OPEN DATA TUTORIAL

Opening and promoting use of budget data

Identify the source and format

SHCP
SECRETARÍA DE HACIENDA Y CRÉDITO PÚBLICO

TRANSPARENCIA PRESUPUESTARIA
Observatorio del gasto

GIFT
GLOBAL INITIATIVE for FISCAL TRANSPARENCY

Identify the source and format

Once the data to be published has been prioritized according to its potential users and their information demands, it is vital to locate the origin and format of the information.

Why does the origin and format of the data matter? Although a recent emphasis has been made in data being open by default, most of the times **the information systems containing the data to be opened were not created having its publication and/or reuse in mind, but to display the daily operations of the government instead**. Thus, working on its openness will involve analysis, debugging and creation of new processes to ensure its sustainability.

Therefore it is worth answering the following questions:

Is the data in **printed format only**?
Is all the data to be published in the **same place**?
How many people are responsible for **generating**, **concentrating** and/or **publishing** the data?
In what **format** is it: PDF, Word, Excel, CSV or another?
Does the data come from an **information system**?
If so, how can we **access** it?
Is it all in the same system or is it necessary **to join data from various technological tools**?

Considering these questions helps establish times for the opening process; identify the areas involved; analyze the quality of the information; as well as defining the challenges and opportunities ahead.

TIP
Make an operative calendar in a collaborative fashion with the areas involved, to define specific dates and actions towards the publication of the information.

Some relevant considerations to identify the source and format of the data are:

1. Identify the actors involved

Spending information may be scattered amongst areas and even government agencies. Nonetheless, it is important to take into account that the more dispersed the information is and the more people are responsible for it, the more will the process of openness depend on the existing cooperation capacities.

Conversely, a set of data that is in a single place with less number of co-responsible areas, can become the spearhead of the opening process, especially if you have already identified public interest on it.

Likewise, consider that the areas responsible for the information may not be familiar with the concepts of data openness and at the same time, the people in charge of the openness may not be very familiar with the contents and structure of the data sets that are being published; so, it is essential to start a continuous two-way learning process, since the usability and quality of the resulting data set and its data dictionary rely on it.

2. Data format

In order to allow the information to be used effectively, it is necessary to consider how malleable or rigid the data is, whether it is ready for processing or if it can be consumed in whole and/or in parts. This will depend, at first glance, on whether the data is unstructured or structured, and from there, you will be able to determine the possibilities of use.

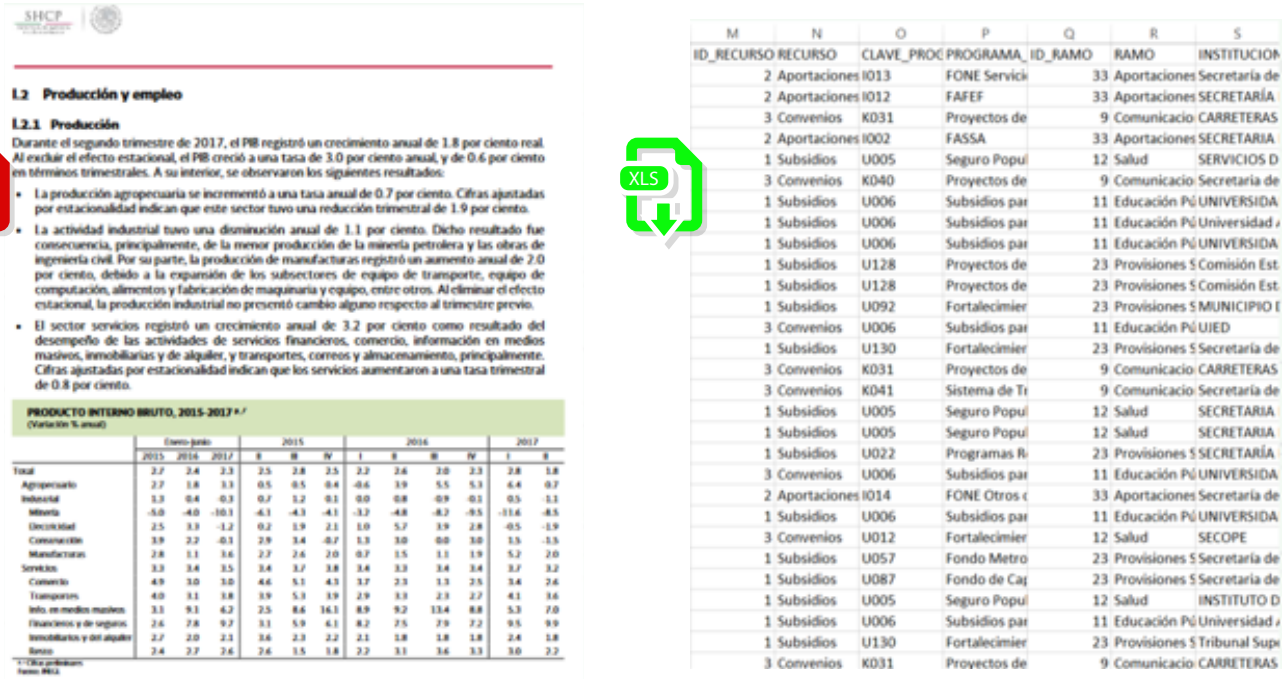
> **Unstructured data** refers to disorganized information, that is, a conglomerate of information from which specific attributes cannot be extracted in an agile and/or automated way.

These data may be public or not, but it is characterized by its storing, for example: information obtained directly from beneficiaries of a service or product derived from official interactions between public servants is often in printed format; information digitized from printed or photographic material is stored as digital images (PDF, JPG, PNG); and reports can be found as text editor files or slide presentations (DOCX, PPTX), among others.

Because of our daily interaction with this formats, we know that it is not possible to obtain sums or any mathematical operation from the information as it is more complicated to process it, so, it will be necessary to transform the data and fit it into a structured format.

In contrast, **structured data** refers to information readable by machines (via open sourced or licensed software), since it has an easily recognizable arrangement. For example, if the structure of the information consists of rows and columns, the spreadsheet or linear display formats allow its agile processing (XLSX, CSV, etc.); while for tree structured-information, the most common formats for information hierarchy are JSON or XML.

If the data that will be published is in a structured format, the openness process can be speeded up.



Example of structured vs unstructured data. On the left, data of the Gross Domestic Product that is public but difficult to process; on the right side, data on transfers of federal resources to sub-national governments in a structured format in rows and columns (XLXS and CSV) for agile processing.

3. Link between computer systems

When the information is already structured in storage systems, it is advisable to proceed to diagnose its quality and know its contents; design a proposal for the public release of the information (in case the format and structure are not usable) and perform tests to finally publish it.

In this situation, it is convenient to aspire to the highest levels of openness from the beginning of the process, in order to have an advantage in its publication. To learn more about the levels of data openness, review the video and reference document "What is open data?".

With these considerations on the identification of the source and format of the data, **the next step would be to start its debugging and cleaning**.

Check the next review document to learn about this process.

References

Kyocera. Difference between structured and unstructured data. Available at: https://smarterworkspaces.kyocera.es/blog/diferencia-datos-estructurados-no-estructurados/

School of Data. Know the format of your data. Available at: https://es.schoolofdata.org/2013/10/31/conoce-el-formato-de-tus-datos/

SHCP
SECRETARÍA DE HACIENDA Y CRÉDITO PÚBLICO

TRANSPARENCIA PRESUPUESTARIA
Observatorio del gasto

GIFT
GLOBAL INITIATIVE FOR FISCAL TRANSPARENCY