UNITED NATIONS Working Paper No. 1
ECONOMIC COMMISSION FOR EUROPE

Workshop on Statistics for SDGs
29 - 30 March 2022, Geneva

Session 1: Data transmission including SDMX and automation

Automated data transmission across UNECE

Prepared by Lucy Gwilliam, Office for National Statistics (ONS)

Abstract

Automation can deliver benefits within data transmission, for example, more efficient processes and improved data quality. Two of the main things that can be used to help automate data transmission processes are application programming interfaces (APIs) and the use of coding. Both of these have their own benefits and would depend on the availability of data in a certain format. The UK Office for National Statistics (ONS), the Statistical Agency under President of the Republic of Tajikistan and the Open SDG platform are just three examples of how automation is used to facilitate the transmission of SDG data across the UNECE region.

I. WHY USE AUTOMATION FOR DATA TRANSMISSION?

1. Data transmission automation helps:

- speed up processes that would otherwise be very time-consuming if carried out manually
- reduce errors that may be introduced when manually entering data
- conserve resources for other projects that cannot be automated

2. It's important to consider the resource it takes to set up an automated process against the expected benefits including overall time spent and the improved quality that automation may bring.

II. TYPES OF AUTOMATION USED FOR DATA TRANSMISSION

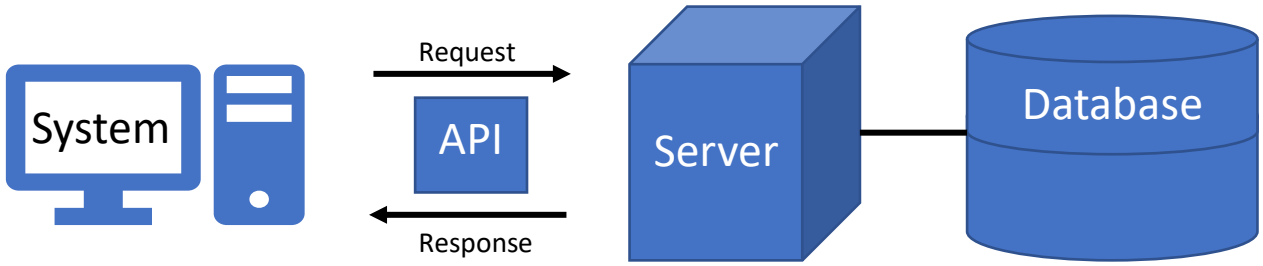
Use of application programming interfaces (APIs)

3. An API is a piece of software that allows two different applications to interact with each other in order to automate processes. For example, to pull data from one system into another. Figure 1 shows the high-level design of a REST (representational state transfer) API. RESTful APIs must conform to a set of criteria including information must be transferred in a standard form.

4. Depending on the format and structure in which the data is stored in the database and that is needed by the end user/system, some manipulation may still need to be done after the data has been received by the system.

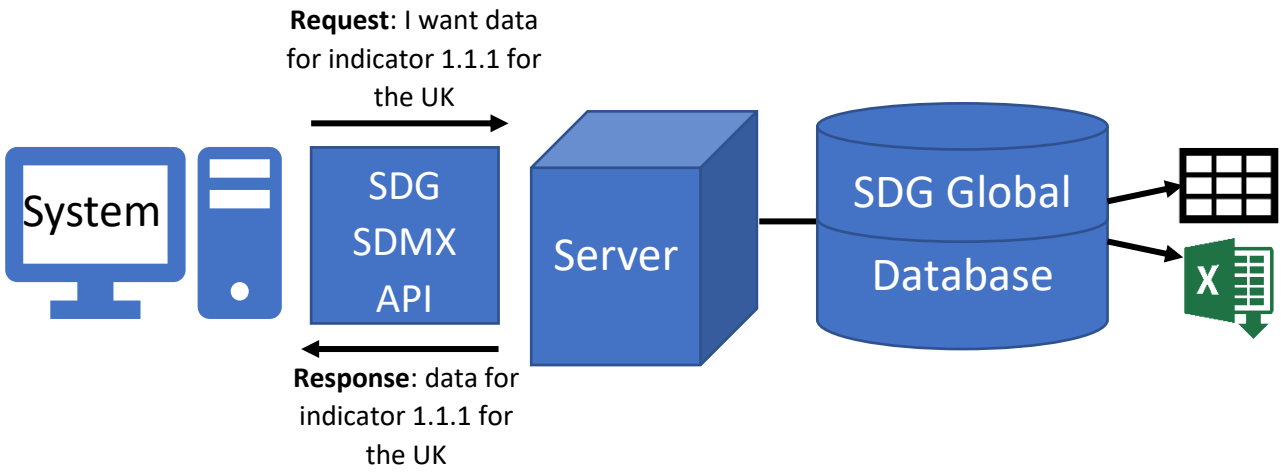
5. Even when data is stored in a database, an API isn't always provided and when one is, it may be limited or have a cost implication. Therefore, it is not also possible to use an API to access data.

Figure 1
Diagram showing the design of a REST API



6. As shown in Figure 2, an example of an API is the UN SDG SDMX API, which allows global SDG indicator data to be accessed by other systems in a machine-readable format, whereas the web interface for the SDG global database provides data in a human-readable format, in tables and Excel file downloads.

Figure 2
Diagram showing UN SDG API and Global Database interactions

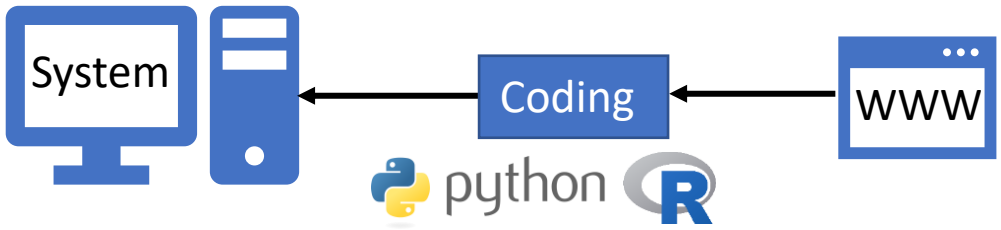


7. While it is not always possible to use an API to access data, as long as the data is available at a predictable URL, programming languages such as R or Python can be used to access the required data and manipulate it into the required format, from any open website. Figure 3 shows the high-level design of pulling data from an online source.

8. Even when the URL at which data the data is available isn't predictable, automation can be used with little human input.

9. When using the process, the coder should follow coding standards and best practice and use reusable functions. Otherwise, depending on the format and structure of the data being provided and the format and structure of the data required, this process can be very time consuming.

Figure 3
Diagram showing the design using coding to access online data



III. EXAMPLES OF AUTOMATED DATA TRANSMISSION ACROSS UNECE

UK example: using coding to access and prepare data in required format

10. The UK Office for National Statistics (ONS) uses R and Python to automate some parts of the data preparation needed to report the UK's Global SDG data. Figure 4 shows how automated data updates compare with the manual update process.

11. The manual data acquisition and preparation process (following identification of suitable source) for the UK involves:

1. Download source data file(s)
2. Copy required data from source data file(s) into dedicated indicator file
3. Carry out calculations using Excel formulas (if needed)
4. Restructure data in long data format i.e., one value per row
5. Quality assurance (QA) of data: different member of team recreates the data and compares
6. Upload data to Open SDG platform

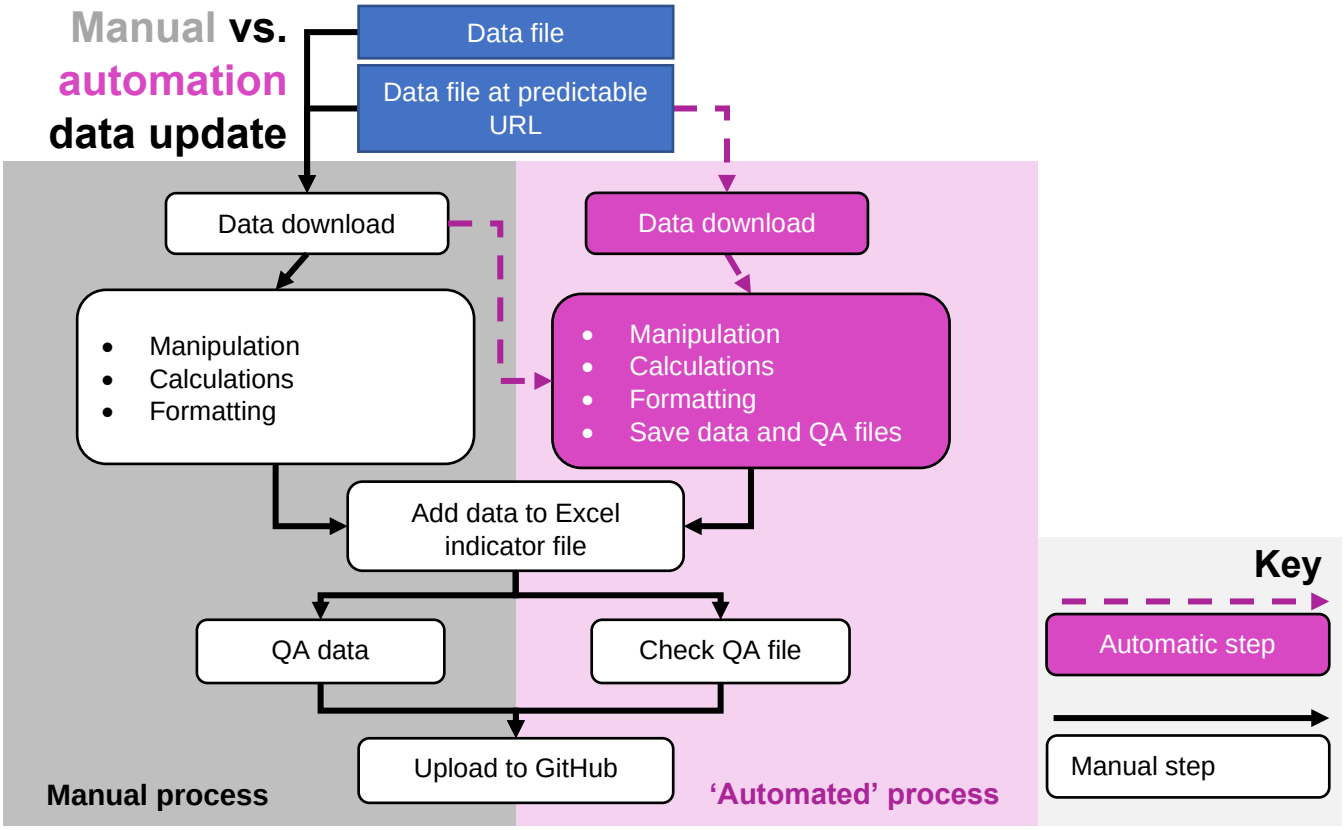
12. The manual process can be very time consuming, especially as more data points get introduced when adding disaggregations.

13. The UK team are automating different parts of the process to reduce the time spent on updating indicators. However, each indicator is considered on a case-by-case basis to ensure the automation of the indicator will save time in the long term. For example, if an indicator only has a small number of data points, it may be quicker to carry out the manual process each time, rather than spend time writing code to automate the process.

14. Depending on where the data is stored, sometimes the automation starts after the data has been manually downloaded. This is often the case when the data is not available at a predictable URL.

15. The UK's data acquisition process has been made much more efficient by reducing the amount of time spent on formatting data as well as reducing the amount of time spent producing files for quality assurance and carrying out quality assurance.

Figure 4
Diagram of UK automated data update process compared with the manual update process



16. Early figures suggest that manual updates often take about 9 hours including QA. Writing code to automate an indicator (including a QA document) takes on average about 5 hours, and running the code takes less than 15 minutes. So, there is a clear time improvement almost immediately. These timings don't account for time spent upskilling the team and conducting code reviews but as the code can be used in reused every year, the efficiency will increase over time.

17. The UK SDG team are sharing the code they use for the automation on GitHub[1] so that others can replicate the process if desired.

18. Next, the UK SDG team will be exploring how they can pull relevant data from a cross-government data service, which outputs data from an API.

Open SDG example: Using APIs for the transmission of data and metadata

19. Open SDG provides lots of options for using automated processes to populate a platform as well as for transmitting data from a platform. Figure 5 shows the functionality that Open SDG users can benefit from.

20. Open SDG platforms can:

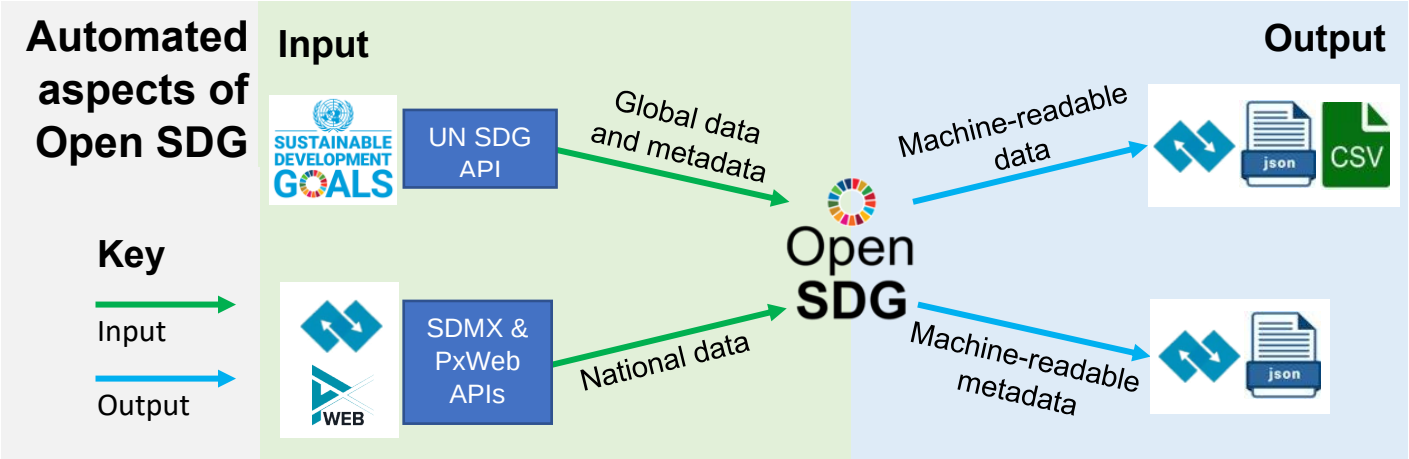
[1] https://github.com/ONSdigital/sdg_data_updates

- pull global data and metadata from UN SDG SDMX API
- pull national data from a country's SDMX API or PxWeb API

21. All Open SDG platforms output machine-readable data (CSV and JSON, and SDMX when configured) and metadata (JSON, and SDMX when configured) to predictable URLs to help with automated data transmission from platforms. For example, if a country wanted to use an Open SDG platform to make their data available in SDMX format, the output could then be used uploaded to the UN SDGs Data Lab.

22. The use of Open SDG often limits duplication of work by users due to the interoperability that it provides.

Figure 5
Diagram of Open SDG automated data transmission functionality



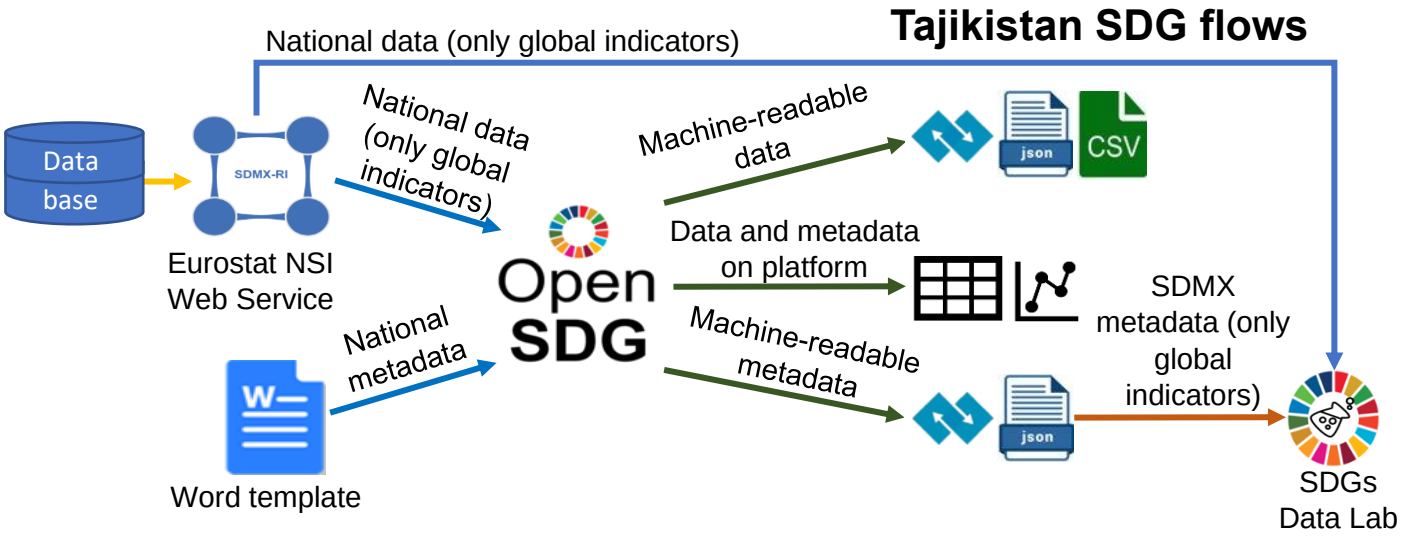
Tajikistan example: automated transmission throughout SDG flows

23. The Statistical Agency under President of the Republic of Tajikistan uses automated data transmission throughout their SDG data and metadata flows, as shown in Figure 6.

24. Tajikistan have configured an API over a database containing their SDG data, which allows their data to be easily accessed by other systems, including Open SDG and the SDGs Data Lab. This prevents them from having to upload the same files to many places and ensures consistency across all systems.

25. Following the pulling of data and the uploading of metadata into their Open SDG platform, Tajikistan's data and metadata is available in machine readable format. The Open SDG converts their Word metadata files (which complies with the SDG metadata structure template (MSD)) into SDMX-ML files, which can then be uploaded to the SDGs Data Lab. This automatic conversion saves Tajikistan from having to use a separate tool for converting their metadata into SDMX format.

Figure 6
Diagram of Tajikistan's SDG data and metadata flows



IV. ADDITIONAL RESOURCES

26. Road Map on Statistics for Sustainable Development Goals - Second Edition[2] (see chapter 4.3).

27. Kyrgyzstan's experience of implementing Statistical Data and Metadata eXchange (SDMX) – background paper[3] and recording[4].

28. Task Team on Data Transmission Wiki[5].

***

[2] https://unece.org/sites/default/files/2022-02/Road_Map_2_E_web.pdf
[3] https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.32/2020/mtg1/W_4_3_ENG_UNECE_SDMX_paper_Nazira.pdf
[4] https://youtu.be/-qQcsgR3sWg?t=2844
[5] https://statswiki.unece.org/display/SFSDG/Task+Team+on+Data+Transmission