# Bridging the Semantic Gap between Official Statistics and AI-Mediated Use

Riitta Piela, Statistics Finland, riitta.piela@stat.fi

**Abstract.**

*The rapid adoption of large language models, AI-based search systems and AI agents is changing how users' access and interpret official statistics. Statistical information may increasingly be retrieved through conversational interfaces, AI-generated search answers, commercial AI platforms and agentic tools rather than directly from official dissemination portals. This creates new risks. AI systems may rely on secondary or outdated sources, or they may use official data while misunderstanding the statistical concepts, populations, classifications, time references and quality limitations behind the figures.*

*A key challenge also lies in the structure of statistical dissemination itself. Official statistics are often published as individual tables, each optimized for a specific publication purpose. Concepts, classifications, populations and time references may vary across statistics, even when they describe related phenomena. When accessed through AI systems or agents, this fragmentation can lead to plausible but misleading answers.*

*This paper argues that AI-ready dissemination requires more than better AI models, natural language query techniques or technical access to data. It proposes a semantic layer and a broader context layer for official statistics. The semantic layer connects user questions to official concepts, valid measures, classifications and operationalization. The context layer adds provenance, quality notes, usage constraints, clarification rules, governed access paths and cases where the system should not answer. Together, these layers help align user intent with statistical meaning before an answer is produced.*

***Keywords:*** *semantic layer, context layer, AI, AI-mediated dissemination, statistical validity*

## 1. Introduction: AI-mediated dissemination

Large language models, AI-based search systems, conversational interfaces and emerging agentic tools are changing how users access and interpret official statistics. Increasingly, statistical information may be retrieved through ChatGPT, Gemini, AI-generated search answers, commercial AI platforms or agentic interfaces rather than directly from the dissemination portals of statistical offices.

For users, this may make statistical information appear more immediate and easier to obtain. However, from the perspective of official statistics, this does not necessarily mean better access to official statistics. AI systems may retrieve statistical information from secondary sources, snippets, outdated summaries, media articles or other intermediaries instead of the statistical office's own website or dissemination services. Even when official sources are used, the answer may still become misleading if the system does not understand the statistical meaning, scope, constraints or methodological boundaries behind the data. The risk is therefore both source-related and interpretation-related.

In an AI-mediated environment, dissemination can no longer be understood only as the final stage of statistical production, where data and metadata are published in tables, databases, APIs, reports and web pages. Dissemination also includes preparing statistical information for discovery, interpretation, retrieval, reformulation and reuse by AI systems.

Three complementary technical access paths can be distinguished. First, public web content remains important because many commercial AI systems rely on web pages, search engine indexes, crawlable content and structured metadata. Second, APIs remain essential for structured and up-to-date access to statistical data. Third, agent-oriented integration layers, such as MCP-, skills- or function-based mechanisms, may provide ways for AI agents to interact with statistical data infrastructures.

However, these access paths are not separate solutions to AI-ready dissemination. They are routes through which AI systems may reach official statistics, but they do not by themselves ensure correct interpretation or statistically valid reuse. A visible web page can still be misunderstood, an API can return correct values without sufficient meaning, and an agent interface can execute technically valid but statistically wrong queries.

AI-ready dissemination is therefore a layered capability. The technical access paths must be supported by semantic and contextual layers that define concepts, valid measures, classifications, provenance, quality notes, usage constraints and clarification rules. Without these layers, none of the three access paths can reliably support AI-mediated use of official statistics.

## 2. The semantic and contextual gap

AI-ready dissemination requires more than syntactic interoperability. Syntactic interoperability means that a machine can read the data structure through APIs, machine-readable file formats, standard schemas or structured tables. Semantic interoperability means that the machine can understand what the data represents: the concepts, variables, measures and classifications. Contextual interoperability goes further. It makes explicit how the data may be used, what limitations apply, what it can and cannot be combined with, and when a question should not be answered without clarification.

This distinction is important because a table may be syntactically accessible through an API but still be semantically and contextually insufficient for AI-mediated use. The system may be able to read the table, but not know what the measure represents, which population it applies to, which classifications are valid, whether it is comparable with another table, or whether the user's question is underspecified. For official statistics, AI-readiness means moving from syntactic access to semantic and contextual interoperability.

The semantic and contextual gap is the difference between the way users express information needs in everyday language and the way official statistics define, structure, constrain and qualify statistical information. For example, a question such as "How has youth unemployment developed in recent years?" is statistically underspecified unless the relevant age group, unemployment concept, time span, data source and comparability over time are made explicit.

The central problem is that natural language allows flexibility and vagueness, while statistical interpretation requires precision. A fluent AI-generated answer may hide the fact that the system has silently selected one interpretation among several possible ones. The result may look useful but be statistically invalid. The gap is therefore not simply a linguistic problem; it is a statistical validity problem.

The gap is amplified by the structure of statistical dissemination itself. Official statistics are often published as individual tables, each optimized for a particular publication purpose. Across tables and publications, closely related phenomena may be described using different concepts, measures, classifications, populations, aggregation levels, time references, methodological assumptions and update cycles. Each table may be correct within its own context, but the relationships between tables are not always explicit enough for AI systems.

In a conventional database, ambiguity may concern which table or column should be selected. In official statistics, ambiguity also concerns whether two measures are conceptually equivalent, whether classifications can be mapped, whether time references are comparable and whether populations are defined in the same way. AI systems therefore need explicit information about whether and how tables, measures, classifications and quality constraints can be interpreted together.

## 3. The semantic and contextual layer

This paper proposes a governed semantic and contextual layer between natural language interfaces and statistical data structures. The semantic layer is the meaning component of this broader context layer. It connects natural language questions to statistical concepts, classifications, operationalizations, constraints and data access paths. It does not replace existing dissemination systems such as PXWeb, SDMX or JSON-based APIs, statistical databases or web pages. Instead, it sits between user intent and statistical data, making the meaning and constraints of the data explicit before an answer is produced.

The semantic layer can be described as a translation layer between natural language and the statistical data model. It answers three practical questions: how users talk about a phenomenon, which statistical concepts and data elements these expressions correspond to, and when the question is ambiguous enough to require clarification. The relationship between the semantic layer and the context layer can be stated simply: the semantic layer defines statistical meaning, while the context layer governs how that meaning is maintained, retrieved, applied and updated across AI-mediated use.

The context layer for official statistics should cover four functions: meaning and statistical structure, use and constraints, quality and trust, and technical and interaction support. Together these define concepts, vocabularies, classifications, populations, measures, units, time references, validation rules, comparability, quality notes, provenance, operational state, persistent identifiers, query parameters and machine-readable rules. This modular structure makes it possible to evaluate where interpretation failures occur.

A particularly important component is operationalization metadata. In official statistics, the same everyday concept can be implemented differently in different statistics. Without explicit operationalization metadata, an AI system may select a technically valid table while answering the wrong statistical question.

For aggregate statistical data, the context layer must make explicit the official measure used, the parameters under which it is valid, the combinations that are allowed or prohibited, and the situations in which the available information is insufficient to answer. The ability to clarify or not answer is part of trustworthy AI-mediated dissemination.

The context layer should therefore be treated as living infrastructure, not as a one-time metadata project. It brings together metadata from production and dissemination contexts and organizes it for AI-mediated use of aggregate statistical outputs. A practical implementation pathway may start from existing dissemination structures, and continue through metadata extraction, selected aggregate data modelling, semantic and contextual enrichment in a controlled data environment, curated views or controlled query functions, and AI services or agents that use the context layer rather than raw tables.

This also has implications for expert work and governance. The context layer should not be governed as a purely technical artefact. It requires expert input on concepts, operationalization, classifications, quality notes, comparability statements and clarification rules. At the same time, the role of experts should be focused on decisions that materially affect statistical interpretation and trust, rather than on manually reviewing every metadata field. In this sense, the governance of the context layer should combine domain expertise, methodological knowledge, dissemination expertise and technical implementation.

## 4. Controlled and agent-safe access

A key implication of the context layer is that AI systems should not receive unrestricted access to statistical tables. They should access data through curated views, controlled query functions or validated access paths. In this context, functions are controlled operations that translate an AI system's request into approved statistical actions. Instead of allowing the system to query raw tables freely, these functions help it resolve concepts, identify suitable indicators, retrieve values, compare time series, validate whether the question is statistically meaningful, and attach the relevant context to the answer.

The agent's workflow should therefore follow a governed path from context retrieval and concept resolution to validated query execution and sourced answer generation. The model should not independently decide what data to retrieve or how to interpret it. It should operate within a semantic and contextual framework that constrains and guides its actions.

A central function of the context layer is not only to help the system answer questions, but also to determine when a question should not be answered. For AI-mediated access to official statistics, asking for clarification or refusing to answer when the available information is insufficient may be a more trustworthy outcome than producing a fluent but statistically invalid answer.

## 5. Evaluation and quality

AI-ready dissemination requires systematic evaluation and monitoring. It is not enough to assess whether an answer sounds plausible or whether a numerical value is technically correct. Evaluation must cover whether the system selects authoritative sources, maps user language to the right concepts and operationalization, handles classifications and time references correctly, communicates quality and provenance, respects constraints, and knows when to clarify or not answer.

The quality of AI-mediated dissemination has two related dimensions. The first is the AI-readiness of the information provided by statistical offices: whether data, metadata, concepts, sources, licenses, access mechanisms and constraints are clear and structured enough for AI-mediated use. The second is the quality of the AI-generated answer: whether the system correctly identifies, interprets, cites and communicates official statistics.

Monitoring should therefore focus not only on technical performance, but also on recurring interpretation failures, missing metadata, outdated context, weak source attribution and cases where the system should have asked for clarification or refused to answer. In this sense, evaluation becomes part of statistical quality management and interpretative risk management.

## 6. Conclusion

AI-mediated dissemination introduces a new form of interpretative risk for official statistics. Users increasingly receive statistical information through systems that may not retrieve the official source, may rely on outdated or secondary material, or may not understand the conceptual, methodological and contextual boundaries of the data they retrieve.

The response cannot be limited to improving language models, prompts, search techniques or technical access paths. Official statistics need a semantic and contextual foundation that makes statistical meaning, source authority, quality, provenance, usage constraints, comparability and cases where not to answer explicit and machine actionable. Publication-specific table design remains possible, but only if the relationships between concepts, measures, classifications, populations, time references and quality constraints are clear enough for AI-mediated use.

## Appendix: List of references


Atlan. (2026, April 6). *The Core Components of a Context Layer for AI Agents Explained*. Atlan. https://atlan.com/know/core-components-context-layer/

Atlan. (2026). *Context Layer for AI Agents: Enterprise Guide 2026*. Atlan. https://atlan.com/know/context-layer-for-ai-agents/

Bandi, P. K. R. (2025). The role of metadata in making data AI-ready: Enhancing data discoverability and usability. *Journal of Computer Science and Technology Studies, 7*(5), 954–963. https://al-kindipublishers.org/index.php/jcsts/article/view/9919

Cube. (2026, April 28). *Why semantic layers make LLM analytics reliable: A paired benchmark across three frontier models*. Cube Blog. https://cube.dev/blog/why-semantic-layers-make-llm-analytics-reliable-a-paired-benchmark-across-three-frontier-models

Gartner. (2026, May 11). *Gartner Says Lack of Semantics Causes Inaccurate AI Agents and Wasted Spending*. Gartner Newsroom. https://www.gartner.com/en/newsroom/press-releases/2026-05-11-gartner-says-lack-of-semantics-causes-inaccurate-artificial-intelligence-agents-and-wasted-spending

Rumiantsau, M., & Fokeev, I. (2026). *Semantic Layers for Reliable LLM-Powered Data Analytics: A Paired Benchmark of Accuracy and Hallucination Across Three Frontier Models*. arXiv. https://arxiv.org/pdf/2604.25149

Winks, E. (2026, June 15). *What Is a Semantic Layer for AI Agents? A Complete 2026 Guide*. Atlan. https://atlan.com/know/ai-agent/semantic-layer-for-ai-agents/