Co-Development of Open source solutions

The .Stat Suite Business Case

High Level Group on Modernisation of Statistics (HLG-MOS)

UNECE, Geneva, November 2023

Prepared and presented by Eric Anvar, OECD – eric.anvar@oecd.org

#OpenSource #OpenKnowledge #CoInnovation #CoInvestment #SDMX #Community
“Our project is open source... The code is posted on GitHub...”

Great, but can this really be called “open source”?

“Posting on GitHub” contributes to a fraction of the “Open Source” value proposition!
...there are many more ingredients to deliver “Open Source” assets.
#OpenSource #OpenKnowledge #Community

**THE RECIPE: FORMULATE A (PUBLIC) STRATEGY**

- **FULL DATA LIFECYCLE**
- **COMPONENT-BASED ARCHITECTURE**
- **COMMUNITY DRIVEN**

**STRETCHING THE VALUE**

**OPEN SOURCE DELIVERY**
There are many great #OpenSource projects driven by statistical orgs

#OpenSource projects are great opportunities for #CoInvestment and #CoInnovation
SIS-CC ("the Community") is a reference open-source community for official statistics, focusing on product excellence and delivering concrete solutions to common problems through co-investment and co-innovation.
if you want to go fast, go alone; if you want to go far, go together
if you want to go slow or not go at all, go alone;
if you want to go fast and far, go together

Source: Who first said: if you want to go fast, go alone; if you want to go far, go together? | Andrew Whitby
#OpenSource #OpenKnowledge #Community Ingredients

THE RECIPE: FORMULATE A (PUBLIC) STRATEGY
Formulate a strategy and make it public

SIS-CC 2020-2025 strategy results from multiple workshops and ongoing conversations over the 2015-2020 period.

Source: SIS-CC 2020-2025 Strategy
SIS-CC 2020-2025 strategy

Achieving common goals

Stretching the value

Consolidating the foundations

Source: SIS-CC 2025-2030 Strategy
#OpenSource #OpenKnowledge #Community Ingredients

COMPONENT-BASED ARCHITECTURE
.Stat Suite as an implementation of #SDMX

- .Stat Core: to store data accessible through SDMX APIs
- .Stat Data Explorer: to explore data from any SDMX endpoint(s)
- .Stat Lifecycle Manager: to manage data and metadata through the lifecycle
#SDMX drives the user experience

Step 1 Home page
Lexical search or Topics

Step 2 Search results
Filtering on dynamic facets

Step 3 Visualisation page
Selecting dimensions / variables
#SDMX enables a cooperative “Business Model”
Main focus so far on the Dissemination use case…
The #SDMX Backbone

MANAGE DATA PROJECT & ALGORITHMS

Source Data

Collect

Process

Diss. Staging

Diss. Final

Design

Archive

Powered By

StatSuite

Domain X-Collect

Domain X-Process

Data Hubs

MANAGE & COLLECT DATA

DISSEMINATE & ENGAGE

Data Hubs
The “Data Integration” use case at the National Bank of Belgium

Source: SALSA Project – National Bank of Belgium
“open” does not mean “free”

You need a sustainable Community model to fund Support, Coordination and Promotion activities As well as Product Development and Innovation
A multi-tier #Community model

**Benefit from Tech-partner commercial support**
(e.g. Malta, IEA..)

**Benefit from Tier-1 member support**
(e.g. LMIS 30+ countries – El Salvador, Jordan…, UNGP – Maldives, Cambodia…, SPC – Pacific Islands)

**Contribution**
80k€ initially + 30k€/year + ad hoc grants on need basis

**Product Development**
Strategy (SLG) Priorities (MLG)
Technico-Functional design (ATF)
User research (UTF) – new!

**Support**
OECD central team
Peer support between members
Support by key partners (ES, BIS)

**Tier 1**
core members
SLG/MLG/ATF

**Technology Partners**
Tier 2 members (.Stat as a service)
supported by reference partner(s)

**Tier 2 members**
regional/sectorial networks
supported by a core member

**Tier 3**
open source
users with no support

**On their own**
(e.g. Thailand, UNESCAP…)

**Tier 1+**
# Community model means full transparency to build trust

High level product overview and flight planner

Kanban Board incl. scheduled backlog + wish list

Changelog & Extensive documentation

Code repository
#OpenSource product requires to get the licensing right

Open source model and operations defined through a specific research work with [OSS Watch](https://oss-watch.org) to examine the conditions (operations, governance, legal) for going open source.

Stat Suite moved to MIT License in June 2019

Stat Suite is now Open!

June 24, 2019

The Community has reached a major milestone, officially launching in open source the Stat Suite product, a native SaaS modular set of components that aims to progressively cover the full statistical lifecycle (GSDPM: Design, Collect, Process, Analyze, Disseminate).

By doing so, the Community now makes the Stat Suite accessible to a wider group of organisations, at lower cost (free, MIT license), leveraging existing open innovations and communities such as REACTS, R, Qlik, Tableau, etc., and through software reuse and interoperability—eschewing market standards such as Excel or Stats

The software is now delivered through state-of-the-art DevOps mechanisms, combining effective agility, multiplicity of contributors, simplicity of deployment on the cloud, software quality and security.

We invite interested organisations or individuals to explore further our open source project to know if you would like to engage and potentially contribute. Please visit the ‘Developers’ section if you would like to know more.
#OpenSource product requires an #OpenSource stack

SDMX connectors

SDMX libraries

ArcGIS, Power BI, Excel Add-in, SDMX tools, SDMX libraries, .Stat Data Explorer, .Stat Data Lifecycle Manager, .Stat Core

BIS, FMR

Cloud ready
#OpenSource requires mastering DevSecOps

Fast delivery of quality features requires test automation at scale

1.09 release/month on average in 2023

Source: .Stat Suite delivery dashboard, Nov 2023
Automated Functional tests

60% coverage – target 70%
in Unit Testing automation
+ Extensive test framework for
SDMX-RI europa
Integration testing with
End2End with GUI testing
Shared functional testing
environment between members
Member-specific testing scenarios

Automated Performance tests

K6 Framework
Smoke Testing
Load Testing
Stress Testing
Soak Testing

Automated Security tests

Extensive vulnerability checking
Static app. security testing (SAST)
Dynamic app. security testing (DAST)
Software Composition Analysis (SCA)
(Dependency Scanning)
Container & Docker image scanning
Regular security audits by members
Regular market reviews to pick best-in-class tools

#OpenSource requires mastering DevSecOps

Focus on accessibility
Ethical and moral obligation
Legal compliance
Business benefits

Main challenges
Balancing test coverage and cost
Continuously reduce marginal cost of testing
Continuously measure and improve
Enable wider cooperative testing

Using AI for Software QA
Generate test data
Generate scenarios for automated testing
Scan code for vulnerabilities

Source: non functional topics– SIS-CC workshop, March 2023
DevSecOps determines Cloud Readiness and multiplicity of deployment strategies

Full-cloud (AWS, Azure or GCP), Hybrid, On-premise

Source: FAO .Stat Suite infrastructure and GitOps– SIS-CC workshop, March 2023
#OpenSource #OpenKnowledge #Community Ingredients
Support your experts to become good trainers, coaches

#OpenKnowledge
START BUILDING KNOWLEDGE

Free online training and resources to support customers use the .Stat Suite and Data modelling in SDMX for data producers

What do you want to learn today?

Learners from across the globe.

More to come...
Develop tools to structure your Data governance
#OpenKnowledge
#SDMX
SDMX provides the tools (conceptual and technical) to operate your governance and progressively harmonise information models.
SDMX backbone brings commoditised SDMX storage, SDMX API-centricity, and SDMX semantics as part of the “Data infrastructure as a Platform”.

SDMX governance as cornerstone of data governance, enabling a progressive harmonisation of data semantics across domains and the data lifecycle.

Each business domain to build their data process and products in relative autonomy, drawing on common information models and technical building blocks.
Establish a systematic User Research
#OpenKnowledge
#CoInnovation
#OpenSource requires mastering User Research techniques

Usability: “A measure of how well a specific user in a specific context can use a product/design to achieve a defined goal effectively, efficiently, and satisfactorily.”

Source: What is Usability - The Ultimate Guide
The User Research Task Force (UTF)

UTF to draw insights from across (external and internal) user bases of all participating organisations.

- Identify and organise joint user research activities, opportunities and outcomes
- Set measure of success for the .Stat Suite, identify best opportunities for improvement.
- Build a common body of knowledge, tools and techniques that each organisation can draw on

Source: ModernStats 2023 - How a strong community and a User-Centered Design approach is key to .Stat Suite product excellence
Experiment opportunities for #Colinnovation
Build coalitions for #Colinvestment
#SDMX #AI
The Data Lifecycle: will be transformed with

SOURCE & COLLECT (META)DATA

PROCESS & ANALYSE (META)DATA

DISSEMINATE DATA & ENGAGE

MANAGE (META)DATA

MANAGE DATA PROJECT & ALGORITHMS
How many booster doses of covid vaccines administered in New Caledonia in 2023?

The Pacific Data Hub - stats.pacificdata.org
Develop a natural language capable Data Explorer (I)

Proof of Concept ongoing, mixing Generative AI & SolR techniques.

List of all topics and dimensions returned (SolR index wrapping sdmx structures)

List of additional reformulations (openAI generation of synonyms and related terms)

List of most relevant topics and dimensions (openAI extraction of topics and dimensions)

List of dataflows SolR query (generated reformulations, extracted topics/dimensions)

List of dimension values (openAI extraction of dimension values)

Source: courtesy of A. Benedetti, SEASE.
The StatsBot is within reach…we could develop it as an open and reusable asset.

The StatsBot flow drawing on SDMX APIs

The StatsBot stack: Azure Open AI, Langchain, Python

Source: courtesy of M. Garge & S. Polke, e-Zest.
# CoInvestment # CoInnovation opportunities

- Develop a Natural language capable Data Explorer (I)
- Develop the universal StatsBot to access any SDMX source (II)
- Develop a “prompt engineering in an SDMX context” practice
- Develop the specialised, “SDMX-literate” open source LLM
#OpenSource #Community Ingredients

THE RECIPE: FORMULATE A (PUBLIC) STRATEGY

FULL DATA LIFECYCLE

COMPONENT-BASED ARCHITECTURE

COMMUNITY DRIVEN

STRETCHING THE VALUE
OPEN KNOWLEDGE & CO-INNOVATION

OPEN SOURCE DELIVERY
Reference resources

**SIS-CC Community**
- SIS-CC Community
- SIS-CC 2020-25 Strategy
- Why SDMX Matters

**SDMX.IO Ecosystem**
- The SDMX.IO ecosystem

**SDMX Community**
- SDMX 2021-25 Strategy
- The business case for SDMX
- SDMX 3.0 packages under review
- SDMX Guideline: *Modelling statistical domains and exchange frameworks in SDMX*
- SDMX Guidelines: *A Reference Framework for SDMX Structural Metadata Governance*

**.Stat Suite, open source platform**
- .Stat Suite – High level product overview and flight planner
- .Stat Suite – Kanban Board
- .Stat Suite – Changelog
- .Stat Suite – Extensive documentation
- .Stat Suite – Code repository

**.Stat Academy, open knowledge platform**
- .Stat Academy – Learning Paths for Data Producer; Data Tooler
- .Stat Academy – Courses
- .Stat Academy – Webinars
- .Stat Academy – Become an instructor