UNECE High-level Group for the Modernisation of Official Statistics

Business Case for StatsBots – Bringing reference facts to the online conversation

This business case was prepared by Jonathan Challener, OECD, and is submitted to the HLG-MOS for their approval.

Type of Activity			
☒	New project	☐	New activity
☐	Extension of existing project	☒	Extension of existing activity Background: "StatsBot project – report to HLG-MOS Exec Board", Annex 3

Purpose

Data deluge continues to grow at an exponentially rate where no longer Statistical organisations, at national and international levels, being the main source for data that feeds the policy making machine. This growth, driven by user demand for data, where users have the ability to search and consume information of any kind through online channels and digital devices, calls for a need to innovate in the dissemination of official statistics, so as to bring reference facts to the conversation.

Statistical organisations, as producers of official statistics, must continue to ensure that the right data is available to the right person, at the right time, in the right way. Policy-makers and policy shapers (notably the media) have a key role to play in this regard, as they are the ones who structure the conversation on policy and expose the facts underlying a particular political issue. They need reliable facts that they can access easily, and rely on in their wider conversation with the public. Statistical organisations have a particular role in enabling this conversation around facts with new digital technology emerging that can greatly support this.

Description of the activity

Drawing on the experiences and results of a proof of concept to develop a Chatbot for Official Statistics undertaken by the OECD, CBS NL, and StatCan, it is proposed to **launch a large scale project for the joint development of a generic StatsBot that draws on statistical structured sources and rich semantics**.

Having focused on the strategic question of scalability, that is, how could linguistic / symbolic AI capabilities, combined with data richness (semantic structure according to SDMX especially), lead to an approach that can scale, from one topic / language / organisation to another, with limited cost. The PoC was limited to one domain (Labour statistics, made available in SDMX format through a .Stat Suite API, sourced from the 3 organisations) in order to identify roadblocks to scalability, and derive from the analysis an approach to build a generic, scalable StatsBot that could expand across topic, languages and source organisations with limited additional cost.

At the end of the PoC, the remaining functional gaps were assessed, along with recommendations in terms of optimal data modelling (in SDMX) to enable efficient data sourcing by the StatsBot. It is proposed to first carry out a larger scale market consultation that would lead to a) the selection of a target technology (either the one selected for the PoC, or another one) and b) the joint development of the generic StatsBot in a second step, ultimately c) the maintenance over time of the StatsBot. The resulting StatsBot should be able to connect to any well shaped SDMX source and be able to assist in the exploration of the data through a conversation – at a relatively limited, beyond the initial investment, for new topics/domains/languages.

In order to reach a production grade StatsBot, including a much improved user interface that could adapt to different sources, languages and topics with limited incremental cost, it is estimated that this would take

approximately quadruple the time spent for the PoC (so, 8 sprints of 2 weeks vs 4 sprints of 1 week). It is assumed, based on the PoC technology partners (Golem.ai) advice, that the technology under the hood (symbolic AI, as opposed to traditional ML in most chatbot technologies on the market) allows for capitalisation (e.g. when one organisation has proofed the chatbot for tourism statistics in a given language, other organisation should be able to leverage the knowledge with limited additional cost) – leveraging the StatsBot back office for non-technical experts to assess the chatbot performance and improve the configuration of it over time.

Alternatives considered

Alternative to the development of a generic StatsBot could be to co-invest first in a joint user research program (see alternative proposal). The two approaches can of course be combined. In this alternative approach, statistical organisations would try to apply best-of-breed user research techniques, in a systematic and global way, in order to first define scenarios for the next generation of data experience, of which conversational access to data through bots is but one channel. The emphasis would then be more on the overall analysis and experiemntations related to the invention of the next generation of data experience, leveraging AI and other techniques – rather than immediately delivering a production grade solution.

How does it relate to the HLG-MOS vision and other activities under the HLG-MOS?

The idea of a collaboration among statistical organisations emerged in late 2018, after CBS presented their concept of a voice assistant responding to questions on socio-economic data, during the HLG-MOS meeting in Geneva. Bilateral discussions ensued, where several statistical offices shared results on their experiences in the field. Eventually, the project took the shape of an HLG-MOS activity leading to the formation of a 'working group developing a Statistical Chatbot for Official Statistics'.

A call to action soon followed and the OECD (as lead organisation of SIS-CC community, whose aim is precisely to co-invest in common platforms) helped to push this agenda forward with the formation of the WG during their first meeting in Paris in March 2019. This then took the shape of a Proof of Concept, developed as a BSTN activity and has been monitored as such since then. The 2020 BSTN activity is to be closed on Nov 20th, during the HLG-MOS workshop dedicated to the topic.

Accepting this proposal would entail to position the project no longer as an exploration activity, but as a project to deliver a production grade platform, and involving fund raising and coordinated execution. An international organisation could typically host the project (eg support project coordination, fund raising, and procurement and delivery activities); the OECD could take on that role, in the context of the HLG-MOS governance, and leveraging know-how and infrastructure developed in the context of the .Stat Suite project.

With a main goal to support the modernisation of official statistics through new and innovative ways in which our statistical organisations operate as well as disseminate data, this aligns well to the HLG-MOS vision.

Proposed start and end dates

Start: January 2021	**End:** December 2021

The estimated budget for the project is in the range of 150 to 300k€, with a recurring cost in the range of 20% of the initial investment. If, say, 10 organisations would unite to develop the generic StatsBot, the resource burden could be split and maximal knowledge sharing and cross-fertilisation could happen. Such an investment should be released in an iterative and agile manner, whereby value is delivered at each iteration and can be checked with actual end users.