PROCESS BRIEF

Data Reconciliation: Process, Standards, and Lessons

A CASE STUDY OF INTEGRATING SUB-NATIONAL PRIVATE SECTOR DATA INTO COLOMBIA'S NATIONAL STATISTICS

RESEARCH LEADS: FREDY RODRIGUEZ AND PHILIPP SCHÖNROCK, CENTRO DE PENSAMIENTO ESTRATÉGICO INTERNACIONAL – CEPEI

EDITING SUPPORT: MELIKA EDQUIST, JESSICA ESPEY, JAY NEUNER, AND LESLIE RAE, SUSTAINABLE DEVELOPMENT SOLUTIONS NETWORK – SDSN

JULY 2018

TRENDS
Thematic Research Network on Data and Statistics

Cepei
Knowledge factory for the world we deserve

ABOUT SDSN TRENDS

SDSN TReNDS is an expert thematic network of the UN Sustainable Development Solutions Network. It is made up of 20-plus leading academics, policymakers, and practitioners in the field of data and statistics for sustainable development. This multidisciplinary group of high-level experts works together to catalyze learning and investment in data for sustainable development.

Specifically, TReNDS:

Helps strengthen the data ecosystem: Generating and curating ideas on how to strengthen local, national, and global statistical systems and governance to harness the data revolution to achieve the Sustainable Development Goals, e.g. through short briefings, blogs, and the Local Data Action Solutions Initiative.

Improves learning on data sharing: Analyzing policies, conditions, and investments that enable data sharing success to generate more frequent and disaggregated data, e.g. through the compilation of in-depth case studies, short briefings, and by documenting learning from the Colombia Data Reconciliation Solutions Initiative.

Encourages standard setting amongst new data actors: Incubating technical coalitions that aim to establish practice standards for new data approaches, e.g. supporting the POPGRID initiative to help establish standards for high-resolution population estimation.

Informs investment in the emerging data opportunities: Supporting policy-relevant research for advocacy, e.g. the Case for Investing in Data briefing series with the Global Partnership for Sustainable Development Data.

CONCEPT

THE CHALLENGE

The Sustainable Development Goals (SDGs) require a huge amount of data to measure progress and achieve the goals. In many countries around the world this is an acute challenge due to financial, institutional, and capacity constraints. This is prompting the development of innovative ways for governments, and specifically National Statistical Offices (NSOs), to generate or curate additional data. Many are looking to the private sector as potential partners to share a wide range of environmental, commercial, and social data that can help monitor attainment of the SDGs. Government policymakers and official statisticians have expressed concern, however, about the feasibility of using third-party data. Worries include data existing in different formats and therefore being irreconcilable, and the feasibility and security of sharing data across public and private institutions.

In 2016, the Colombian National Administrative Department of Statistics (known as DANE) conducted a data gap analysis of the data currently available to monitor Colombia's progress on the SDGs. It found that within the Colombian government, there was information available for 54 percent of the global SDG indicators, partial information for 30 percent, and no data (and/or no methodological clarity on how to measure the indicator) for 16 percent (Departamento Administrativo Nacional de Estadística - DANE, n.d.).

THE SOLUTION

In response to this analysis, the Sustainable Development Solutions Network (SDSN) and Centro de Pensamiento Estratégico Internacional (Cepei) decided to collaborate to test new potential methods for collecting missing SDG data from the private sector. The aim was to transmit private sector data addressing a deficiency defined by the data gap analysis to DANE, the national institution in Colombia responsible for measuring SDG indicators. The project explored the best governance arrangements to facilitate data exchange, as well as the technical challenges associated with data reconciliation. The group hypothesized that a technological data reconciliation platform would support the data sharing process. The Bogotá Chamber of Commerce (BCoC) was selected as a partner for the project based on its willingness to share a wide range of highly useful datasets and strong management-level commitment to the SDGs. These efforts have included: a commitment to the SDGs by the President of BCoC in 2015 (on the sidelines of the Summit for the Adoption of the 2030 Agenda) (Cámara de Comercio de Bogotá 2018); BCoC becoming a member of the Global Partnership for Sustainable Development Data (GPSDD), also in 2015 (Global Partnership for Sustainable Development Data 2016); programs on SDG implementation in Bogotá and the surrounding metropolitan area, currently in development by BCoC (The World Bank 2017); and the alignment of the outcomes of BCoC's 2016 sustainable report with the SDGs (Cámara de Comercio de Bogotá 2016).

In Colombia, all organizations (private sector companies, nongovernmental organizations, and foundations) must register with their city's Chamber of Commerce and are required to update their registrations on a yearly basis. The BCoC has key information on more than 650,000 companies or "productive units" collected through the Mercantile Register (Cámara de Comercio de Bogotá 2017). Its databases have more than 2.6 million records of business performance. The breadth of the datasets available makes Bogotá's private-public data ecosystem an ideal test case, with partners able to examine the feasibility of sharing diverse types of data. BCoC is also a strategic partner, as other Chambers of Commerce could replicate the collaboration across the country.

DESIRED OUTCOMES OF THE PROJECT

- Identify data from a private entity that would help fill gaps in the national government's "official" SDG indicator set
- Assess the utility of using data from a sub-national source for national monitoring
- Identify appropriate governance arrangements to facilitate data sharing across parties
- Identify the technical means by which to share data securely and ensure comparable data standards
- Build capacity among project partners

PROCESS

TECHNICAL DESIGN

Integrating non-traditional data sources, such as data generated by the private sector, into the formal statistical community and official national statistics is still a relatively unproven process with a range of technical challenges. To take on this challenge, one of the original project partners alongside SDSN, Cepei, and DANE was Amida Technology Solutions (hereafter "Amida"), a software developer with expertise in developing technical platforms for data reconciliation. The objective of the collaboration was to build on Amida's experience automating health data reconciliation in the United States of America (Amida Technology Solutions, n.d.) and determine if such a process might be feasible for private enterprise data coming from the Bogotá Chamber of Commerce and being transmitted to DANE, the national statistical office. While the scale, scope, and policy challenges associated with reconciling data from hospitals, labs, pharmacies, doctors, and patients exceeded the complexity of the Colombian project, the project partners believed there are core elements of any data reconciliation project that could provide important lessons for this technical work.

Following early discussions with representatives from DANE, the BCoC, Cepei, and Amida, it was agreed that Amida would develop and pilot the use of an open-source data platform that could ultimately be employed by multiple countries' NSOs or line ministries, automatically gathering non-traditional or sub-national data streams to complement their reporting platforms regardless of their current capacity levels.

Beyond the technical design of such a tool, practices from different actors were studied in order to project the potential of scaling up a prototype that relies not only on technological matters, but also seeks answers about sub-national information gaps, governance structures for these types of initiatives, and the possibility of data sharing if the quality of data complies with expectations.

SDSN and Cepei began work with BCoC and DANE to identify institutional barriers to data sharing and ensure a supportive policy environment for the data exchange to take place. SDSN and Cepei documented project learning and, subject to successful completion, aimed to replicate the effort in other cities across Colombia.

INSTITUTIONAL READINESS

The project began in early 2016. Cepei, as the local implementing partner, initially faced challenges related to institutional readiness and acceptance of the initiative inside the BCoC. To overcome this institutional skepticism, Cepei was required to build relationships with key members of BCoC and convene events to spark interest. Cepei also drew on preparatory work conducted in 2015, including a series of high-level events in Bogotá focused on data sharing and public-private collaboration.

Although the process of building relationships and garnering institutional trust is not quantifiable or measurable in numeric terms, there are four distinct stages of work that were accomplished before the project even began. This initial work laid the foundation for the project before it formally commenced, and helped ensure a successful agreement between DANE and BCoC:

Raising Awareness (2015)

Objective

To convene key stakeholders at global and local events to present the relevance of the data revolution and the role of development actors (including the private sector and NSOs).

Activities

- United Nations Statistical Commission – Side-events with DANE (Centro de Pensamiento Estratégico Internacional 2015a). Cepei organized two events to discuss the data revolution in Latin America and the Caribbean and the need for a multi-stakeholder approach to measure sustainable development.
- Cartagena Data Festival – DANE as key partner, Cepei as co-organizer of the festival (Cartagena Data Festival 2015). This event was hosted in Cartagena, Colombia and brought together organizations from different sectors in order to promote the data revolution and find means of collaboration.
- Initial meetings with the BCoC, considered a high-level overture within the Chamber of Commerce to lay the groundwork on the importance of data for sustainable development.

Linking Global and National Levels: Part One (2015)

Objective

To demonstrate the importance of the Sustainable Development Goals (taking into account Colombia's pivotal role in the inception of the SDGs (Sustainable Development Solutions Network 2016)) to national actors and the benefits of working with international partners.

Activities

- Cepei formed institutional linkages with the GPSDD and SDSN.
- Cepei organized several working meetings with both SDSN and GPSDD representatives as keynote speakers. All meetings were held at the Chamber of Commerce and the Chamber's CEO and/or vice presidents, as well as DANE's senior leadership, attended the meetings (Centro de Pensamiento Estratégico Internacional 2015b).
- Several workshops were organized with DANE and the National SDG Commission's technical team (United Nations Development Programme 2015).
- Both the Colombian government (through DANE) and the BCoC became members of the GPSDD in September/October 2015 (Global Partnership for Sustainable Development Data 2017).

Building Trust and General Bilateral Agreements (2016)

Objective

To formalize the partnership among actors and foster initiatives around data for the SDGs, Cepei advocated to have Memorandums of Understanding (MoUs) with BCoC and DANE to create an enabling environment towards a joint collaboration for the SDGs.

Activities

- In December 2015, Cepei and BCoC agreed to collaborate on SDG projects and activities beyond data (Centro de Pensamiento Estratégico Internacional 2015c). In April 2016, Cepei and DANE agreed to collaborate to achieve better data for SDG monitoring (Centro de Pensamiento Estratégico Internacional 2016b).

Linking Global and National Levels: Part Two (2016)

Objective

To demonstrate the willingness and capacities available in Colombia to develop projects and initiatives on data for the SDGs.

Activities

- In March 2016, Cepei arranged for the Colombian delegation to attend the United Nations Statistical Commission. This delegation included the CEO of BCoC and a member of the board, the vice president of Telefónica Colombia, and the director of DANE. DANE and BCoC participated in a GPSDD members meeting and national workshop that Cepei organized (Centro de Pensamiento Estratégico Internacional 2016a).
- SDSN, DANE, BCoC, and Cepei initiated discussions for the data reconciliation project. In May 2016, the initiative was approved by all partners.
- Following the approval, Cepei and SDSN began to develop a more technical approach to the project with BCoC and relevant partners.

PROJECT IMPLEMENTATION

IDENTIFYING DATABASES

The BCoC commenced the project by conducting a general review of available databases that met the quality standards for data collection and processing stipulated by DANE. This review determined which databases were most relevant for the project. It illustrated significant gaps in quality in terms of collection methods or survey design. This analysis allowed BCoC and Cepei to identify the steps needed to ensure each database was in a useable and sharable standard. Based on the results of the analysis, BCoC proposed using the Mercantile Register database, which has the largest datasets with time series compiled by the Chamber. Once the database was selected, those responsible for collecting and processing the Mercantile Register organized follow-up meetings to define and evaluate the methodologies used, including a review of web versus paper data collection.

SELECTING INDICATORS

Cepei held several consultations and meetings with DANE to determine which indicators they most needed assistance with monitoring and for which BCoC might be able to provide data. This process enabled the partners to further define the project's scope and identify their needs at both the national and sub-national levels (specifically focused on Bogotá). DANE and Cepei started by analyzing the global indicators provided by the United Nations Inter-Agency and Expert Group on SDG Indicators (IAEG-SDGs), which were then reviewed by the BCoC. The partners determined that BCoC's research contained data from the private sector that was applicable to SDGs 1, 5, 7, 8, 9, 11, 12, 13, 16, and 17 (United Nations 2015).

Following BCoC's decision to make the Mercantile Register available for the project, Cepei reviewed the administrative record questionnaire. The team analyzed the variables, detailed questions, and the annexes in the questionnaire. It concluded that the register could provide the best quality data for SDGs 8 and 9, which are priority goals for Colombia. Cepei determined that the most useful data from the Mercantile Registry were for the six indicators in Figure 1.

8.10	Number of commercial bank branches (per 100,000 adults in cities with large populations, and by municipalities for those with less than 100,000 inhabitants)
8.10	Number of enterprises specializing in electronic deposits and payments (per 100,000 inhabitants in cities with large populations, and by municipalities for those with less than 100,000 inhabitants)
8.10	Number of bank correspondents (per 100,000 adults in cities with large populations, and by municipalities for those with less than 100,000 inhabitants)
9.2	Manufacturing employment as a proportion of total employment
9.3	Percentage of SMEs that have obtained a loan or a credit line
9.3	Percentage of SMEs exporters

Figure 1. Select indicators from the Sustainable Development Goals (United Nations Statistics Division 2017).

Additional consultations with BCoC highlighted other data quality issues, including the following:

- Although there are guidelines for filling out the questionnaire, within BCoC there is no documented methodology for data collection, processing, and dissemination of data obtained from the Mercantile Register.
- There are significant questions on the form that are not mandatory for all businesses to answer, leading to gaps or inconsistencies in the dataset; for instance, the number of employees can be entered using a zero or any value with any formatting, such as a value of "40.000.000" employees.
- While the current procedure requires businesses to register by March 31 of a given year, it is possible for companies to register throughout the year, allowing the database to have a constant dynamic data entry.

TECHNICAL DEVELOPMENT

In its inception, the project aimed to develop a tool that could solve the data sharing challenges between BCoC and DANE. This was based on the theory that better technology would make the data cleaning and standardization process much easier, which was one of the major barriers to institutional data sharing. However, within a few months of commencing the project it became clear that it was more a question of process rather than a lack of sufficient technology. As a result, the local partners concluded that a robust technology tool was not what was required to improve data sharing, but rather a simplified, open-source tool that would enable transparent sharing amongst partners.

Therefore, the project's focus shifted as local partners targeted institutional agreements, as well as redefining the kind of tool that might support the data sharing process. To achieve the objectives of the pilot, the partners agreed to work with a local partner who could provide a more responsive solution using the open-source statistical software R. Using an open-source tool would not only make this process easier for the institutions involved, but also make it easier for future partners to reuse the programming code to run calculations of the selected indicators. Importantly, the open-source code also allowed future users to identify quality issues in the database, such as atypical information and levels of completeness.

CAPACITY BUILDING

Throughout the project, a major objective was to build capacity among local actors within BCoC and DANE, as well as ensure that any successes were highly replicable in other contexts. The main capacities that allowed this pilot to succeed were:

1. R programming

Data analysis and programming teams provided training to staff in BCoC and DANE in order to reuse the algorithms and coding made in R in similar databases within BCoC.

2. Institutional best practices for data sharing

Administrative and legal units developed new procedures to make data sharing within other organizations feasible in the future.

3. Use of non-traditional data sources

The process enabled the NSO, in this case DANE, to understand how to better communicate their data needs and requests to organizations beyond national statistical systems. Furthermore, the data reconciliation pilot created a better understanding of expectations for both sides of the partnership to generate accurate information for the SDGs in Colombia.

LESSONS LEARNED

Though it was mainly conceived as a technology solution, this pilot demonstrated that there is a lack of institutional readiness to perform these types of projects, particularly as related to the administrative and legal units of an organization. Partners should bear this in mind in the initial planning phase, particularly when piloting this type of project. Main lessons learned from the project:

Building trust between partners

- Dialogue between the private sector and the NSO was an important element before initiating the technical aspects of the work.
- There are protocols within institutions that are not easily calculable in programmatic planning; e.g. what was perceived of as a simple request for a sample of a given database in fact required analysis by and approval from BCoC's legal unit.
- Raising awareness about the SDGs as a starting point for discussions with stakeholders is a crucial element.
- Although support from senior management inside the institutions is necessary for speed and practicality, it does not guarantee the smooth running of the project. Senior management is not necessarily aware of the constraints that staff are confronted with when it comes to data sharing.

Harmonization

- Databases need to have a dictionary of variables or metadata prior to using a data reconciliation tool.
- Partners must define clear definitions of concepts to avoid misinterpretations in the process.

Trade-offs between data access and confidentiality

- Partners need to make legal units aware of the statistical purposes of data without compromising confidentiality of the data.
- Administrative and legal units require more time than knowledge management or data analysis units to run analyses and feasibility studies regarding the data sharing process.

Technology tools are not always the answer

- To ensure high-quality, accurate information, BCoc needs to improve its collection, validation, and quality processes regarding the Mercantile Register prior to sending the data to DANE.

NEXT STEPS

All companies have a legal requirement to register and maintain updated data in their records. Given this requirement, the Mercantile Register and other data collected by other Chambers of Commerce in Colombia can also be involved in this process. Due to the positive outcomes from the first pilot exercise, Cepei has started to replicate the Bogotá exercise in other Colombian cities, such as Medellín and Cali, with the support of SDSN.

Cepei is also exploring regional scenarios to scale up the initiative among organizations leading the use of the Mercantile Register, such as the Association of Registers of Latin America and the Caribbean (Asorlac), that could become potential partners in monitoring and reviewing the SDGs at different levels.

REFERENCES

Amida Technology Solutions. n.d. “Health.”

Cámara de Comercio de Bogotá. 2016. “Quinto Informe de Sostenibilidad.”

Cámara de Comercio de Bogotá. 2017. “Bogotá - Región Cerró El 2016 Con Más de 694.000 Empresas y Establecimientos de Comercio Activos.” February 2017. https://www.ccb.org.co/Sala-de-prensa/Noticias-CCB/2017/Febrero/Bogota-Region-cerro-el-2016-con-mas-de-694.000-empresas-y-establecimientos-de-comercio-activos.

Cámara de Comercio de Bogotá. 2018. “Avances e Implementación de La Agenda 2030 En Colombia.” 2018. https://www.ccb.org.co/Sala-de-prensa/Noticias-Red-de-Embajadores/2018/Avances-e-implementacion-de-la-Agenda-2030-en-Colombia

Cartagena Data Festival. 2015. “Cartagena Data Festival.” April 20, 2015.

Centro de Pensamiento Estratégico Internacional. 2015a. “Enfoque Multidimensional Para Medir Los ODS.” March 5, 2015. http://cepei.org/eventos_cepei/46ava-sesion-de-la-comision-estadistica-de-nu/.

Centro de Pensamiento Estratégico Internacional. 2015b. “Ecosistema de Datos Para El Desarrollo Sostenible.” September 17, 2015. http://cepei.org/eventos_cepei/ecosistema-de-datos-para-el-desarrollo-sostenible-2/.

Centro de Pensamiento Estratégico Internacional. 2015c. “Primer Laboratorio de Datos Para El Desarrollo Sostenible.” December 10, 2015. http://cepei.org/eventos_cepei/primer-laboratorio-de-datos-para-el-desarrollo-sostenible/.

Centro de Pensamiento Estratégico Internacional. 2016a. “Colombia En La 47o Sesión de La Comisión Estadística de La ONU.” 2016. http://cepei.org/gobernanzas/colombia-en-la-47o-sesion-de-la-comision-estadistica-de-naciones-unidas/.

Centro de Pensamiento Estratégico Internacional. 2016b. “Convenio CEPEI-DANE Para La Revolución de Datos.” April 12, 2016. http://cepei.org/datos/convenio-de-asociacion-para-la-revolucion-de-datos-cepei-dane-2016-2/.

Departamento Administrativo Nacional de Estadística - DANE. n.d. “Primer Congreso Andino de Datos Para Los Objetivos de Desarrollo Sostenible.” http://www.dane.gov.co/files/images/eventos/ods/Memorias-Congreso-Andino-de-Datos.pdf.

Global Partnership for Sustainable Development Data. 2016. “Bogotá Chamber of Commerce - Cámara de Comercio de Bogotá.” 2016.

Global Partnership on Sustainable Development Data. 2017. “Partners.” 2017. http://www.data4sdgs.org/partner-listing.

Sustainable Development Solutions Network. 2016. “A Case Study of Colombia: Data Driving Action on the SDGs.” May 6, 2016.

The World Bank. 2017. “SDGs Targets Used to Focus Colombia’s Local, Global Goals [Blog Post].” August 23, 2017. http://www.worldbank.org/en/news/feature/2017/08/23/sdgs-targets-used-to-focus-colombias-local-global-goals.

United Nations. 2015. “Sustainable Development Goals.” 2015. https://sustainabledevelopment.un.org/sdgs.

United Nations Development Programme. 2015. “Ecosistema de Datos Para El Desarrollo Sostenible En Colombia.” October 26, 2015. http://www.co.undp.org/content/colombia/es/home/presscenter/articles/2015/10/26/ecosistema-de-datos-para-el-desarrollo-sostenible-en-colombia/.

United Nations Statistics Division. 2017. “SDG Indicators.” 2017. https://unstats.un.org/sdgs/indicators/indicators-list/.