UNECE High-level Group for the Modernisation of Official Statistics

Business Case for a Statistical Open Source Software Project

This business case was prepared by Barteld Braaksma (Statistics Netherlands) and is submitted to the HLG-MOS for their approval.

Type of Activity			
☒	New project	☐	Extension of existing project

Purpose

The Official Statistics community is increasingly looking into the opportunities of statistical open source software (SOSS) to replace existing (legacy) IT solutions based on closed (proprietary) software. The reasons for this interest are manifold: cost-efficiency, collaboration opportunities, avoiding vendor lock-in, quality improvements and fast development cycles. Moreover, the shift to OSS aligns with the fact that newly recruited staff members often have thorough knowledge in languages like R and Python, that are often used to develop statistical (and data science) open source solutions. At the same time, there may be challenges concerning support, maintenance, training, sharing conditions, legal aspects, etc. Given the generic and universal nature of the issues that arise and the huge gains that can be obtained, there is a strong case for collaboration in the official statistics community and formulating common policies. A series of knowledge sharing and discussion sessions, organised as a BSTN activity in 2023, already showed clear interest in various aspect of the OSS topic. Simultaneously the CES and the ESS organised activities around SOSS.

The purpose of the Statistical Open Source Software project is to develop a better common understanding of the pros and cons, do's and don'ts of moving to further and more comprehensive use of open source software as a cornerstone for official statistics production, based on concrete experiences through use cases of broad interest.

The experiences with CSPA have shown that developing common use cases is quite difficult. Even within a single organisation, finding a common basis for open source cases of broad appeal is not trivial. There are, however, concrete examples and also new opportunities both in the input (data collection), throughput (analysis and processing) and the output (dissemination) domains. Suitable use cases may be found in all of these domains.

In the input domain, statistical organisations are looking into the use of apps to support/replace existing data collection systems. One example concerns Rapid Survey Systems (RSS), to collect data in areas where there is a sudden (and often urgent) need for additional information. Several institutes have built their own RSS solutions. It may be worthwhile to explore if there is a common ground here. As another example, the ESSnet Smart Surveys has conducted research on the development of apps to support the ESS Household Budget Survey and the ESS Time Use Survey. The transition from research to production would benefit tremendously from a collaboration model that includes an open source approach.

In the throughput domain many open source R-packages have been developed by multiple NSIs and are used throughout the ESS. See the Awesome List of official statistics software (www.awesomeofficialstatistics.org) for many examples.

In the output domain, the .Stat suite[1] is used by quite a number of statistical institutes, and that number is growing. The suite is supported by the SIS-CC community. Many users develop their own code on top of the core .Stat functionality, to integrate it better in their own environments. It may be useful to share code, experiences and practices in this area and identify which role a community like SIS-CC could play. Another interesting example is PxWeb (formerly PCAxis), used for many years by several users, originally developed by Statistics Sweden and now released as open source with the support of an ESSnet. Two further widely used open source examples in the output domain are Demetra for seasonal adjustment and the Argus package for statistical disclosure control.

Description of the project and the Work Packages/sub-activities

The Statistical Open Source Software project for 2024 could consist of several work packages. The exact definition can be left somewhat open at this stage and be made more concrete if the project is launched and the interests of participants are better known.

Workpackage 0 is a preliminary activity to decide on the **scope and ambition** level of the project. Which generic aspects to consider? Which particular use cases to deal with? A specific question is also if open source software developed in other communities (Pandas, PySyft, Spark, TensorFlow, …) should be considered.

Workpackage 1 focuses ***on generic aspects*** of the systematic use of open source-based approaches for official statistics and follows a top-down approach. Several sub-packages may be defined, covering aspects like organisation of maintenance, support and training; standards and principles; legal aspects and liabilities/responsibilities; licensing model and fair distribution of costs; community building, communication and engagement with e.g. the scientific community and private sector.

Workpackage 2 deals with concrete open source-related ***use cases*** in the input, throughput, or output domains, in a bottom-up way. Use cases to be covered can be largely developed separately, while it remains important to define a way to learn from each other and interact with WP1.

Workpackage 3 deals with the usual project-wide activities like ***management and communication*** (internal and external), but also has an important role to facilitate and stimulate interaction between WP1 and WP2 and their sub-packages/use cases, to make sure there is cross-fertilization between practical experiences and theoretical ideas. Also, this WP will be the linking pin with other SOSS activities, such as the ESS OS4OS group), that has e.g. defined a set of principles for OSS in the ESS; and the Use of R in Official Statistics (uRos) conference.

Deliverables and timeline

The main deliverables for **Workpackage 1** will be templates/frameworks/checklists/guidelines for a standard approach for statistical open source software. The exact contents can only be decided along the way. It might be necessary to include an update strategy for the deliverables as well.

The main deliverables of **Workpackage 2** are concrete open source packages for the use cases considered, including (steps towards) core communities to support them.

The main external deliverable of **Workpackage 3** is a communication plan to spread the results from the project, which may include documentation/guidelines, training courses (e.g. related to the 2022 Meta Academy or 2023 Carpentries projects) and/or one or more workshops.

[1] From the SIS-CC site: "The **.Stat Suite** is a standard-based, componentised, open-source platform for the efficient production and dissemination of high-quality statistical data. The product is based on the General Statistical Business Process Model (GSBPM) and the Statistical Data and Metadata eXchange (SDMX) standards."

Offices/Countries committed	
Right now, no offices or countries have formally committed but we expect sufficient interest based on recent activities in this area. In 2023, the BSTN organised a series of online meetings around open source (transformation)-related issues which attracted a wide audience. Moreover, in the European Statistical System a group around open source (OS4OS) was active that e.g. developed open source principles. Moreover, open source was one of the topics at the CES plenary meeting in June 2023. And there is a lot of interest in using ChatGPT-like systems to translate proprietary code to open source (SAS-to-R). It is important to choose a 'leave no-one behind' approach: countries are at different maturity levels and resources may differ depending on e.g. their sizes.	
Alternatives considered	
An alternative could be to continue the activities under the BSTN umbrella or wait for further initiatives from the European Statistical System, but 1) this topic attracts a lot of interest beyond the European Union and 2) this means that the visibility created by a formal project will not be achieved, and we miss the opportunity to create international guidelines and common perspectives that are broadly shared and supported.	
How does it relate to the HLG-MOS vision and other activities under the HLG-MOS?	
Further and more comprehensive use of (statistical) open source software directly contributes to modernisation of official statistics production and is therefore directly aligned with the HLG-MOS vision. There are potential synergies with past and present projects, as well as with existing modernisation committees and thematic meetings.	
Proposed start and end dates	
Start: January 2024	**End:** November 2024
Further work plan will need to be elaborated by the project manager and project team, under guidance of the HLG-MOS and the Executive Board	