

Workshop on the Potential of NLP in SDMX

Summary Report

March 2024

pm

30pm

Online

SDMX + AI:

UNLOCKING THE POTENTIAL
OF NLP TO ENHANCE
DATA ACCESS



OECD



sdmx



BIS

01 Introduction

Following several use cases being identified and experimented, as well as information exchanges in different fora, including the 9th SDMX Global Conference in Bahrain in November 2023, a workshop was convened over two half days on 27th and 28th March 2024 to share current practices, projects, and knowledge in the area of natural language techniques applied to data accessibility – more specifically, in **leveraging Generative AI to enhance access to official statistics disseminated via SDMX services**. The primary use case focused on data dissemination by an individual organisation – but the discussion also extended to the larger context of accessing a number of SDMX sources through an AI-enabled “universal data concierge”.














The primary objective of the workshop was to assess the main use case(s) identified in the area of enhanced access to data, delving into emerging solutions, and identifying priority areas for collaboration and co-investment. It brought more than 70 participants together including experts, practitioners, and stakeholders to foster a deeper understanding of the subject matter and explore potential avenues for working together.

This workshop was jointly organised by the OECD and BIS in the context of SIS-CC and SDMX.IO communities.

The background to the workshop can be seen in the [agenda](#).

02 Presentations and recordings

Below are the list of presentations from day 1 and day 2 as well as the recordings up to session closed to the tech providers. Please note the presentations and recordings have been made available to registered participants only with **access to non-registered participants to be granted upon request** – clicking on the presentation of interest will generate an access request, and once granted will give access to all presentations.

Presentation(s)	Link(s)
Keynote by Markku Huttunen, Statistics Finland, on AI usage in data dissemination	
Presentation of the data accessibility uses cases discussed on Day 1	
1) Rafael Schmidt, BIS	
2) Jim Tebrake, IMF	
3) Jens Dossé, OECD	
Tech providers responding to the use cases	
1) StatGPT by Maksym Samusenka, EPAM	
2) Natural Language Search with LLM by Alessandro Benedetti, SEASE	
3) StatsBot PoC and Expansion by Rahul Wane, E-Zest	
4) Using NLP methods to improve SQL data accessibility by Christoph Bergen and Julian Kurz, HMS Analytical Software	
More use cases Generative AI	
1) Applying generative AI across the data cycle by Olivier Dupriez, World Bank	
2) HLG-MOS project on Gen AI in Official Statistics by Inkyung Choi, UNECE	
3) Generative AI assessment: can we safely and cheaply run LLMs to access the right data? By Mirko Avantaggiato and Giuseppe Bruno, Bank of Italy	
4) OECD.AI Observatory and Policy Explorer by Luis Aranda, OECD	
5) Leveraging AI for metadata management, data discovery and dissemination by Bilyana Bogdanova and Rafael Schmidt, BIS	
6) ECB use cases for the AI offering by Alessandro Bonara, ECB	



Recording day 1



Recording day 2

03 Participant questionnaire results

Following participation in day 1, participants were invited to complete a short questionnaire that was designed in order to collect feedback on the four (4) technical expert presentations as well as capture further information in regard to identified or new use cases and appetite for possible future collaboration and co-investment.

A PDF version of the questionnaire was shared ahead of the workshop so that participants could collaborate with colleagues within their organisation and prepare in advance. **One response was provided per organisation.**

Use cases

How well were your organisation's needs reflected in the use cases presented by the IMF, BIS, and OECD (primarily, focusing on data dissemination with natural language and conversational capability)? (Select the one that applies)

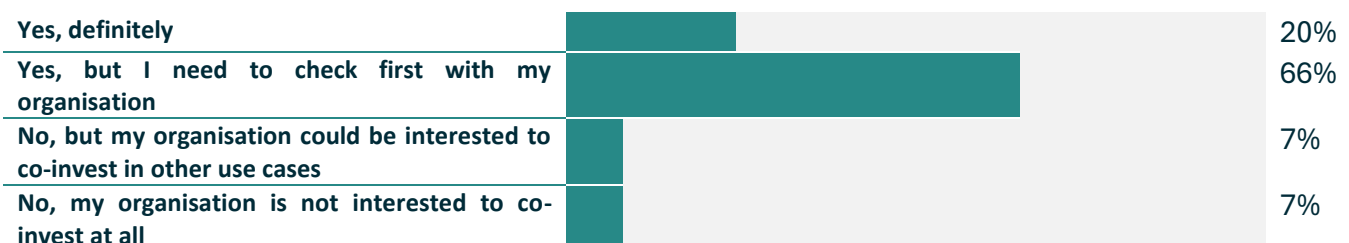


In case you have other use case(s) that are relevant to your organisation or particular risk or challenges and opportunities you see, please add them below with a concise description.

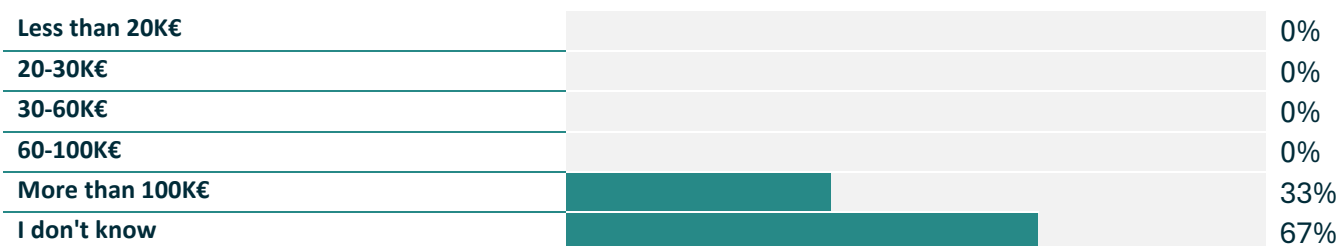
Summary of responses: *Although there was a high percentage of alignment with the use case(s) presented by IMF, BIS, OECD, responses highlighted some additional use cases and considerations including AI-based data modelling assistant, code generation and pair-programming, as well as the risks around understanding the different user requirements, and potential for costs to escalate, especially in regard to cloud based services. The responses also highlight the need for solutions that are relevant for specific organisations, while addressing various challenges and opportunities in the context of their unique requirements. Overall, the use cases and considerations presented provide valuable insights into the specific needs and potential opportunities for leveraging AI and data-related technologies.*

Interest in a co-investment approach

Would you be interested in a co-investment approach e.g., contribute financially, along with other interested organisations, to develop the AI-enabled data accessibility use case?

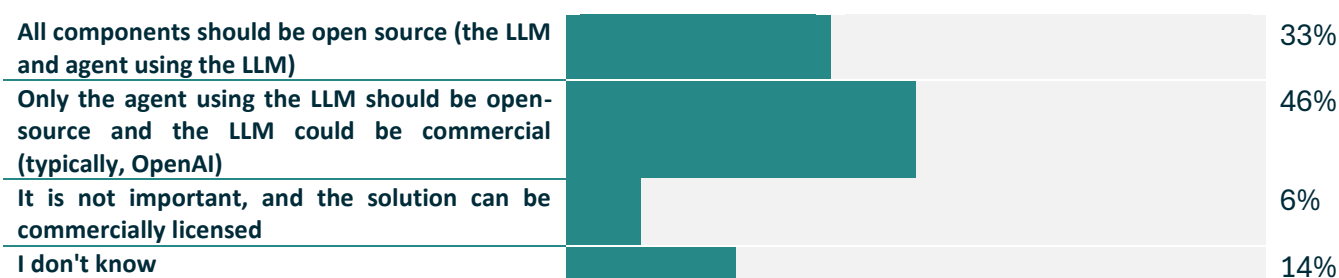


If you answered “**yes, definitely**”, please provide an indication of range of investment you could potentially mobilise over the 2024-2025 time period (*Note: the figures below are just indicative, they are not commitments but are meant to give an idea of scope*).



Open source versus commercial

How important is an open-source approach (e.g., artefacts developed through a co-investment approach should be open-source and reusable by all).



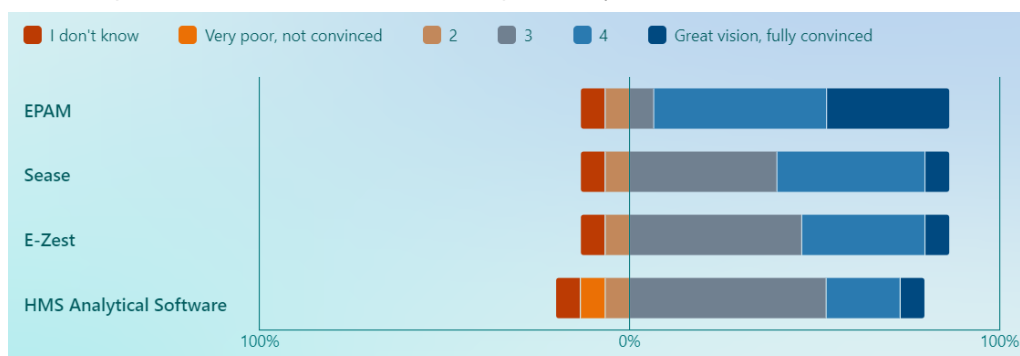
Does your institution prefer self-deployment of AI applications or purchasing managed services.



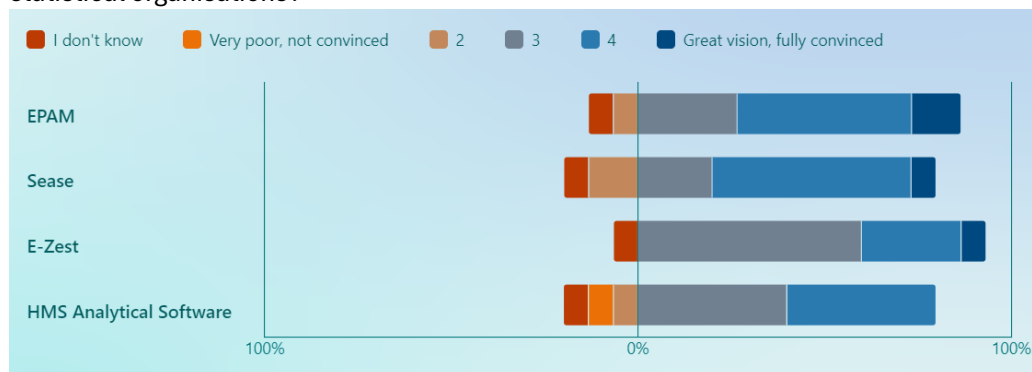
Technical Experts feedback

Four Technical Experts presented their vision for delivering better data accessibility using AI techniques. For each of them we would like you to share your impressions by providing a rating for each of the following questions. (0= I don't know; 1= very poor, not convinced at all; 5= great vision, fully convinced).

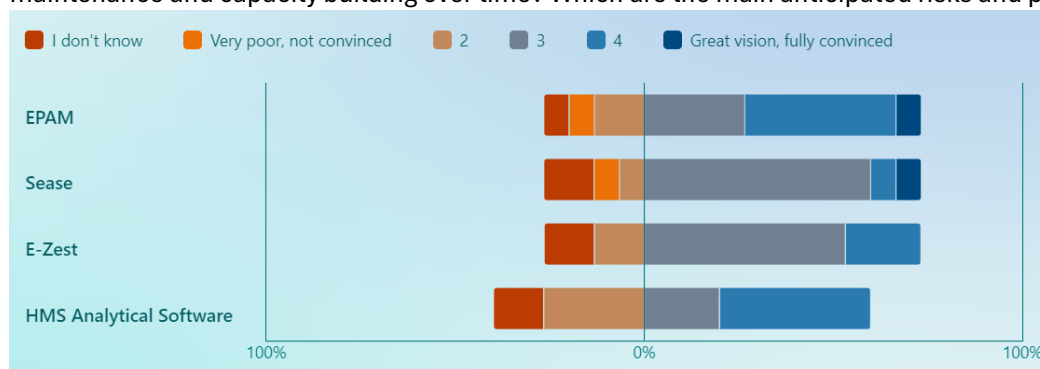
Reformulation of the data accessibility use cases in an official statistics context, as they understand them, including scenarios for the possible user experiences they envisage (combining – or not – regular, augmented search experience with a conversational experience).



Their technical vision, the target technical architecture. How far is SDMX semantics leveraged in combination with textual search and LLM/prompt-engineering techniques? Which are the components that are open source and those that are not? How replicable and scalable is their approach in the varied contexts of statistical organisations?



Their vision for the project to scale beyond the experimental phase. What is the suggested project approach (key deliverables, high level plan) for achieving production-grade services? Can this approach fit in the context of an open-source community, with a view for statistical organisations to co-invest, and mutualise the cost of maintenance and capacity building over time? Which are the main anticipated risks and pitfalls?



If you have another technical expert/provider that you think could be interesting for this group to interact with, please share more details below:

AI group at Azure (offers small POC at their costs too)

Do you have any other comments and/or suggestions or ideas for AI projects that could be developed in a co-investment approach?

Summary of responses: While some organisations are currently relying heavily on self-deployment for operations, there is an openness to adopting managed services within a cloud context. In addition to the fundamental "data accessibility use case," there is interest in exploring the "learning/knowledge use case" and the "data harmonisation enabler use case," as well as other potential use cases like "data ingestion and mapping." To move forward, it's proposed that interested parties align on the specific problem to be addressed, followed by a potential co-investment Proof of Concept (POC) with reusability in mind. Subsequent actions would hinge on the overall cost and the capacity of the final solution to meet organisation specific requirements, including project governance, IT security, and functional offering. Additionally, the development of tools for metadata augmentation and insights on fine-tuning open source Language Model Models (LLM) and Statistical Language Models (SLM) are suggested as valuable areas for attention and improvement.

04 Identifying the use case for co-investment approach

A number of use cases were described and identified over the two day workshop. Keeping in mind the objective of the workshop – identify a scope where co-investment would make sense – one "killer use case" is confirmed: data accessibility via natural language search and discovery, including with conversational mode, of data available in SDMX.

The use case is valid in that it fulfils following criteria:

- (a) the use case is specific to official statistics;
- (b) there is relative consensus on the functional expectation (illustrated by the typical UX scenarios presented at the workshop, based on .Stat Data Explorer) and value to the end user;
- (c) the delivery of a production-grade service seems within reach.

Of course, these assumptions remain to be refined and confirmed but they appeared as realistic to most of the workshop participants, with variants (such as "talk to data", "aggregate data" or "transform data" use cases). One intriguing side use case, connected to data accessibility, was introduced by the World Bank (scoring the relevance of search results or their ordering). Also, ECB mentioned that making our data "AI-ready", ready to be "consumed" by Bing and Google chat agents, is probably of equal importance to offering an AI-based UX as described above. BIS added that an SDXM+AI data accessibility project is also an opportunity to promote SDMX and the value it can bring to statistical organisations, as well as scope areas where the standard should be improved to be more "AI-ready". Lastly, IMF insisted in their presentation on the SDMX registry extension to this service, which could allow to have a central, universal access to any SDMX source, in addition to allowing for natural language / conversational access to a given SDMX endpoint; however, the question of how such a service could work around the issue of semantic incoherences appears as a challenge (a point emphasised by the ILO).

Other important use cases might call for a co-investment at a later stage; but do not appear as immediate candidates. These use cases will be further explored, especially in the context of the HLG-MOS "generative AI for official statistics" project, which was introduced by UNECE during the workshop and which participants are invited to join. Of notable importance, the following use cases were mentioned:

- A second use case, or family of use cases, appeared as highly promising: metadata enrichment, augmentation, harmonisation (and translation). Geared to data producers, they essentially consist in facilitating their work through LLM-based extraction/generation of information, applied to structural metadata (SDMX data model harmonisation) as well as referential or process metadata. These use cases definitely fulfil criteria (a), but more analysis is required to reach consensus (b) and scope co-investment (c).

- A third use case, or family of use cases, retained a lot of attention: leveraging generative AI to support code generation, pair-programming applications... Clear productivity and quality gains are expected from these use cases, as well as possibility to upskill statisticians to newer techniques. However, these use cases appeared less specific to official statistics (criteria (a)) and could possibly rely on generic solutions offered by the market.

05 Framing the co-investment approach

The questionnaire allowed to identify 13 organisations definitely ready (3) or potentially ready (10) to participate in a co-investment scheme to deliver the SDMX+AI data accessibility use case, at a production-grade level. From the tech providers presentations, the characteristics of the target architecture emerge:

- The opportunity of leveraging SDMX semantics to make data “AI-ready” and accessible;
- RAG approach involving LLM to generate or extract information;
- Initial investment need to achieve a production-grade “minimum viable product” are in the same range (150k€, 6 months);
- Architectures illustrated or already deployed are similar but rather complex and require to harness a range of technologies, hence the preference expressed by the majority to resort to a third party to manage the system;
- The question of which LLM to use is still wide open; avoiding lock-in with one particular technology and being able to adapt to different models appears as a must;
- The operating costs of commercial LLMs required to fulfil the use case (e.g. GPT-4) are high and scale strongly with usage. Although it was generally agreed that operating costs are likely to fall as LLM technology becomes more commodity and competition increases, operating costs could present a challenge for institutions wishing to adopt a solution;
- Pace of technology development means solutions are likely to quickly become obsolete – it was agreed that small steps are needed.

The following additional criteria were agreed on in this part of the conversation:

- (d) the solution is inclusive, that is, deployable and/or usable in any statistical office, even least wealthy ones; a criteria particularly emphasised by UNSD, WB and IMF;
- (e) the solution is open source (source code is accessible and reusable with no limitation; being open source can mean a lot more as can be seen [here](#)) but with the important caveat that commercial LLM could be part of the solution (and yet pose a financial challenge as stated above);
- (f) the solution is vendor agnostic (especially, can be deployed over any cloud – AWS, GCP, Azure – and, potentially, on premise).

06 Conclusion and next steps

Two scenarios for co-investment are envisaged:

Scenario 1: leverage work done by IMF/EPAM and make StatGPT a global solution. Provided this scenario is compatible with criteria (a) to (f), it should ideally allow to save time and resources for everyone.

Scenario 2: start from scratch an open source approach (the way presented by SEASE, E-Zest or HMS). This scenario would by design comply with set criteria and seems within reach.

There is a wide consensus amongst participants that scenario 2 should be explored only if scenario 1 proves incompatible with either of the set criteria. Also, in the questionnaire, EPAM appeared as the most convincing in terms of functional understanding, architecture vision and project approach. A hybrid scenario could also be possible in that, alternative providers could contribute to the evaluation of and complement the EPAM value proposition in later stages.

On that basis, the (tentative) envisaged next steps are the following ones:

- (end of April if possible) IMF/EPAM to draft a note explaining the value proposition and project approach based on the workshop exchanges and conclusions. All workshop participants ready to contribute as need be.
- (end of May if possible) Convene a meeting with organisations interested in a co-investment approach, and based on the IMF proposal, to review it and ask questions or complement the approach.
- (end of June 2024) The project scope and intentions of contributors are confirmed.

The project could possibly be formally agreed and start by September-October; that should allow for statGPT being deployed in other contexts by the spring of 2025.

07 List of Participants

Following representatives participated in-person.

First Name	Last Name	Affiliation
Christoph	BERGEN	Analytical Software
Julian	KURZ	Analytical Software
Rafael	SCHMIDT	Bank of International Settlements (BIS)
Bilyana	BOGDANOVA	Bank of International Settlements (BIS)
Glenn Philip	TICE	Bank of International Settlements (BIS)
Giuseppe	BRUNO	Bank of Italy
Maksim	SAMUSENKA	EPAM
Jean-Luc	PLAGNAUD	EU Commission / Eurostat
Alessandro	BONARA	European Central Bank (ECB)
Ana	PADRÃO	European Central Bank (ECB)
Gábor	HORVATH	European Central Bank (ECB)
Edgardo	GREISING	International Labour Organisation (ILO)
Luc	ROETTIGERS	Ministère de l'Economie - STATEC
Petra	MELLAERTS	National Bank of Belgium (NBB)
Frederik	VAN HECKE	National Bank of Belgium (NBB)
Chloe	ACAS	Organisation for Economic Cooperation and Development (OECD)
Eric	ANVAR	Organisation for Economic Cooperation and Development (OECD)
Luis	ARANDA	Organisation for Economic Cooperation and Development (OECD)
Laura	BELLI	Organisation for Economic Cooperation and Development (OECD)
Jonathan	CHALLENGER	Organisation for Economic Cooperation and Development (OECD)
Jens	DOSSÉ	Organisation for Economic Cooperation and Development (OECD)
Petko	YANEV	Office fédéral de la statistique (OFS)
Timothee	RONDEZ	Office fédéral de la statistique (OFS)
Jonas	DEPLAZES	Office fédéral de la statistique (OFS)
Alessandro	BENEDETTI	Sease
Markku	HUTTUNEN	Statistics Finland
Olivier	DUPRIEZ	World Bank

Following representatives participated online.

First Name	Last Name	Affiliation
Olivier	SIRELLO	Bank for International Settlements (BIS)
Stratos	NIKOLOUTSOS	Bank for International Settlements (BIS)
Brian	BUFFETT	Bank for International Settlements (BIS)
Edward	LAMBE	Bank for International Settlements (BIS)
Daniele	OLIVOTTI	UNICEF
Abdulla	GOZALOV	UN Statistics Division (UNSD)
Amir	KHATIB	Bank of Israel
Aymen	CHAREF	Food and Agriculture Organisation of the United Nations (FAO)
Sergio	PENA	Food and Agriculture Organisation of the United Nations (FAO)
Fadhila	NAJEH	Food and Agriculture Organisation of the United Nations (FAO)
Frano	ILICIC	Food and Agriculture Organisation of the United Nations (FAO)
Galya	STATEVA	Eurostat

Liam	BYRNE	Australian Bureau of Statistics
Hugh	STEHLIK	Australian Bureau of Statistics
Jim	TEBRAKE	International Monetary Fund (IMF)
Jeff	DANFORTH	International Monetary Fund (IMF)
Jerome	SOLIS	MILA Quebec
Josep	ESPASA REIG	LIS Cross-National Data Center in Luxembourg
Luca	GRAMAGLIA	Eurostat
Manuel	CUELLAR-RIO	Instituto Nacional de Estadística y Geografía (INEGI)
Mara	SÁNCHEZ	Instituto Nacional de Estadística y Geografía (INEGI)
Martha	GARCIA ACEVEDO	Instituto Nacional de Estadística y Geografía (INEGI)
Irving	CABRERA	Instituto Nacional de Estadística y Geografía (INEGI)
Almir	DELIC	European Central Bank (ECB)
Zlatina	HOFMEISTER	European Central Bank (ECB)
Ole	SORENSEN	European Central Bank (ECB)
Stefano	PAMBIANCO	European Central Bank (ECB)
Juan	ALBERTO SÁNCHEZ	European Central Bank (ECB)
Paddy	POWER	Statistic New Zealand
Gyorgy	GYOMAI	Organisation for Economic Cooperation and Development (OECD)
Patrice	ORLIANGE	Organisation for Economic Cooperation and Development (OECD)
David	BARRACLOUGH	Organisation for Economic Cooperation and Development (OECD)
Pedro	CARRANZA	Organisation for Economic Cooperation and Development (OECD)
Sandrine	SOUFLIS	Organisation for Economic Cooperation and Development (OECD)
Jean-Baptiste	NONIN	Organisation for Economic Cooperation and Development (OECD)
Anastassia	SAMSONOVA	Organisation for Economic Cooperation and Development (OECD)
Amal	LAHMAR	Organisation for Economic Cooperation and Development (OECD)
Yeojin	YOON	Organisation for Economic Cooperation and Development (OECD)
Sophie	MONNIER	Institut national de la statistique et des études économiques (INSEE)
Patrick	CLOUTIER	Statistics Canada
Stephane	CRETE	Statistics Canada
Steve	ARMSTRONG	Statistics Canada
Mark	FALVO	Statistics Canada
Christian	MASSICOTTE	Statistics Canada
Sudipta	DUTTA	Reserve Bank of India
Daniel	GOLDFUß	Analytical Software
Ilya	GORELIK	EPAM
Michael	GAVRONSKY	EPAM
Mandar	GARGE	e-Zest