Irving Fisher Committee on
Central Bank Statistics

BIS

12th biennial IFC Conference: "Statistics and beyond: new data for decision making in central banks"

22-23 August 2024

Enhancing metadata with generative AI: the case of BIS statistics[1]

Bilyana Bogdanova, Magdalena Erdem, Bianca Ligani and Olivier Sirello,
Bank for International Settlements

[1] This contribution was prepared for the conference. 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.

Enhancing metadata with generative AI: the case of BIS statistics

Bilyana Bogdanova, Magdalena Erdem, Bianca Ligani, Olivier Sirello[1]

Abstract

The rise of generative artificial intelligence (AI) carries significant potential for metadata in official statistics. This paper sheds light on the role of generative AI as both user and producer of metadata. We describe the symbiotic link that exists between the two roles as a circle where generative AI employs metadata to, in turn, generate them in its responses. We illustrate this circle in action by documenting two use cases from the Bank for International Settlements. Finally, we contextualise this circle within the evolving information paradigm and call for official statisticians to rethink their role in the production, usage and dissemination of metadata in the age of AI.

Keywords: official statistics, metadata, artificial intelligence, process efficiency

JEL classification: C80, C88, C63, O33

Contents

Introduction

The rise of generative artificial intelligence (AI) and large language models (LLMs) carries significant potential for official statistics. This frontier technology can perform a variety of tasks across the entire statistical business process, spanning from collection, processing, analysis and dissemination. In practice, the race to use generative AI in official statistics has already started, with a burgeoning number of use cases across statistical organisations. Typical applications include statbots, coding assistants, report generation and searches based on natural language. Although this technology is not without risks, its benefits are increased automation of routine processes with substantial improvements in the accuracy, timeliness and quality of statistics.

Yet one area where generative AI is particularly relevant and cost-effective is metadata generation and editing. There are several reasons for this, on a conceptual and practical level. On the one hand, it is well known that LLMs can reach human-like editing and text generation proficiency due to the transformers' capability to grasp relationships between words. On the other hand, processing metadata is a daunting task for compilers, as it is usually time consuming and requires manual review. Finally, AI not only *produces* metadata but, most importantly, also *uses* them. In this context, AI *for* and *from* metadata can significantly change the landscape of official statistics.

The goal of this paper is threefold. First, it sheds light on the circular relationship between metadata and generative AI and discusses the key implications for official statistics, particularly in terms of data interpretability, discoverability and dissemination. Second, from theory it dives into practice and explores two use cases of generative AI as producer and user of metadata, taking as examples the recent work on Bank for International Settlements (BIS) statistics. Third, acknowledging the risks and limitations, it argues for a new active role for official statisticians in the age of AI, particularly to feed a virtuous circle between generative AI and metadata.

[1] Respectively, Head of Macroeconomic Statistical Support, Monetary and Economic Department , Bank for International Settlements, and co-chair of the AI and Communication Task Team, UNECE High-Level Group for the Modernisation of Official Statistics (bilyana.bogdanova@bis.org); Head of Departmental Research Support, Monetary and Economic Department, Bank for International Settlements (magdalena.erdem@bis.org); Senior Macroeconomic Analyst, Monetary and Economic Department, Bank for International Settlements (bianca.ligani@bis.org); Senior Statistical Analyst, Monetary and Economic Department, Bank for International Settlements, and co-chair of the Generative AI for Official Statistics Project, UNECE High-Level Group for the Modernisation of Official Statistics (olivier.sirello@bis.org).

The authors thank Kevin Tracol and Krzysztof Zdanowicz for helpful comments and suggestions. We also thank Frank Herrmannsdoerfer and Yannis Seemann for their excellent support for the BIS Metadata AI Editor. The views expressed here are those of the authors and do not necessarily reflect those of the Bank for International Settlements.

1. Generative AI as a user and producer of metadata

From data to information: the pivotal role of metadata in official statistics

In official statistics, metadata play a pivotal role in interpreting, analysing and sharing data. Often defined as structural or descriptive information about the primary data, metadata typically allow users "to get the facts right" (UNECE (2013)).[2] From a functional perspective, they play at least three key roles. First, they enable users to understand, identify and interpret the data by providing the semantic context. For example, a given figure by itself does not have any meaning, unless it is augmented with some additional information or *structural* metadata (Dippo and Sundgren (2000)). From this point of view, metadata give shape to the meaning of data. A second important role is to describe the statistical process. *Reference* metadata help in understanding, for example, the collection, compilation, processing and dissemination phases across the statistical production chain. This information is particularly relevant to evaluate the data, for instance in terms of accuracy, completeness, reliability and, ultimately, trustworthiness. Finally, data without metadata can hardly, if at all, be exchanged and shared. As enablers for interpretability, they play a vital role in ensuring that data are exchanged and communicated in a correct, consistent and accurate way.

More fundamentally, metadata transform data into information. With the growing availability of granular and alternative data, metadata have become more essential than ever before (IFC (2024b)). On the one hand, compiling and using microdata often require sound understanding of the information model to perform correct business aggregations. In fact, with microdata, "the devil is often in the details" necessitating a "data-centric metadata" approach to provide information not only on the data themselves but also, more importantly, on the related workflows to access and analyse them (Israël and Tissot (2021)). On the other hand, harnessing the wealth of alternative sources, including new data types, also requires rich, multidimensional and contextual information, particularly to integrate and link data across multiple sources and domains while also ensuring harmonisation and consistency (Laaboudi et al (2023); UNECE (2024d)). More broadly, providing comprehensive and transparent metadata pertains to the reliability and trustworthiness of official statistics (Krizman and Tissot (2021)). Not surprisingly, the Fundamental Principles of Official Statistics emphasize the objective of facilitating a correct interpretation of the data,[3] while the

[2] According to the Generic Law on Official Statistics (article 4), metadata are "means data and other documentation that describe statistical data and statistical processes in a standardised way by providing information on data sources, methods, definitions, classifications and data quality". Another definition refers to metadata as an "excerpt of meta-information", with "meta-information" defined as data descriptions (see Froeschl (1997) for comprehensive definitions of metadata).

[3] More specifically, principles 3 and 4 of the Fundamental Principles of Official Statistics read, respectively: "To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics" and "The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics." See UN, A/RES/68/261.

Principles Governing International Statistical Activities advocate for transparency in concepts, definitions and classifications as well as in sources and methods.[4]

In this context, achieving proper management and stewardship of metadata is critical. Although ensuring effective metadata management is not a new issue (Froeschl (1999)), the deluge of data, notably propelled by technology over the past few years, raises new fundamental questions. First, one could argue that the ever-increasing volume of information, coupled with widespread use of alternative sources, requires the systematic collection and curation of metadata across the data lifecycle (Gregory (2024)). This demands adequate production systems and software able to generate fine-grained metadata programmatically. These metadata should also be machine-readable and actionable, particularly to integrate with metadata-driven processes and streamline data workflows through automated decision-making (Batista et al (2022)). A second, related consideration is the fundamental role of standards to properly manage and govern metadata. In fact, standards are not only key to ensuring consistency, integrity and accurate classification of concepts. They also play a critical role in achieving proper governance and management of statistical concepts, for example by ensuring that metadata authoring and maintenance follow predefined sets of rules.[5] In this regard, standards are a key condition to foster the findability, accessibility, interoperability and reusability of data (FAIR) (Wilkinson et al (2016), UNECE (2024c)). Finally, perhaps more importantly, metadata are becoming increasingly *active*, rather than *passive*, as technology facilitates two-way interactions between producers and users. In practice, this means that metadata are not solely created by producers, for example to document a certain asset in a data catalogue, but users also contribute to their generation. Typical examples are machine learning and, more broadly, AI solutions which leverage users' input. In this context, establishing data governance frameworks is a critical condition for leveraging the full potential of these frontier technologies (BIS (2024); UNECE (2024)).

The circle between generative AI and metadata: implications and benefits

The rise and democratisation of generative AI have opened up new avenues for official statistics. If AI is not a new phenomenon, recent advances in computational power as well as techniques such as tree-based methods, neural networks, transformers and natural language processing have enabled the unprecedented expansion of generative AI over the past few years. This technology can typically generate a broad range of content such as visuals, audio and texts, for instance through LLMs which are specifically trained to predict sequences of words from large sets of data. Driven by the recent roll out of chatbots using generative AI,[6] an

[4] The Committee for the Coordination of Statistical Initiatives states in its Principles Governing International Statistical Activities that "Concepts, definitions, classifications, sources, methods and procedures employed in the production of international statistics are chosen to meet professional scientific standards and are made transparent for the users." It also stresses the good practices of "Documenting and publishing concepts, definitions, classifications and metadata used by the organization [… and] how data are collected, processed and disseminated by the organization".

[5] For example, the Dublin Core Metadata Initiative (DCMI) is responsible for governing and curating the Dublin Core vocabulary, one of the major metadata vocabularies used by the World Wide Web. In the field of official statistics, standards such as the Statistical Data and Metadata eXchange (SDMX) also have specific terms for governing metadata (see SDMX (2023)).

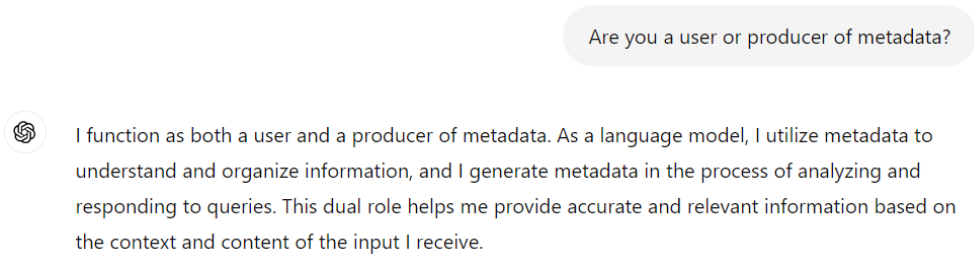
[6] A noteworthy example is the launch in November 2022 of ChatGPT, which reached over one million users in less than a week, making it the fastest-growing application in history.

unavoidable race has started across domains and organisations, including central banks, on how best to leverage this frontier technology to modernise and streamline work (Araujo et al (2024)). Official statistics have not been spared by this wave of innovation: solutions and ideas have quickly emerged that apply generative AI across the entire statistical production chain, from data collection and information extraction and processing to dissemination and communication (UNECE (2023); Fraisl (2024)).

More specifically, the use of generative AI is particularly cost-effective in generating and editing metadata, for at least three main reasons. First, by their very nature, LLMs can achieve human-like generation and editing of text. This can be typically explained by the capability of transformers to capture relationships between words (Araujo et al (2023)). A second reason is that processing metadata is particularly challenging for compilers. In fact, metadata editing is usually highly time-consuming, often requiring statisticians to manually review content. This process involves multiple checks, from basic rules such as capitalisation, formatting and layout, to more advanced tasks like verifying grammar, spelling and syntax, or ensuring text fluency and consistency. To give an order of magnitude, depending on the complexity of the text, statisticians may spend several minutes validating only a single attribute, such as a short text documenting the compilation of a time series. Finally, generative AI has proven to be particularly beneficial across several domains, such as data discoverability (Dupriez et al (2024)), documentation and, more importantly, communication, for example to transform technical jargon into plain language. It may therefore help to adapt technicalities contained in metadata to the literacy level of users (IFC (2023; 2024a)).

Beyond cost-effectiveness, curating and disseminating high-quality metadata are more crucial than ever, as generative AI is both a *producer* and *user* of metadata. The relationship between generative AI and data is in fact circular and symbiotic. It is circular since, as producer, generative AI *creates* content or information, leveraging LLMs that are pre-trained on extensive data sets. The responses from these algorithms are then *used* by the technology and continuously fed back into the system, creating a cycle. It is symbiotic since generative AI cannot work properly without metadata. Metadata play indeed a pivotal role as they *inform* the models about how to interpret and, more importantly, *understand* the data (Dupriez et al (2024)). And, indeed, when asked "Are you a user or a producer of metadata?", ChatGPT answers: "I function as both a user and a producer of metadata. As a language model, I utilise metadata to *understand* and organize information, and I generate metadata in the process of analysing and responding to queries" (Figure 1).

ChatGPT as a user and producer of metadata — Figure 1



Source: ChatGPT 3.5 (as of June 2024).

In practice, this circular relationship carries important implications for official statistics. First, one could say that statisticians are the ones initiating the loop by providing the initial data and metadata, thereby setting the cycle in motion. Disseminating high-quality statistics with accurate metadata becomes more pressing than ever before: the reason is that any wrong input would propagate across the system in a never-ending loop. Pushing to the utmost extreme, statisticians leveraging AI could generate low-quality metadata based on initial inaccurate input. A second, perhaps even more fundamental, implication is that generative AI should be able to properly understand the data disseminated by statistical offices: at stake lie issues such as integrity, accessibility, misuse of data and, ultimately, misinformation. In fact, while humans rely on shared and general knowledge to understand data, LLMs do not. To make matters worse, LLMs may suffer from hallucination problems and currently exhibit significant cognitive limitations (Perez-Cruz and Shin (2024)). Metadata can help AI to better understand the data: in this way, they play a crucial role in making data AI-ready while ensuring that no semantic meaning is lost (Noe (2024)).

2. From theory to practice: the case of BIS statistics

Given the pivotal role of metadata in the new information ecosystem, the BIS has been innovating its statistical workflows to leverage AI, as both a producer and user of metadata.

Leveraging AI to produce metadata: the BIS Metadata AI Editor

The BIS disseminates data and metadata on key financial and macroeconomic indicators through the BIS Data Portal. As part of its mandate to support global monetary and financial stability, the BIS compiles a broad range of statistics, spanning from credit to the non-financial sector statistics, as well as on property prices, exchange rates and international banking, in close cooperation with central banks and other national authorities. These time series are released through the BIS Data Portal,[7] where users can access a variety of tools and curated statistical content for guided exploration. Undoubtedly, metadata play an important role here. First, they account for key aspects of the compilation, such as methodology, data collection, coverage and sources. Second, metadata are fully integrated across various features of the BIS Data Portal by leveraging the SDMX standard (Lambe and Park (2024)). This standardised approach is particularly helpful to prevent content duplication and enables users to seamlessly navigate complex information, including through dashboards and tables.

In this context, leveraging AI-powered assistants can be helpful to balance the need for high-quality metadata with their resource-intensive curation.[8] First, assistants are relatively easy to configure and customise. For example, the assistant

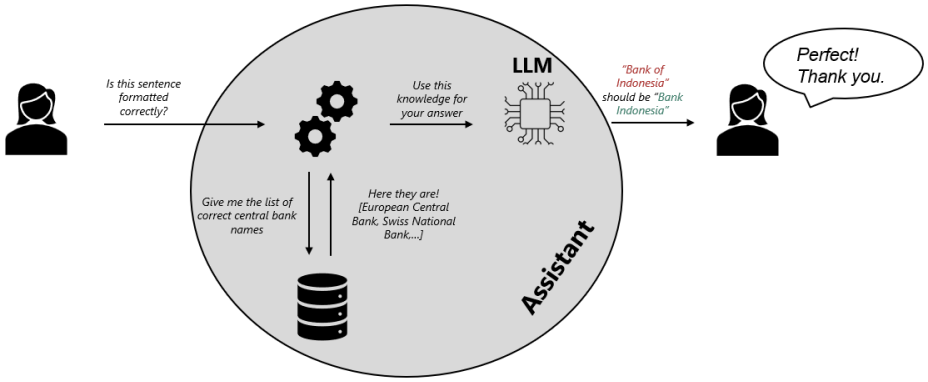
[7] The BIS Data Portal is the entry point to BIS statistics and can be accessed at data.bis.org.

[8] We refer here to OpenAI Assistants API. As specified in the OpenAI documentation, an assistant follows specific sets of instructions, leverages custom tools and can access files to be used in its output. See more at OpenAI Assistants.

can be preconfigured directly from the OpenAI platform. Here, statisticians can specify the model and system instructions as well as add more custom functions. Second, statisticians can upload files to augment the knowledge base of assistants. In fact, these are OpenAI's version of retrieval-augmented generation (RAG). This feature is important for maximising the accuracy of the output. For example, assistants can adhere to specific guidelines, for example to align with BIS editing rules (Figure 2). Integrating custom files, such as those containing the exact names of BIS member central banks, facilitates the verification process and ensures metadata consistency. Finally, assistants can be easily integrated within an application through API (application programming interface) calls.

A basic editor assistant workflow

Figure 2

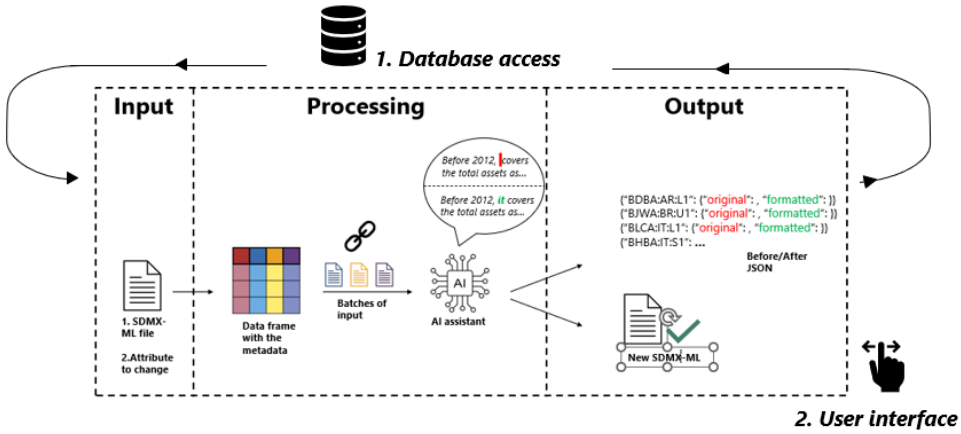


Source: authors.

Built upon assistants, the BIS Metadata AI Editor is an application to process metadata developed by statisticians for statisticians. The application has three major components. First, it leverages the SDMX standard end to end: at the very start, users can upload an SDMX file containing the metadata, for example specified in a given attribute at the time series level. Once the processing has been completed, the application generates an SDMX file with the results of the metadata processing. Integration with SDMX is key to streamlining the workflow as well as to ensuring full compatibility with existing statistical pipelines (SDMX (2024)), which typically rely on SDMX to ingest and validate data (Figure 3). Second, statisticians may provide the application with their own assistant through a unique ID and generate and edit the metadata attributes based on a set of custom instructions. Third, to facilitate the workflow, the application comes with both a command-line interface (CLI) and a web interface, based on Streamlit. The user interface has been designed to be very straightforward and easy to maintain. Here users are presented with two pages: the first contains the input parameters such as uploaders, assistant ID, a metadata field to edit and other advanced parameters to finetune the process. The second page presents the responses returned by the assistant and grouped by time series. The interface also offers users the chance to track the changes made by the assistant. In this context, statisticians are given the choice to accept, revert or freely edit the generated text. This is a critical feature, not least to ensure human-in-the-loop, for example to validate the results.

More broadly, the BIS Metadata AI Editor follows three principles: agile, modular and opportunity-driven. Given the fast pace of AI developments, it is key to avoid "reinventing the wheel". Rather, what is needed is assembling existing solutions under a common framework. In this regard, the BIS Metadata AI Editor relies on off-the-shelf solutions whenever possible. Another significant benefit of this modular approach is the combination of low initial costs with high customisation. Finally, the solution has been opportunity- rather than issue-driven. In fact, the project has created opportunities beyond the initial scope, including in terms of capacity development.

The BIS Metadata AI Editor: end-to-end pipeline Figure 3



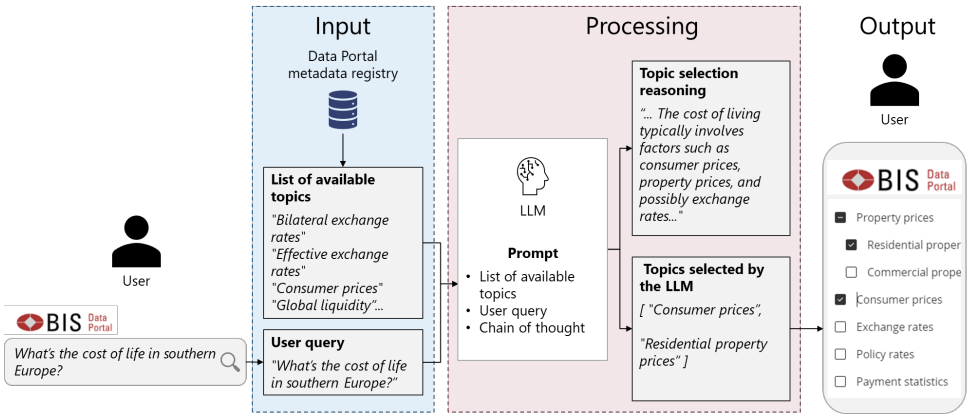
Source: authors.

Leveraging AI to better use metadata: improving the findability and accessibility of data

As a user of metadata, generative AI can significantly improve data findability and accessibility. One of the key strengths of natural language processing (NLP) is precisely its ability to link general content with respective domain-specific content. For example, a question asked in simple words by a non-expert user can be matched with more technical jargon. In this regard, AI as a user of metadata can boost the findability of the data. In fact, in contrast to traditional searches based on keywords, an AI-powered semantic search interprets the metadata: in short, it allows for *searching with meaning* (Dupriez and Solatorio (2024)). In practice, this entails many advantages, not least search efficiency and simplification of queries. More importantly, searching with meaning yields a powerful advantage in terms of accuracy: for example, searching for "security" in the *context* of financial statistics may return data on securities – intended as financial instruments – while this same keyword could show results on physical security threats in the context of statistics on criminality. Enhanced data findability fosters, in turn, greater data discoverability and, more importantly, accessibility. Since semantic searches are typically able to return relevant results even when provided with imprecise queries, they help to broaden the data reach, particularly for non-expert users who may not employ the correct keywords in their queries.

Against this setting, NLP-based techniques are being tested to enhance the search and glossary functionalities of the BIS Data Portal. Traditionally, the BIS Data Portal has helped users filter data within predefined categories, also called data sets. Yet, navigating through those filters may require users to have some domain-specific knowledge, for example to choose between *bilateral* or *effective* exchange rates. To address this challenge, AI, and in particular language models (LM), can assist. For instance, a user's simple data query is passed to the LM, and the LM leverages the descriptive metadata attached to each category to identify the most relevant data set based on the user prompt (Figure 4).

The BIS Data Portal query assistant: browsing data sets Figure 4

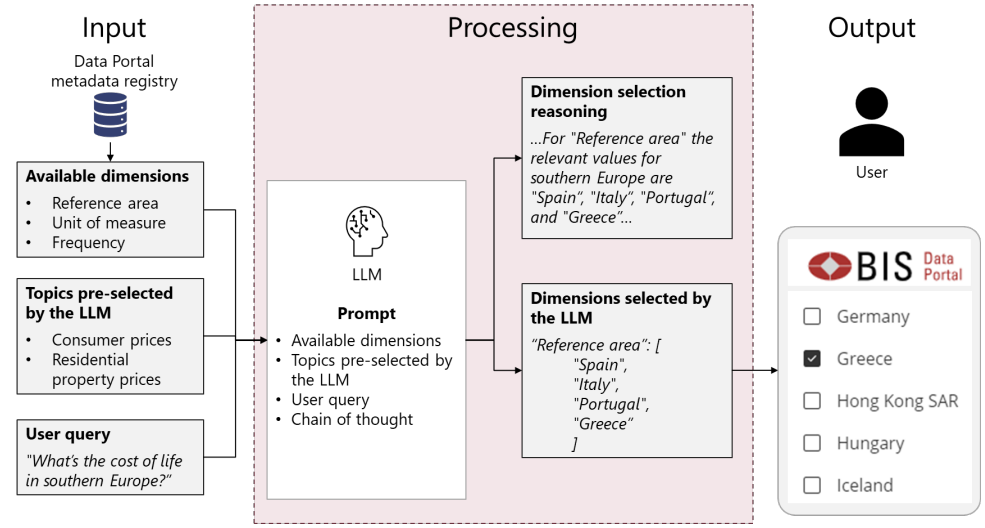


Source: authors.

To optimize results, the proof of concept employs an LLM and chain-of-thought prompting which enhances the LLM's ability to perform complex reasoning by generating intermediate reasoning steps (Wei et al (2022)). In practice, once selected, the data set represents a smaller subset from which the model can filter the relevant data further. For instance, if the selected data sets cover multiple geographical locations, chain-of-thought prompting can be useful to identify the data elements associated with the user's query (Figure 5). Although this approach is not without challenges, for example due to incomplete, ambiguous or excessively technical jargon, it also clearly shows that enriched metadata not only benefit users but also support LMs in answering questions from users.

The BIS Data Portal query assistant: refining the results

Figure 5



Source: authors.

3. Beyond challenges: towards an active role of official statistics for a virtuous circle between AI and metadata

Leveraging generative AI for producing and using metadata poses a certain number of limitations and risks. Being aware of these challenges is a necessary but not sufficient condition to foster a virtuous circle between AI and metadata.

Facing risks and ethical implications

Although generative AI presents many benefits, it also features several limitations that cannot be ignored. The first is the lack of transparency and explainability, often referred to as the "black box" problem. This issue arises because the algorithms and models used in generative AI can be highly complex and opaque, making it difficult for users to understand how decisions are made or results derived (UNECE (2024d)). And, in fact, when asked if metadata are used in the training, ChatGPT answers: "As an AI language model, I do not have direct access to my training data or know exactly how it was curated."[9] Recent and ongoing work by the High-Level Group for the Modernisation of Official Statistics has highlighted these concerns in the field of official statistics, emphasising the importance of transparency for ensuring the reliability and trustworthiness of AI in official statistical processes (UNECE (2023)).

More specifically, **hallucinations and repetitions may be problematic for producing and using metadata with AI**. Hallucinations occur whenever an LLM generates incorrect content by perceiving wrong patterns, for instance due to insufficient training data or biases. This risk is high for metadata: it could generate incorrect outputs which can, in turn, feed back into the system as inputs that may be

[9] The prompt is "Do you use metadata in your training and how do you use them?", provided to ChatGPT (GPT-4o), free version, as of June 2024.

used for statistical purposes. This never-ending cycle would propagate and contaminate future AI trained on inaccurate metadata. Ultimately, hallucinations in AI have the potential to undermine knowledge (Garrett (2024)). The risk of repetition might also be very problematic as AI-generated content typically reproduces the relationships found in its original training set.

A third, related strand of limitations revolves around ethical and reputational considerations. On the one hand, language dependency is known to be a significant issue as AI systems trained on data from one linguistic or cultural context may not perform well in another. For instance, language divides are critical: one in two training data sets are in English, creating challenges for any application where English is not the main language (OECD (2024)). Moreover, generative AI as a user and producer of metadata can also lead to greater data misuse if not properly governed. Typical cases involve using metadata unethically or inappropriately, including to manipulate the training processes of AI systems, for example through model poisoning and prompt attacks. Taken together, these risks give rise to reputational concerns, which stand out as a significant challenge for leveraging generative AI in official statistics.

Finally, copyright, confidentiality and governance are crucial aspects that must be addressed when leveraging generative AI in official statistics. First, one fundamental point is whether content produced from LLMs is identical to proprietary inputs, such as articles, published elsewhere. Second, metadata are typically not given the same level of protection and governance as the data themselves. This can lead to breaches of confidentiality and privacy. The use of metadata to train AI systems also poses questions about unauthorized data access and the best security protocols to mitigate such risks. Finally, in terms of quality, metadata are typically less subject to data quality management processes traditionally applied to data. Yet, given their critical role in AI applications, the lack of strict quality checks on metadata may raise significant risks, not least in terms of misinformation. This calls for urgently tailoring data governance and management frameworks to AI requirements for metadata, with clear identification of responsibilities, roles and ownership to secure the highest quality and compliance with existing global statistical standards, including the Fundamental Principles of Official Statistics.

Feeding the virtuous circle between generative AI and metadata: calling for an active role for official statistics in the age of AI

As we move into the era of AI from the age of the internet, the paradigm of information retrieval is radically changing. With the advent of the internet, finding ways to index and link information became critical, for instance by using search engines, so that users can quickly find their way into the information they are look for. It should not come as a surprise that most of the discussion on metadata-based information management frameworks gained relevance precisely in this period (Froeschl (1999)). In the age of AI, what will matter is also the modelling of data, on top of their indexing. In short, machines do not just return search results, they answer questions. In fact, a chatbot answers the user directly without requiring any content browsing. Transitioning from data indexing to data modelling is anything but a simple move. It fundamentally shifts the focus from *where* the data are to *how* they are. In practice, the data are not necessarily trustworthy because they have been published on an official data portal. Rather, they are trusted because of extensive metadata

detailing *how* they are compiled and sourced. In other words, traceability of the inputs to the outputs becomes key.

This fundamental shift sheds light on the central role of official statisticians in the production and dissemination of high-quality data and reference information. In fact, there may be at least three implications. First, at a practical level, compilers may need to produce more, detailed and better metadata that describe the meaning, context and, more importantly, the origin of the data. Harmonised information or attributes such as the source, owner and lineage are particularly relevant to correctly trace, track and interpret the data. Fortunately, official statisticians are well equipped to address these challenges, starting with leveraging their expertise in using statistical standards, such as SDMX. Second, this evolution in the information landscape is a good reminder of the importance of literacy, particularly in statistics, data and, obviously, AI. Although the general trend in recent years has been to simplify statistical information to reach broader audiences, the rise of AI may centrifugally propel the trend in the opposite direction: it is precisely because AI, by its very nature, averages or levels information down that official statisticians have a role in levelling up users' knowledge. Metadata can foster literacy: rich and accurate metadata tailored to users' needs are key to avoid compromising data quality, integrity and transparency. This point brings up a third and perhaps more strategic consideration: dissemination of official statistics should adapt to the age of AI. Because AI-powered agents, such as chatbots or copilots, are likely to be the main entry door for data (Allen (2016)), platforms such as data portals are likely to be diminished in the same way that statistical bulletins have been gradually replaced by statistical portals or, more broadly, newspapers by social media. In practice, dissemination portals by statistical authorities may serve as endpoints for AI systems to tap into high-quality and authentic data. Key developments, such as model context protocols (MCP), can be part of the solution looking forward. Relatedly, a critical aspect of official portals will be to ensure that data can be properly reused and, more importantly, can be understood by machines, with metadata playing a key role. This requires an important change in the dissemination culture, as statistics should be not only machine-*readable* but also machine-*understandable* (US Commerce Department (2024)).

All in all, in the quest to produce more metadata, statisticians will ensure a virtuous circle with their usage by AI. In fact, given the circular relationship between generative AI and metadata, statisticians will have an important role in maintaining a virtuous circle. Should statisticians fail in their roles of creators, guardians and custodians of reliable metadata, AI will be trained with inaccurate or, worse, incorrect data. These will, in turn, be used by statisticians themselves, unavoidably precipitating a descent into deepfakes, misinformation and disinformation. The very essence of official statistics as a public good central to the transparency and accountability of democratic societies would be severely compromised. Making sure that the circle is virtuous, rather than vicious, will ultimately require two critical considerations. First, quality must be taken even more seriously. It might not be enough to describe data quality through accurate metadata. Rather, it is indispensable to prove that quality by quantifying the data's relative precision and traceability. Adding a standardised, quantitative layer in this regard will provide robust evidence and strong incentives for users, both humans and machines, to use data coming from authoritative sources. A second, perhaps more relevant, consideration is that in a world marked by a sustained deluge of data and models, machines are undoubtedly outpacing humans as direct users of official statistics. This means that it will not be enough for statisticians to

communicate and engage with only humans. Rather, statisticians will need to take on a new role as trainers of AI models and machines: metadata standards, open linked data and knowledge graphs can already be very helpful tools in this regard. Bridging the statistical literacy gap for machines may well become the top priority in the years ahead if we wish to preserve trust in official statistics in the age of AI.

References

Allen, B (2016): "The role of metadata in the second machine age", *International Conference on Dublin Core and Metadata Applications*, DC-2016, Copenhagen.

Araujo, D, S Doerr, L Gambacorta and B Tissot (2024): "Artificial intelligence in central banking", *BIS Bulletin*, no 84, January.

Araujo, D, S Probst, R Schimdt, B Vitez and M Zoss (2023): "Box 2: Harnessing recent breakthroughs in large language models (LLMs)", in D Araujo, G Bruno, J Marcucci, R Schmidt and B Tissot, "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October, pp 17–20.

Bank for International Settlements (BIS) (2024): "III. Artificial intelligence and the economy: implications for central banks", *Annual Economic Report 2024*, June, pp 91–127.

Batista, D, A Gonzalez-Beltran, S-A Sansone and P Rocca-Serra (2022): "Machine actionable metadata models", *Scientific Data*, vol 9, no 1.

Dippo C and B Sundgren (2000): "The Role of Metadata in Statistics", US Bureau of Labor Statistics, *Statistical Survey Paper*.

Dupriez, O, H Fu, C Hammer and A Solatorio (2024): "The transformative role of AI for development data", *World Bank Blogs*, April.

Dupriez, O and A Solatorio (2024): "Beyond keywords: AI-driven approaches to improve data discoverability", *World Bank Blogs*, May.

Fraisl, D (2024): "The potential of artificial intelligence for the SDGs and official statistics", *PARIS21 Working Paper*, April.

Froeschl, K (1997): *Metadata management in statistical information processing*, Springer.

——— (1999): "Metadata management in official statistics – an IT-based methodology approach", *Austrian Journal of Statistics*, vol 28, no 2, pp 49–79.

Garrett, A (2024): "The devil, the detail, and the data", *Journal of the Royal Statistical Society*, Series A: Statistics in Society, no qnae063, June.

Irving Fisher Committee on Central Bank Statistics (IFC) (2023): "How central banks communicate on official statistics", *IFC Report,* no 15, February.

——— (2024a): "Communication on central bank statistics", *IFC Bulletin*, no 60, April.

——— (2024b): "Granular data: new horizons and challenges", *IFC Bulletin*, no 61, July.

Israël, J-M and B Tissot (2021): "Incorporating micro data into macro policy decision-making", *Journal of Digital Banking*, vol 6, no 3, pp 233–50.

Krizman, I and Tissot B (2021): "Data governance: issues for the National Statistical System", *IFC Bulletin*, no 54, July.

Laaboudi, C, M Karlberg and M Islam (2024): "Open data dissemination at Eurostat: state of the art", *Statistical Journal of the IAOS*, vol 40, no 1, pp 81–90.

Lambe, E and T Park (2024): "The BIS Data Portal project – delivering the next generation platform for BIS statistics", *IFC Bulletin*, no 60, April.

Noe, L (2024): "Are official statistics machine-understandable enough? Analyzing ODIN to understand data readiness for AI", *Open Data Watch Blog*, April.

Organisation for Economic Co-operation and Development (OECD) (2024): "Embracing the technology frontier", *OECD Digital Economy Outlook*, vol 1, May.

Perez-Cruz F and H S Shin (2024): "Testing the cognitive limits of large language models", *BIS Bulletin*, no 83.

Statistical Data and Metadata Exchange (SDMX) (2023), *A reference framework for SDMX structural metadata governance*, March.

——— (2024): *SDMX+AI: unlocking the potential of NLP to enhance data access*, March.

United Nations Economic Commission for Europe (UNECE) (2013): *Getting the facts right – a guide to presenting metadata*, ECE/CES/29.

——— (2023): *Large language models for official statistics*, High-Level Group for the Modernisation of Official Statistics White Paper, December.

——— (2024a): *Data stewardship and the role of national statistical offices in the new data ecosystem*, ECE/CES/STAT/2023/4, April.

——— (2024b): *Organisational aspects of implementing ML based data editing in statistical production*, High-Level Group for the Modernisation of Official Statistics, February.

——— (2024c): *Data governance framework for statistical interoperability*, High-Level Group for the Modernisation of Official Statistics, March.

——— (2024d): *In-depth review of linking data across domains and sources*, Conference of European Statisticians, 72nd plenary session, ECE/CES/2024/5, June.

US Department of Commerce (2024): "AI and open government data assets request for information", *Federal Register*, vol 89, no 75, pp 27411–13.

Wei, J, M Bosma, E Chi, B Ichter, Q Le, D Schuurmans, X Wang, F Xia and D Zhou (2022): "Chain-of-thought prompting elicits reasoning in large language models", *Advances in neural information processing systems*, no 35.

Wilkinson, M et al (2016): "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, vol 3, no 160018.

Irving Fisher Committee on Central Bank Statistics

BIS

Enhancing metadata with generative AI:
the case of the BIS statistics

Bilyana Bogdanova, Magdalena Erdem, Bianca Ligani and **Olivier Sirello**

XII Biennial of the Irving Fisher Committee on Central Bank Statistics

Basel, 22 August 2024

Metadata in the age of AI: an introduction

- Metadata are critical yet costly
  - They are **critical**: they give shape and meaning to data
  - There are **costly**: time consuming, hard to maintain and standardise
- Yet, the data deluge and the rise of AI mark a new epoch:
  - Data are **everywhere**
  - Data **more and more generated, used and interpreted by machines**, not just humans
  - Data are **undifferentiated** and meaningless if not transformed into **information**
- What is the **role** and **place of metadata in the age of AI**?
  - Could generative AI help producers and users to achieve better metadata?
  - What is our role as official statisticians in central banks?

Outline

I. Generative AI as a user and producer of metadata

II. From theory to practice: the case of BIS statistics

III. Towards an active role of official statistics for a virtuous circle between AI and metadata

I. Generative AI as a user and producer of metadata

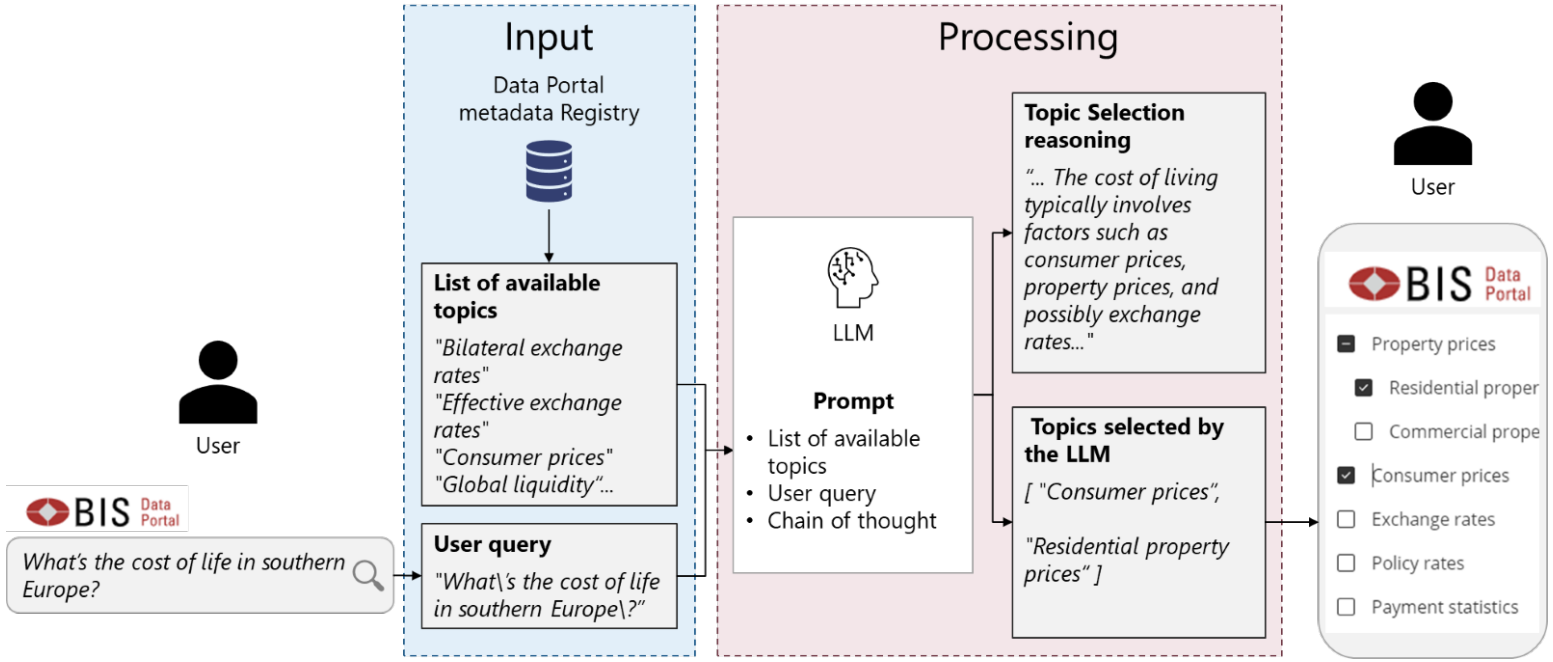
- The relationship between generative AI and (meta)data is circular and symbiotic
  - It is **circular**: generative AI *produces* content which is then *used* by the technology in a continuous loop or cycle
  - It is **symbiotic**: generative AI cannot work without metadata: they inform models to interpret the data
- **Two** major **implications**:
  - **Statisticians initiate the loop**: as providers of ground truth, any inaccuracy would propagate in a never-ending loop, with consequences for data misuse and misinformation
  - In the age of AI, machines should be trained e to *properly* understand the data disseminated by statistical offices: **from machine-*readable* to machine-*understandable* data**

II. From theory to practice: the case of BIS statistics

- **Producing metadata** with generative AI: the BIS Metadata AI Editor
  - Check metadata of time series, for example compilation, methodology description to ensure proper enforcement of editing rules
  - Leveraging OpenAI assistants: custom instructions and tailored knowledge base
  - Leveraging SDMX end to end: from ingestion to dissemination
  - Simple, modular and agile
    - By statisticians for statisticians
    - Framework rather than solution approach

II. From theory to practice: the case of BIS statistics

- **Using metadata** with generative AI: improving the BIS Data Portal search
  - NLP-based techniques to improve the search: search with meaning, not keywords



III. Towards an active role of official statistics: for a virtuous circle between AI and metadata

- **What is the role of official statisticians in this context?**
  - **Be aware that this circle is not free of limitations and risks**
    - Black box, lack of transparency, hallucinations and repetitions
    - Ethical considerations, including language divides and propagation of bias
    - Confidentiality
    - Governance and stewardship

III. Towards an active role of official statistics: for a virtuous circle between AI and metadata

- **Use frontier technologies *strategically*** for the development of official statistics
- A proactive approach is needed to **keep** and **feed the virtuous circle between generative AI and metadata**
  - Moving into the age of AI: **data modelling** or *how* data are
  - This calls for:
    - More, finer and better metadata: be **accurate**
    - More dissemination and communication: be **relevant**
    - More, stronger and better literacy of users, both humans and machines: to **ensure trust** in official statistics

Irving Fisher Committee on
Central Bank Statistics

BIS

Thank you

Bilyana Bogdanova - bilyana.bogdanova@bis.org
Magdalena Erdem - magdalena.erdem@bis.org
Bianca Ligani - bianca.ligani@bis.org
Olivier Sirello - olivier.sirello@bis.org