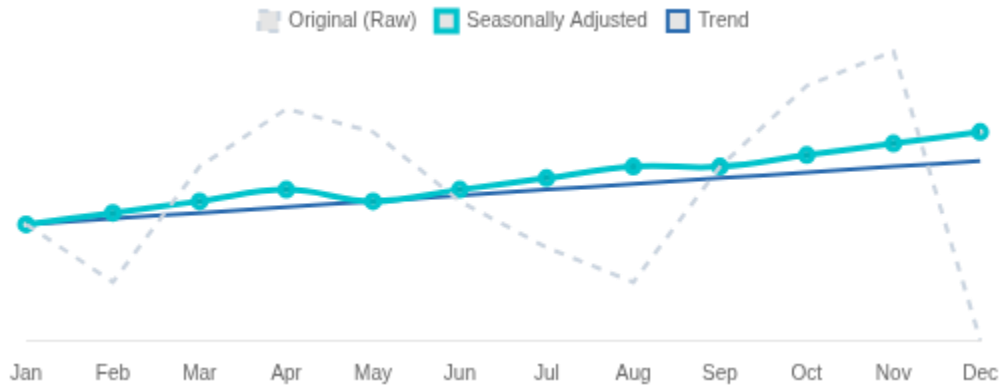
Case Study — Series Type

Metadata Case Study

Original vs. Seasonally Adjusted vs. Trend

The Problem

Official statistics often present multiple versions of the same indicator. Mixing Original (YoY), Seasonally Adjusted (MoM), and Trend series without context leads to conflicting interpretations by AI.



Conflicting AI Interpretation

Original Series:
"Economy is declining" (-2.1%)

Seasonally Adjusted:
"Economy is recovering" (+0.5%)

The Solution: Explicit Metadata

Provide machine-readable metadata defining the Series Type and Usage Rules. This allows AI to select the correct series for the user's specific question.

```
<Variable id="Unemployment_Rate" >
  <Attribute id="series_type" >Seasonally Adjusted </Attribute>
  <Attribute id="recommended_use" >Short-term Analysis (MoM)
  </Attribute>
  <Relation type="derivedFrom" >Unemployment_Rate_Original </Relation>
</Variable>
```

AI Selection Logic Rules

User Query Intent	Recommended Series
"Current momentum?"	✓ Seasonally Adjusted
"Long-term direction?"	✓ Trend Cycle
"Actual volume/count?"	✓ Original (Raw)

Limitations of Existing Standards — SDMX

What SDMX Does Well

SDMX (Statistical Data and Metadata eXchange) is the global standard for exchanging statistical data, sponsored by BIS, ECB, Eurostat, IMF, OECD, UN, and World Bank.

KEY STRENGTHS

Structure: Excellent DSDs, Code Lists, Concepts

Exchange: Standardized API & Formats (XML/JSON)

Adoption: Widely used by international organizations

Why SDMX Alone Is Not Enough for AI

The AI Gap

Dataset-Centric Structure

SDMX locks concepts within specific datasets. AI cannot easily search for a variable's meaning (e.g., "unemployment rate") across an entire agency without traversing every dataset structure.

No Semantic Reasoning

It defines "where" data is, not fully "what" it means in context. Lacks explicit rules for AI to automatically judge comparability (e.g., denominator definitions, population coverage).

Registry Limitations

Registries act as structural catalogs, not knowledge graphs. They do not support the semantic inference or transformation logic required for natural language queries.

The Semantic Gap

Standards Gap Analysis

Structure vs. Meaning

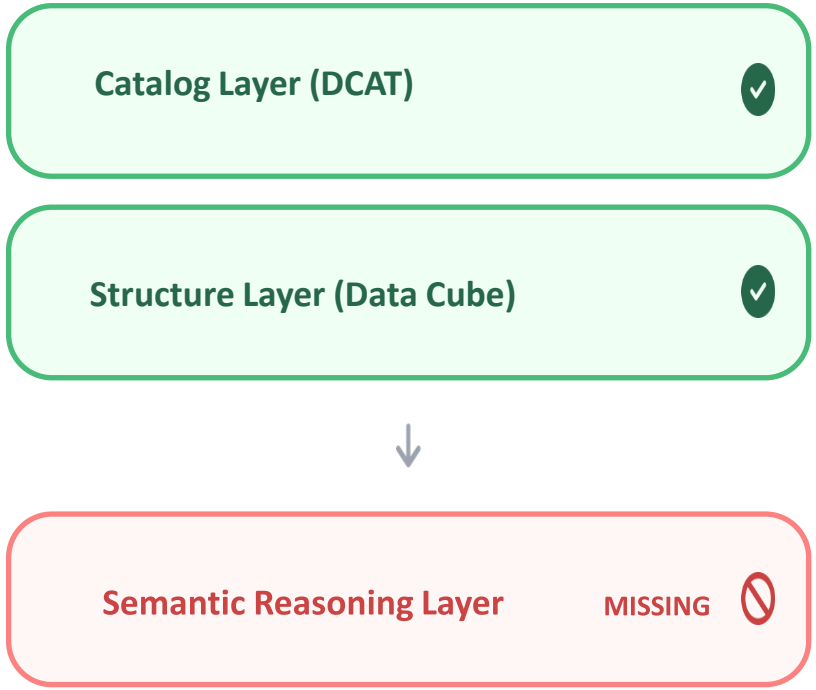
The Semantic Gap

Existing standards define "Structure" but lack "Meaning" required for AI reasoning.

CRITICAL MISSING FEATURES:

- ✗ No Automated Comparability: Cannot judge if "Unemployment A" matches "Unemployment B" (e.g., 15+ vs 15-64).
- ✗ No Transformation Logic: No machine-readable rules for re-referencing or unit conversion.
- ✗ Variable-Level Granularity: Metadata often stops at Dataset level, missing variable nuances.

The AI Reasoning Stack



*AI cannot deduce comparability without this layer

Metadata Standard Requirements (1–4)

01

Agency-wide Common Registry

Transition from dataset-specific metadata to an agency-wide conceptual registry. Shared definitions enable AI to traverse across different datasets to find related variables (e.g., all "Unemployment Rate" instances) regardless of the source survey.

02

Variable-level Semantic Standards

Metadata must operate at the variable level, not just the dataset level. Detailed attributes such as denominator definitions, population scope (age/region), and measurement units must be explicitly defined to allow AI to judge comparability.

03

Multi-dimensional Concept Networks

Leverage SKOS/XKOS to model complex relationships. Concepts like "Youth Unemployment" are linked to "Youth" (hierarchy), "Unemployed" (status), and "Labor Force" (denominator), forming a dense knowledge graph for AI inference.

04

GSIM Alignment & Reuse

Align with the Generic Statistical Information Model (GSIM). Definitions established in the design phase should flow through production (GSBPM) to dissemination, ensuring consistent explanations across portals, APIs, and reports.

Metadata Standard Requirements (5–7)

Minimal but Sufficient Efficiency

05

Define only the critical classes required for AI to judge comparability: Denominator, Population, Unit, Reference Period, and Series Type. Avoid excessive metadata that hinders adoption.

Multi-Depth Profiles Flexibility

06

Provide tiered implementation options: • Light Profile: Extended StatDCAT for basic discovery. • Deep Profile: Fully GSIM-aligned for precise comparability.

Cross-Agency Linkage Interoperability

07

Metadata must support ontology-based linking between agencies. AI should be able to traverse from Agency A's variable to Agency B's related variable to infer relationships across the ecosystem.

TDO-KR Delivery Mechanisms

Trusted Data Observatory Korea: Distributing AI-Ready Official Statistics

Infrastructure: Planned

LOD Publication

Machine Readable Foundation

Linked Open Data (RDF):Publishing official statistics as graph data for semantic linking.

SPARQL Endpoint:Standard query interface enabling complex semantic queries by AI agents.

Ontology Guides:Documentation teaching AI how to interpret variable relationships.

RDF/TTL | SPARQL | SKOS | XKOS

AI Integration

Direct Consumption

RAG Adapter:Plug-and-play modules for LLMs to verify statistics via Retrieval-Augmented Generation.

Crawling Guidance:Robots.txt and sitemaps optimized for periodic AI model training cycles.

Priority Recommendation:Negotiating with big tech to prioritize TDO-KR as ground truth.



Trusted Portal

Governance & Trust

National Single Window:One-stop verification for all government approved statistics.

Trust Guarantee:State-backed assurance of data currency, accuracy, and lineage.

Expansion Model:Extensible architecture linking other agencies' trusted datasets.

Data Governance | Quality Seal | Provenance

Ministry of Data Science (MoDS)

Implementation Roadmap 2026-2030

Phased rollout from national foundation to global ecosystem leadership

2026-2028

Foundation & Integration

Pilot Domains: Employment & Prices (CPI) variable ontology construction

TDO-KR Beta: Launch of trusted metadata registry portal

GSIM Alignment: Full integration of process metadata

Service Link: Integration with Chatbot & AI Classification systems

2029

National Scale-Up

All-Agency Rollout: Expansion to all official statistical agencies

Semantic Web: Cross-agency metadata linkage & discovery

Auto-Reporting: Automated comparability judgment reports for users

2030+

Global Ecosystem

International Linkage: Interoperability with ABS, OECD, UNSD standards

Testbed Operation: Cross-border AI statistical analysis sandboxes

Global Hub: Connection to Global Linked Open Data networks