Organisation for Economic Co-operation and Development

DAF/COMP/WP3/WD(2022)28

Unclassified

English - Or. English

3 November 2022

DIRECTORATE FOR FINANCIAL AND ENTERPRISE AFFAIRS
COMPETITION COMMITTEE

Working Party No. 3 on Co-operation and Enforcement

Data Screening Tools for Competition Investigations – Note by Colombia

28 November 2022

This document reproduces a written contribution from Colombia submitted for Item 3 of the 136th OECD Working Party 3 meeting on 28 November 2022.

More documents related to this discussion can be found at www.oecd.org/daf/competition/data-screening-tools-for-competition-investigations.htm

Ms Despina PACHNOU
[Email: Despina.PACHNOU@oecd.org]

JT03506562

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Colombia

1. Introduction

1. This is a document about the development and implementation of digital data screens by the Colombian Competition Authority. The document is about two digital screening tools that were developed by the Competition Authority's data analytics team. It explains how the tools were originated and their development processes. It also describes their uses, and how they can inform competition assessments and potential investigations. Both analytics data tools show empirical analysis of two methodologies that allow the Authority using public information to evaluate systematically, repeatedly, and verifiable transactions which are part of some markets, and in this way enhance the functions of inspection, surveillance, and control in more markets in Colombia.

2. This contribution emphasizes the relevance of empirical methods like data screening tools for the application of a systematic review of data. This is the core argument of the document. The SIC is improving its digital screening tools to harness existing data and optimize its research methods. Applying a systematic collection and review of data is of great use for the diagnosis of potential distortions in a vast number of markets that would otherwise remain non scrutinized. Through the development of digital screening tools, the SIC aims at a wider approach to the detection of potential distortions to competition in different markets to increase the effectiveness of its functions.

3. To describe each of the tools, this document provides a non-technical overview of the logic the teams followed for their development, which includes an account of the extent of the interaction between the developers and data scientists and the team working on investigations, and an identification of the specific needs acknowledged and translated into solutions. The flow of ideas exchanged between engineers, economists, lawyers, and the forensics unit has been crucial for the creation of tools that can efficiently support the investigative processes conducted by the team at the Competition Authority. Combining the worldviews of developers and investigators is key to guaranteeing that the tools come afloat, and time and resources end up well utilized.

4. The description also includes a review of both tools' data treatment cycles. First, we will review the data used for screening purposes and the rationality of its choosing; we also explain the techniques used to gather such data and the different sources' relevance vis-à-vis testing hypotheses that are relevant for competition assessments.

2. Harnessing Access to Public Data for Competition Protection

5. In recent years, technologies have advanced significantly, and large sets of datasets tracking, and compiling market transactions have become ever-more available. We are now more familiar with "drowning in information" as Rutherford D. Roger suggested, and so, we are starving for knowledge. For competition authorities worldwide this means, knowledge of how markets operate in particular contexts. The Superintendence of Industry and Commerce –henceforth, SIC- has identified the need to harness the increasing data available and has developed tools that would aid investigators to assess relevant data systematically, thus increasing the rate of detection of anticompetitive disruptions to markets, whether they come from regulators or from market agents themselves.

6. The SIC has focused on enhancing the detection functions for industries and markets where data is not only publicly available but also offered in different forms. For example, one of the tools that will be described focuses on procurement markets, the other on the assessment of regulatory projects that are published by regulators on their own webpages. Lately, Colombia has been investing in improving the quality of public procurement data as it is a relevant condition for increasing transparency in the public procurement system. Transparency contributes to create a level playing field and allows stakeholders to effectively participate and monitor procurement market dynamics. This also means that it may increase accountability. The Colombian public procurement system runs through a platform called SECOP II. The platform has been evolving, from existing only for contract publishing purposes to becoming a transactional platform (e-procurement). The first version of SECOP would only include information on awards and notices. Whereas the second version would include "procurement documents and allow electronic communication at all stages of the procurement cycle and allow electronic submission of tenders"[1].

7. In Colombia, access to public information is a constitutional right. One way of guaranteeing the exercise of this constitutional right is by providing open data portals. The official Open Data portal provides "primary or unprocessed data, available in standard formats and interoperable to facilitate access and use, under the custody of public and private entities in charge of exercising public functions. This data is available for any citizen, free of any restrictions, to allow their retrieval, use, and creation of services derived from them."[2] Under the Transparency Law—Law 1712 of 2014--, public entities are compelled to publish open data, except for classified and reserved information.

8. Information on public procurement processes is open to public review, as mandated by the law. This means this data should be published on a web page, free to access and use, and available in formats that enable their utilization, reuse, and operation without any restrictions as they are displayed under an open license.

9. The data that is available today, provides a record of government data for the purposes of research, applications, and visualizations. It stores data from 1100 entities. Open data portal stores datasets that unify public procurement information from all these state entities. These datasets, available at the web portal, compile information on procurement processes, including data on the type of process, the value, the object, participants, and other. Colombia Compra Eficiente (CCE, henceforth) updates the information on the SECOP Open Data portal on a daily basis. CCE performs a transformation of the information to eliminate characters that may generate errors. The data contained in the Open Data Portal are in formats that allow automatic computer processing, in addition to allowing it to be processed and displayed on the platform or downloaded in different formats such as CSV, Excel, JSON, RDF, RSS, TSV, XML[3]. As of September 2022, the unified data set for Colombian public procurement compiles data on integrated

[1] OECD (2016), Towards Efficient Public Procurement in Colombia: Making the Difference. OECD Public Governance Reviews, OECD Publishing, Paris. Retrieved from: https://read.oecd-ilibrary.org/governance/towards-efficient-public-procurement-in-colombia_9789264252103-en#page4

[2] Guía para el uso y aprovechamiento de Datos Abiertos en Colombia https://herramientas.datos.gov.co/sites/default/files/2021-08/Guia%20de%20Datos%20Abiertos%20de%20Colombia.pdf

[3] Manual para el uso de Datos Abiertos del SECOP. Retrieved from: https://www.colombiacompra.gov.co/sites/cce_public/files/files_2020/manual_de_datos_abiertos_actualizado.pdf

information on the public procurement processes that have been registered on the SECOP I and II platforms, which have ended with a contract.

10. It is also very important to note that CCE provides access to this dataset via API. Socrata's Open Data API provides programming-level access to this dataset, including the ability to filter, query and aggregate data. Which helps greatly in the process of collection and treatment of the relevant data for the tools that the authority is developing.

11. On the other hand, the SIC has also developed a tool that reviews the regulatory projects that are available for public access and that are published by regulators.

12. Having data for their significance for the overall economy and the availability of transactional data. As mentioned in the call for contributions to the present discussion, data analytic tools are usually applied ex-ante and can guide investigators into flagging hotspots of irregular or suspicious behaviour from the actors in the markets.

3. Types of Digital Screening Tools: A View into Sherlock and Inspector

13. This section builds on Colombia's contribution to the Stanford Computational Antitrust Implementation Survey[4]. We provide a brief overview of the uses of the screening tools and includes information relevant for the discussion at the roundtable, meaning how does the authority compile and treat the data, and what techniques has the authority used to gather data from public sources.

3.1. Sherlock

14. Sherlock is a project that seeks to support the SIC's investigators in the identification of signs or patterns that suggest potential anticompetitive behaviours with the data available from public procurement processes.

15. The first stage of development of the project consisted in the creation of a tool that could facilitate the access of investigators to public data available on the web. The purpose is to aid case handlers in the identification of procurement processes with similar characteristics to the one (s) of interest; the bidders within these processes; the contracts awarded by the same entity to the same contractors; relationships or nexus between legal representatives and relevant participants, among others. This first stage of the tool comprised capturing the data, organizing it, and providing simple descriptive analytics to the investigators. The software cleans, structures, and transforms the collected data and then loads it into a database within the SIC. It also allows for visualization in dashboards with filters, graphs, and comparisons.

16. Sherlock's second stage involves the automation of the search of the abovementioned signs and patterns within the bulk of public procurement data available online. The data collection is performed after the definition of variables relevant for investigators and identifiable via CCE's API called SOCRATA. Then, the data is cleaned and standardized by string type, numeric type, or URL type. Also, variables are standardized for SECOP I, SECOP II Contracting Processes and SECOP II Electronic Contracts.

[4] Schrepel, Thibault and Groza, Teodora, The Adoption of Computational Antitrust by Agencies: 2021 Report (June 21, 2022). 2 Stanford Computational Antitrust, 78 (2022), VU University Amsterdam Legal Studies Paper Series, Available at SSRN: https://ssrn.com/abstract=4142225

17. This second stage will leverage historical data available in Open Datasets and will analyse and further diagnose information based on indicators already identified by international organizations like the OECD (Organization for Economic Cooperation and Development). As of today, the tool works on generating alerts when these three conditions are found in the data:

- When the awarded budget is equal to or less than 5% of the official budget
- When the same bidder is awarded three or more times in a selection process within the same State Entity (concurrence), identifying within the objects key words in accordance with the economic sectors
- When there is only one bidder identified in the process

18. The project will have the capacity of providing informed red flags. It would automatically categorize a particular piece of data into one of the indicators of collusion widely known. This feature will provide predictions useful for the investigators. This would provide a more dynamic computational antitrust tools flow of valuable information for the researchers, in addition to greater descriptive insights from the data.

3.2. Inspector

19. The highly qualified human talent within the Competition Advocacy Group devotes a significant amount of time to purely operational activities. For example, gathering information on new regulatory projects issued by regulatory entities to analyse later whether such projects may impact free competition. The application developed by the SIC solves the problem by monitoring the regulatory projects published on the websites of state regulators. If changes are detected in the published regulations, the system automatically sends an e-mail to the interested party previously defined to inform about the news.

20. The objective of Inspector is to streamline and facilitate the review process of regulatory projects issued by regulatory entities by the members of the Competition Advocacy Working Group of the Deputy Superintendence for Competition Protection. The aforementioned, in order to promote the adequate fulfilment of the informal function of competition advocacy and the adequate use of the GAC's human resources.

3.3. Interaction - Functionality

21. The tool works automatically from Monday to Friday at 7 am, where it reviews the publication pages of the monitored entities in search of new regulatory projects, performs in the entire value chain, identifying possible new regulatory projects and issuing a consideration on the impact of such regulation on free competition in the market corresponding to the monitored entity. At the end of the process of search, identification, and consideration of regulatory projects of an entity, if it finds one or more documents that meet the conditions to be considered as a regulatory project, it sends an email to the responsible official of the SIC notifying the identification of new projects, the link to download the document and the consideration of the project filtering model described in the value chain. Since Inspector works only with regulatory projects that are published as additional files to the web page, the file concerning the project is downloaded to the server where it is executed and stored temporarily on the server (once the Inspector Day is over, this information is deleted).

22. The tool has been developed in three stages. In the first stage, a proof of concept was carried out with a minimum viable product, where the aim was to verify whether it was possible to carry out the monitoring of regulatory projects for a reduced number of entities.

In conjunction with the GAC, 7 pilot entities were identified, for which the development of the tool for the capture and notification of new regulatory projects of the pilot entities was carried out.

23. In the second stage of the tool, the regulatory project filtering model was developed, consisting of the vectorization model and the document classification model. For the development of these models it was necessary to collect the information available on concepts previously issued by the GAC, generate the document base that would be used for training, develop the models, test the different models and validate the different models according to which had the best performance.

24. Finally, we are in the third stage of development of the tool, where the spectrum of entities monitored by Inspector is being expanded from 7 to the 78 entities that make up the full universe identified by the GAC. In order to cover the 78 entities without generating a specific development for each one of them, an identification of the characteristics of the different web pages was carried out, which allowed us to identify patterns, classify and include each one of the entities in 10 cases, which will need to be developed independently. At this moment, 7 of the 10 cases have been developed, and Inspector is monitoring 62 regulatory entities. The cases were then prioritized according to the relevance of the entities they comprise for the GAC, which allowed the different cases to be developed sequentially. Additionally, the filtering model developed in the second stage of the tool is integrated with the cases that are being tested.