Workshop jointly hosted by

SIS-CC
Statistical Information System Collaboration Community

SDMX.IO

Wed 27 March 1pm-6pm
& Thu 28 March 9am-12.45pm
OECD, Paris, France + Online

SDMX + AI: UNLOCKING THE POTENTIAL OF NLP TO ENHANCE DATA ACCESS

OECD sdmx BIS

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, it was agreed to convene a 1st workshop 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 of interest is that of data dissemination by an individual organisation – but the discussion could be extended to the larger context of accessing a number of SDMX sources through an AI-enabled "universal data concierge".

The primary objective of this workshop is to assess the main use case(s) identified in the area of enhanced access to data, delve into emerging solutions, and identify priority areas for collaboration and co-investment, bringing together experts, practitioners, and stakeholders to foster a deeper understanding of the subject matter and explore potential avenues for working together.

This workshop, jointly organised by the OECD and BIS in the context of SIS-CC and SDMX.IO communities, will be conducted in a hybrid format, spanning two half days. It is scheduled to take place on 27th March 2024 from 1:00 PM to 6:00 PM CET and on 28th March 2024 from 9:00 AM to 12:30 PM CET. For those attending in person, the venue is the OECD Boulogne Meeting Room BB10, located at 46 Quai Alphonse le Gallo, 92100 Boulogne-Billancourt.

Participation in this workshop is by invitation only with the following groups invited, SIS-CC and sdmx.io Community members, invited partners (such as Eurostat), SDMX sponsors (including ECB, IMF, World Bank, UNSD), and UNECE and selected organisations contributing to the HLG-MOS 2024 project Generative AI for Official Statistics. We aim to create a diverse and collaborative environment that encourages fruitful discussions and cross-sectoral cooperation.

To enrich the discussions, four (4) technical providers have been invited, including EPAM, E-Zest, Sease, and HMS Analytical Software, who are actively experimenting in this area. **These experts will join us on the Day 1 of the workshop, as well as the first part on Day 2**, to share their insights, present their approaches, discuss possible collaborative approach to joint developments, and provide an early indication of possible resource implications. Their expertise and experiences will contribute significantly to shaping the discussions and outcomes of the workshop.

To ensure the active participation and engagement of all participants, they are requested to carry out preparatory work ahead of the workshop (see next page). In addition, an online questionnaire will be shared, for participants to assess each proposal presented by the technical experts at the end of Day 1. The questionnaire results will help qualify the individual needs and scenarios of participating organisations and identify appetite and opportunities for co-investment – in the area of data accessibility and more broadly, in applying generative AI to statistical work. The insights gathered will pave the way for meaningful conversations during Day 2, and exploration of possible project scopes, including synergies with the wider communities.

We look forward to your active participation and contributions in shaping the future collaboration in this new and emerging area of work on *SDMX + Generative AI*. **Register by selecting one or both days below** *(registration is required for each day separately)*:

Day 1

Day 2

Participant preparation

Participants are invited to prepare ahead of the workshop in order that the discussions are productive and the overall workshop objectives are met. **Specifically, we request that participants review the background material from each of the technical experts in relation to the project experimentations that will be covered during Day 1**. Knowledge of this background is considered a pre-requisite for all participants. The preparatory time needed for this preparatory work is in the range of 2-3 hours.

In addition, participants should review this questionnaire (not to be completed) that will be shared as an online form at the end of day 1 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 plans and appetite for possible future collaboration and co-investment.

Technical provider and project	Material type	Link
EPAM StatGPT	Previous related presentation (video – watch from 1:57:23)	
	Presentation file of the above	
	Background document 'Discover statistical data with Stat-GPT'	
SEASE Natural Language Search with LLM	Previous related presentation (video – watch from 2:44:46)	
	Presentation file of the above	
	Final report from PoC	
	Source code repository (PoC)	
	Components and sequence diagrams	
	Example queries	
E-Zest StatsBot PoC and Expansion	*Material not yet provided*	
HMS Analytical Software, Using NLP methods to improve SQL data accessibility	*Material not yet provided*	

Participants are also invited to consult the additional reference material listed below.

Additional reference material	Link
HLG-MOS White Paper (draft): Large Language Models for Official Statistics	
2023 SDMX GC Keynote: Why SDMX Matters? A Community journey towards SDMX as an enabler for AI and the Data Mesh by Eric Anvar, Head of Smart Data, OECD	
How many: answering common fuzzy questions with precise data responses – Yves Jaques, UNICEF	

Agenda

Day 1: 27th March 2024, chaired by Rafael Schmidt, BIS

Time*	Session	Content & Presenter(s)
1:00PM	Introduction	**Introductory remarks** by Eric Anvar, OECD, and Rafael Schmidt, BIS
1:10PM	Setting the scene	**Keynote** by Markku Huttunen, Statistics Finland, on AI usage in data dissemination
1:35PM		**Presentation of the data accessibility uses cases discussed on Day 1** 1) Rafael Schmidt, BIS; 2) Jeff Danforth, IMF; 3) Jens Dossé, OECD
2:00PM	Tech providers responding to the use cases (I)	1) **StatGPT** by Maksym Samusenka, EPAM 2) **Natural Language Search with LLM** by Alessandro Benedetti, SEASE
3:30PM	**Coffee break**	
4:00PM	Tech providers responding to the use cases (II)	3) **StatsBot PoC and Expansion** by Rahul Wane, E-Zest 4) **Using NLP methods to improve SQL data accessibility** by Daniel Goldfuss, HMS Analytical Software
5:30PM	Looking ahead to Day 2	**Launch of questionnaire** by Jonathan Challener, OECD, to qualify, for each organisation, interest in the presented approaches and opportunities for co-investment, and identify more broadly use cases of interest in using generative AI in official statistics.
5:45PM	Wrap-up	**Day 1 concluding roundtable** to capture last thoughts of the day and prepare for Day 2
6:00PM	**Close**	
Evening	*Subscription social dinner for interested participants at Le Grand Comptoir de Boulogne.*	

** Please note times are in CET and are indicative and open to change.*

In their response to the presented use cases, the four (4) technical providers are invited to cover the following points:

1. **Provide the prep material that participants are expected to review ahead of the workshop** (see previous page – workshop preparation) – in the shape of videos (including demos if applicable), reports and sharing of codes.
2. **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).
3. **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?
4. **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? Presenters are invited to provide elements on the resource implications of their approach so that participants can have a grasp of the required intensity of investment.

Note: *Technical providers are invited to participate in each other's presentation, although in an observer mode. They were all consulted ahead of the workshop and agreed to this open sharing approach. They will also participate in Day 2, except the last part (discussion on next steps).*

Day 2: 28th March 2024, chaired by Eric Anvar, OECD

Time*	Session	Content & Presenter(s)
08:30AM	*Welcome coffee and pastries available for in-person participants ahead of the meeting start.*	
9:00AM	Recap on Day 1 & Roundtable	**Report on Day 1 questionnaire results** by Jonathan Challener, OECD, followed by a roundtable on the interest in the presented approaches and opportunities for co-investment, and more broadly use cases in using generative AI in official statistics.
09:30AM	More use cases Generative AI	Looking more broadly at applying generative AI in official statistics and creating synergies between networks: 1) **Applying generative AI across the data cycle** by Olivier Dupriez, World Bank 2) **HLGMOS project on Gen AI in Official Statistics** by Inkyung Choi and A. Kipkeeva, 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 Presentations of 12 minutes each, followed by a 20-minute discussion, to include indications on opportunities for synergies between networks represented by the presenter (such as: HLG-MOS, IFC, OECD.AI).
11:05AM	**Coffee break**	
11:30AM	Next steps	1) Discussion to identify opportunities for co-investment in the area of data accessibility 2) Contributing working with broader initiatives and networks such as the ones identified ***Note**: Tech providers will not participate in this last session*
12:45PM	**Close**	
Lunch	*Provided to in-person participants as an opportunity to continue the discussion and network.*	

** Please note times are in CET and are indicative and open to change.*

Information for in-person participants

How to reach the OECD

The workshop is to take place at 46, quai Alphonse Le Gallo – 92100 Boulogne-Billancourt, located on the Seine opposite Parc Saint Cloud.

Several transportation options are available including:

Metro/RER

- Line 10, Pont de Saint-Cloud station (10 minutes by foot)
- Line 9, Pont de Sèvres station (15 minutes by foot)
- Future Grand Paris Express station

Bus routes

- 169, 171, 179, 279, 291 – Pont de Sèvres bus stop
- 52, 72, 126, 160, 175, 467 – Pont de Saint-Cloud bus stop

Tramway

- T2 – Saint-Cloud station (15 minutes by foot)

Bike rentals

- Vélib-metropole: A bicycle hiring option – an alternative way to move in Paris.

More information to make your trip to France easier including list of hotels, visa requirements, and other useful information is available on the OECD web site.

Security and registration

For in-person registration, please indicate during the online registration process for day 1 and day 2. If you have any questions or issues with the registration please send an email to Chloe ACAS, Chloe.ACAS@oecd.org with Jens DOSSÉ, Jens.DOSSÉ@oecd.org on copy.

An electronic visitor pass will be sent to your email address one day prior to the start of the workshop. Please ensure you have this ready, either on your mobile phone or a printed copy. This is needed in order to pass through the first checkpoint before entering the main OECD building and security screening.

Badges will be provided to all visitors to enable those to access the OECD building and meeting room. Please remember to have your passport or identity card ready as proof of your identity. The badges are presented at the welcome entrance desk and are to be worn at all times within the OECD building. On arrival, you must register at the reception desk to obtain your visitor's badge. You will need to bring photo identification with you to receive your badge. All badges will be printed and waiting for you. These badges will give you access to specific areas during the period of the workshop.

Please arrive well in advance of the start of the meeting to allow sufficient time for the security and registration formalities. Participants can start to arrive from 12.15pm and we suggest no later than 12.40pm.