IndiMaker - Open Data Linking Framework

Juan Santiago Preisegger, Alejandro Greco, Ariel Pasini(✉), Marcos Boracchia, and Patricia Pesado

Computer Science Research Institute LIDI (III-LIDI), Facultad de Informática, Universidad Nacional de La Plata, 50 y 120 La Plata, Buenos Aires, Argentina
{jspreisegger,apasini,marcosb,ppesado}@lidi.info.unlp.edu.ar

Abstract. Open data portals make a very important set of information available to the community. Those interested in a particular topic, retrieve data on the topic from different portals, but then, processing them together is difficult due to the different publication criteria used by each portal. To assist in this process, a tool called IndiMaker was developed, together with a framework that helps linking these files to users with little technical experience in data analysis. Within the framework, the tool allows applying different operations on the files, generating graphics in a dashboard, which makes information analysis easier. The framework was applied to the "environment" topic, in particular, to water and air quality and energy generation.

Keywords: Open data · Data linking · Software engineering · Open government · Environment

1 Introduction

A city thrives on the behavior of its citizens. Citizens, through different devices, are able to register more and more information about the activities they carry out. Making intelligent use of the information registered by government authorities to improve the lives of citizens is a great contribution to the community itself. But the contributions that can be achieved from the recorded data may come not only from the government – different agencies or individuals, who are capable of analyzing the information, processing it and proposing improvements, are also important factors in this cycle of city improvement. This difference is what makes a city a smart, sustainable and participatory [1].

To achieve citizen participation in this type of process, organizations make large volumes of open data available to their community, so that those interested in the subject can process them and generate contributions in the process of improving the city [2, 3]. But when accessing the data, technical differences appear such as file formats, file structure, column names, data types, magnitudes, etc., which make it difficult, and in some cases impossible, to analyze the information.

Computer Science Research Institute LIDI (III-LIDI) - Partner Center of the Scientific Research Agency of the Province of Buenos Aires (CIC).
J. S. Preisegger, A. Greco, A. Pasini, M. Boracchia, P. Pesado—Fellow UNLP.

P. Pesado and J. Eterovic (Eds.): CACIC 2020, CCIS 1409, pp. 337–349, 2021.
https://doi.org/10.1007/978-3-030-75836-3_23

The proposed framework is aimed, on the one hand, at linking datasets obtained from open portals and about a specific subject, and, on the other, at allowing a joint analysis of these data in a simple way that does not require advanced technical expertise. This framework is based on five steps: *1) Search, 2) Preliminary analysis, 3) Direct loading, 4) Standardization, and 5) Linkage.* This process is supported by the use of the IndiMaker tool.

IndiMaker has the potential to process files in various formats, apply different operations to their contents, and link the files, creating a dashboard of indicators that helps the user visualize the operations performed on the files. This article incorporates a more detailed description of the tool presented in [4].

To validate the process, the framework was applied to open data related to the environment, in particular, to data sets obtained from different public portals on air quality, water quality and energy consumption.

The second section presents the concepts of sustainable smart cities and open data. Then, in the third section, the general concepts about the indicator dashboards and the IndiMaker tool are discussed. The fourth section introduces the open data linking framework. In the fifth section, the framework is applied to environment-related data, and finally, in the sixth section, our conclusions and future work are discussed.

2 Sustainable Smart Cities and Open Data

Cities thrive on the participation of their communities. Citizens constantly generate information that can later be used in making decisions about the development of that city and, after a while, will affect the lives of those citizens. Achieving that citizens have access to data and being allowed to participate in their analysis, thus contributing to the development of the city is an important contribution in order to turn a city into a smart city.

2.1 Smart Sustainable Cities

In general, the concept of smart cities is related to the use of technology to carry out city activities, but, in reality, it is much more than that. According to [1], Smart Sustainable Cities represent the last stages of progression through digital cities and smart cities, and are considered as a continuous transforming process, based on the collaboration and commitment of different actors, building different capacities (human, technical and institutional) in a way that improves quality of life, protects natural resources, and pursues socio-economic development. The International Telecommunications Union (ITU) of the United Nations established one of the pioneering definitions of a smart sustainable city: "A Smart Sustainable City is an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social, environmental as well as cultural aspects".

2.2 Open Data

Society expects increasingly more from its government and government officials. These demands include transparency and an efficient management of public goods, as well as collaboration with different sectors of society and participation in the decision-making process [5]. Based on these requirements, and assisted by new technologies, a new type of government with greater citizen inclusion came to be, allowing citizens to contribute to public policies and participate in the decision-making process [6, 7].

The implementation of an open government resulted in data opening, which consists in making available to society data about common citizen interests so that, in any way, they can develop new ideas or applications that will deliver new data, knowledge, or other services that the government is unable to deliver [3, 8, 9].

The data that made available to society are very diverse, and this sharing process is not only carried out by government agencies – international organizations, NGOs, and other organizations promote various measures to make gradually more data sources available to society, not only related to government management, but also to other areas, such as the rational use of resources and the protection of the environment.

3 Using Open Data to Create Dashboards

The data are used to generate indicators that are in turn used to build dashboards. For the generation of these dashboards, data sources and datasets must be analyzed, and tools that allow establishing relations among data must be available, which is used to build relevant information that becomes an indicator that allows improving the decision-making process.

3.1 Data Sources

The new paradigms, developed by the different organizations, coordinate actions to improve quality of life for society through data opening and the improvements proposed in city infrastructure. Various applications and tools were generated, from multiple sectors, to provide support and automate, or improve, the process of publishing, searching and, sometimes, processing information for the different sectors of society. Among these tools, catalogs and open data portals stand out. Organizations use these to publish data on different aspects of their activities and the environment in which they operate. For example, some countries, provinces, municipalities or organizations have portals where they unify data from the different regions or topics in which they specialize. Those that are most advanced in the area publish their data and describe them using data schemes, which helps make data more descriptive in terms of content.

A tool that uses this information to allow data to be searched more globally is Google Dataset Search, which is a search engine specialized in finding datasets stored on the web, through keywords, as long as they use schema.org dataset tags or equivalent structures represented in the Data Catalog Vocabulary (DCAT) format [10].

3.2 Datasets

A dataset is a collection of data that is usually tabulated, that is, it corresponds to the contents of a single database table or a single data matrix, where each column of the table represents a specific variable and each row represents a specific member of the data set.

Datasets are the backbone of data portals and catalogs. They group one or more data resources and, for their publication, they require prior preparation in order to be processed and reused by third parties. According to [11], this includes three activities: *1) Documentation*: this activity consists in defining the metadata that each of the datasets to be published will have. Metadata describe the basics of the dataset, and are used to organize, classify, relate, and find the necessary data (e.g., title, description, institution, license, category, publication date, etc.). *2) Structuring*: it consists in preparing the dataset to be published with a structured format, without inaccurate or empty fields, which allows reusing and processing in any software. *3) Data loading*: it consists of publishing the data on a platform that allows organization and easy access by those who are going to reuse the data.

3.3 Dashboards

An indicator can be defined as a piece of data, or a set of data, that helps to objectively measure the evolution of a process or an activity corresponding to any organization. Indicators can be organized and connected to form a dashboard. These dashboards allow a more exhaustive monitoring and evaluation of the process or activity. In addition, they generally allow visualizing their evolution graphically, which helps interpret the results.

3.4 Tool for Dashboard Generation - IndiMaker

IndiMaker is a system that can be accessed from any web browser and allows connecting datasets to build custom indicators in dashboards.

To start using the tool, a username and password are required. When you open the login screen, you will be asked for these credentials to begin.

Before building an indicator, it will be necessary to generate a dashboard, which will contain a set of indicators that will generally have common objectives to measure.

In addition to the set of indicators mentioned above, dashboards also have a name that identifies them, a description of their purpose, and a set of datasets that must be loaded by the user when building each dashboard. It should be noted that datasets can be removed if they are no longer necessary, and new ones can be added. This tool allows importing datasets in various formats (.xls, .xlsx, .xml, .ods and .csv).

An operation that is very interesting is the combination of datasets. This operation can be done using columns with common content between the datasets. The result of this operation will be a new dataset with richer information.

It can also be used to homogenize all this information and store it in the database; then, the user can perform operations on the data stored and build indicators. The indicators must be generated appropriately; otherwise, they can generate inaccurate, incorrect or subjective information, which would hinder data analysis. One of the great virtues of

this tool is its simplicity when generating indicators, since it is designed to make use of the information in a way that makes it easy for the user to perform complex operations between the different data, avoiding *a priori* the generation of inaccurate indicators. Therefore, the indicators that are generated will allow developing a quantitative measure that will have meaning for those who analyze it.

At the same time, the tool can be adapted to different devices, like a desktop computer, a tablet or a smartphone, and it can also be used in both Spanish and English.

The tool has a role management system, which means that there will be different users in the system with different permissions and, thus, different operations available to them.

When accessing the tool, there is a menu on the left of the screen with all the options available to the user. These options include:

- General dashboard: This will be where, based on the different charts, the results of the indicators from the different dashboards can be viewed.
- Dashboards: Here, users can manage their dashboards, add new dashboards and remove or edit existing ones. This view will show a paginated list with all the dashboards for the logged-in user. The number of records to display per page can be configured, results can be sorted by name, and dashboard searches can be sorted by name, which improves access speed and favors simplicity if there are many results.

To build the indicators, their name, type of indicator, owner, measurement frequency, description, and reference levels must be indicated, in addition to the operation on which the corresponding indicator is based. The operations that can be performed on data include addition, subtraction, division, percentages, data grouping, and various logical comparisons, including the ability to use regular expressions for more advanced users. Users can build indicators by combining these operations.

The name field is used to identify the indicator.

The type field allows selecting a type from those already loaded in the system, and it will be used to classify the indicator.

The owner is the person responsible for defining the different roles associated with the indicator:

- Who is responsible for generating information
- Who gathers the information
- Who analyzes the information
- Who reports or presents the information obtained with the indicator

The frequency field allows selecting how often the indicator will be recalculated.

Calculation frequency should not be confused with information collection frequency. For example, to analyze the work of a supplier, it may be convenient to calculate the indicator every six months. However, is the information on the work of that provider going to be gathered after they have been working for six months? The answer is no. It will be more convenient to have a supplier control sheet where their weekly management is tracked.

The description field, optional but recommended, can be completed with free text by the user, and it is used to describe the objectives of the indicator.

Within a section called "Indicator formula" there are some fields that will allow carrying out operations on the data belonging to the sources in the dashboard. This is the process used to create the calculation for the indicator, known as its "formula".

An example of formula construction can be seen in Fig. 1. In this example, the "USA_Air" dataset was used, which contains information about the air in different cities of the United States for different years. This formula allows knowing the total number of days in 2020 for each of the cities in which air pollution is considered acceptable.

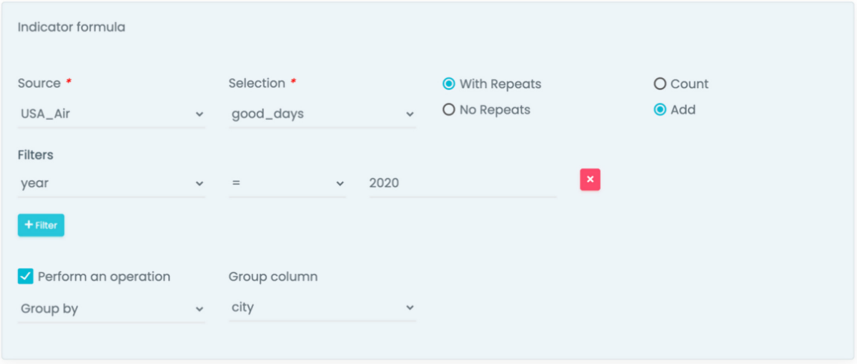

Fig. 1. Building the formula for an indicator

Finally, both the "Critical Level" and "Satisfactory Level" fields are optional.

These fields allow entering reference values for the user when calculating the indicator. These entered values will be represented when creating the chart for the indicator, and will allow alerting the user about possible deviations from the objectives established by the indicator.

It should be noted that, at all times, the user can search for a particular indicator. To do so, there is a search button located in a top menu bar that, when clicked on, will display a text field to be completed, where users can enter the name of the indicator (or part of it) that they want to search for.

The indicators in each dashboard can be represented in various ways, including different charts, which allows users to easily be alerted to potential deviations from previously established objectives. Figure 2 shows a sample representation for an indicator as a line graph.

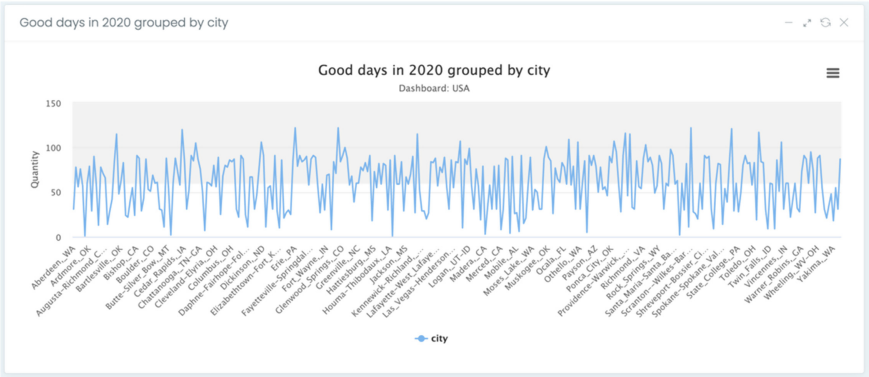

Fig. 2. Representation of an indicator as a line graph.

4 Open Data Linking Framework

When analyzing the data from the different catalogs or portals, there are incompatibility issues between the data formats used by the different providers. Given this situation, to achieve successful data linking, it is essential to analyze data sources and formats beforehand. Consequently, progress was made in generating a five-step data linking framework: *1) Search,* 2) *Preliminary analysis, 3) Direct loading, 4) Standardization, and 5) Linkage.*

1. *Search*: Searching for datasets related to the area of interest, through organizations' data portals or catalogs, or using the Google Dataset Search tool.
2. *Preliminary analysis:* Analyzing the datasets obtained, verifying that information is in a format supported by the tool and that all the columns have a header. Also, checking content format (completing rows and columns, transposing rows by columns in any of the datasets, etc.) for a successful linkage.
3. *File loading:* Loading the file into IndiMaker. When loading the file, the tool will perform a series of checks on the content. If validated, the comparison instance will begin. In the event that inconsistencies are detected, the content will be standardized.
4. *Standardization:* The standardization process can be done manually or automatically (using an external tool), depending on the size of the file. In this instance it is verified:

    a. The data in the columns of the datasets are of the same type.
    b. The data in the columns have some value.
    c. The max amount of data in each row is not exceeded.

5. *Linkage*: Consolidating the tables based on a common parameter, if necessary, to obtain more information and to be able to relate the data to analyze them in a simple and direct fashion.

5 Environment Open Data Linkage – Case Study

Governmental and non-governmental agencies make available to the community numerous sets of environmental data from their geographic region. When analyzing the information as a whole to obtain regional values, there are incompatibility issues between the data formats used by the different organizations. Data sources and their formats should be analyzed to make the process of information linkage possible.

Below, the five steps of the framework will be applied:

Search. The environment was selected as a case study. Among the different aspects that are relevant for this topic, it was decided to limit data to the aspects that affect society every day, such as water and air quality, which are essential for society and its future. Also, energy generation was included, which is often related to quality of life.

Data searches were carried out using the following terms: *Drinking water quality*, *Energy generation*, *Air quality*. In Table 1, below, shows the datasets selected among those obtained from the different countries.

Table 1. Datasets obtained per country.

Country	Water	Air	Energy
USA	✓	✓	✓
Chile	✓	✓	✓
Brazil	✓	✓	✓
Uruguay	×	✓	✓
Paraguay	×	×	✓
Bolivia	×	×	×
Peru	×	×	×
Colombia	✓	✓	×
Argentina	✓	×	✓

Preliminary análisis. From the data obtained, data type, format and structure were analyzed in each case, so as to be able to link them using the proposed tool to carry out a more in-depth analysis of the environmental situation in the different regions.

Air and Water Data. In the case of air and water quality, it was possible to identify a certain standard to analyze existing magnitudes for different characteristics. It was observed that almost the same tests are carried out on different samples to analyze different characteristics and determine if they are within healthy margins for human consumption.

Energy Data. In the case of datasets regarding energy, the difference between the data published by the different organizations is greater. It was observed that some countries

simply publish an annual percentage and the type of energy production on which the different published energy generation plants are based, while others simply publish the percentages of their energy generation sources without plant-level granularity, and yet others publish totals corresponding to each energy plant capacity, energy plant types, and even the companies that own them.

Direct Loading. After collecting the data, a direct loading process was carried out, as a starting point, to analyze whether the untreated data met tool requirements. It was found that not all files had the data in an orderly manner and without errors – only 35% of the datasets passed this minimum control, indicating that the data published by the organizations required further standardization before they could be processed. If the data are analyzed by area, it can be seen that, based on this standardization, progression is as follows: Water: 80% direct loading 20% indirect loading; Energy: 66% direct loading and 33% indirect loading; Air: 50% direct loading and 50% indirect loading.

Standardization. In all three areas, it was observed that, beyond the standardization of certain published data, data are in a very "raw" format that makes optimal processing impossible. In many cases, they have empty spaces, or even have more data in the rows than is declared. In each case we can see:

Air and Water Data. There are differences in the structure of the datasets, the units used for the different quantities that are analyzed, and how these are stored in the datasets. For example, variations in columns to rows or one field split into several ones. This makes datasets mapping difficult.

Energy Data. In this case, it was observed that the measurement units used in the different datasets vary greatly, which makes a linear comparison of the data impossible. At the same time, due to the great difference in terms of the data published by each organization, a selection of certain common fields to all datasets is required, based on their type, so as to be able to represent them in a typified way. This is difficult due to the existing dispersion.

Linkage. Once the data are standardized, it is possible to link different datasets that have data in common, allowing a regional analysis of the information. With these datasets, the tool allows selecting columns to perform operations and obtain values with which various charts can be generated for a linear analysis of magnitudes of interest. Then, the analysis carried out with each of the data sets can be made visualized.

Water Data. Among the information collected on water quality, data from Colombia and Brazil were selected, as an example. The number of measurements carried out in both countries were added up, and the vertical bar charts that can be seen in Figs. 3 and 4 were generated. As it can be seen, the information published by Colombia is split by state and, on the other hand, the information from Brazil is split based on the different natural effluents in the country. The tool allows, in any case, visualizing this information regardless of these differences.

In a further analysis, effluents could be mapped to their corresponding Brazilian states and data could be compared by geographic region.

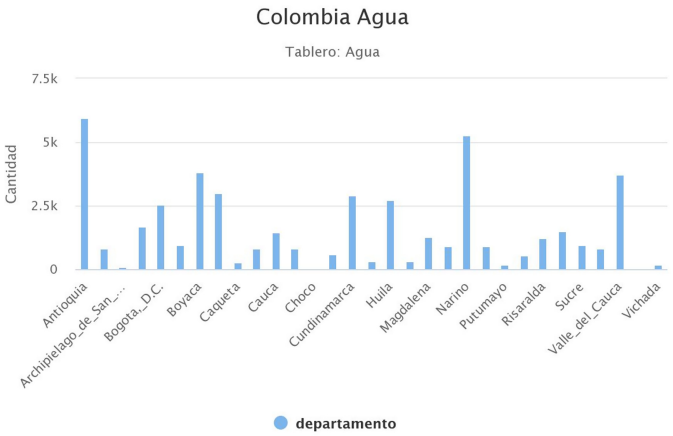

Fig. 3. Number of water quality measurements in Colombia

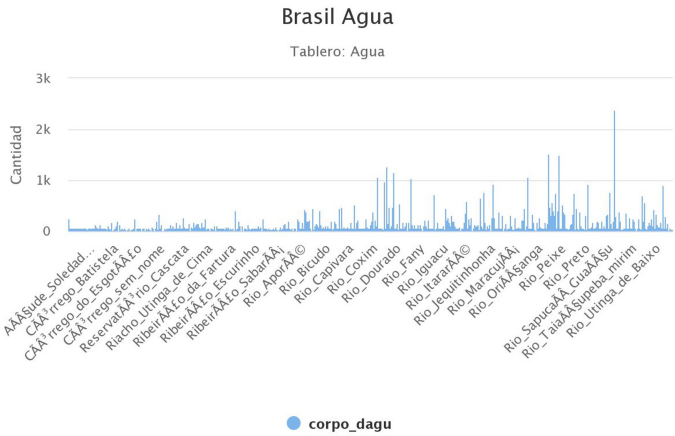

Fig. 4. Number of water quality measurements in Brazil

Air Data. As regards the information collected on air quality, suspended particulate values from the cities of Bogota in Colombia and Montevideo in Uruguay were selected to show how the tool can be used. Annual average values were calculated for Bogotá (this step was not necessary with Montevideo data because this city already records average values), and the horizontal bar charts shown in Figs. 5 and 6 were generated. In this case, the information published by Colombia was more detailed based on multiple measurements per year, so it had to be processed further to obtain a format similar to that published by Uruguay. The tool allows, in any case, visualizing this information regardless of these differences, for the same yearly periods.

Energy Data. Among the information collected on energy production, data from Paraguay and New York City were selected as an example. Beyond their differences in

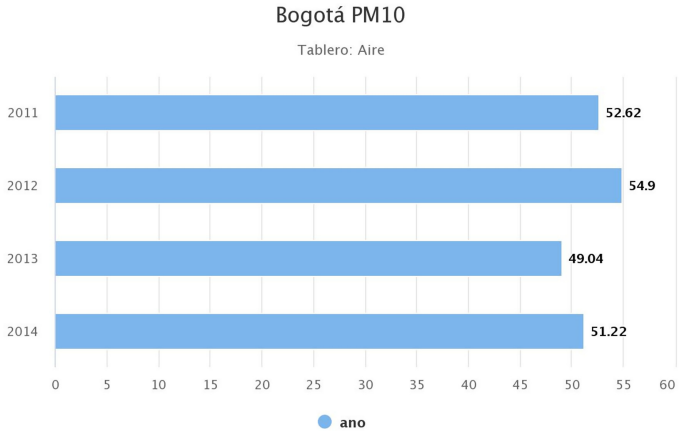

Fig. 5. PM10 measurements in Bogota.

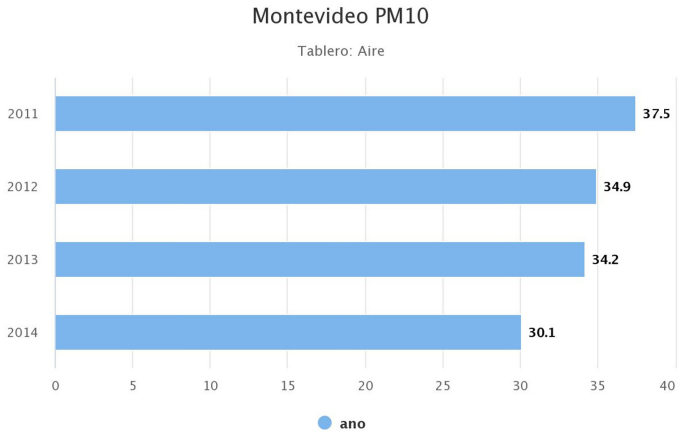

Fig. 6. PM10 measurements in Montevideo.

size and population, a comparison was made with these data to observe the differences in terms of energy production values, in GWh. In this case, the information published by both sources was very similar, although NYC also specified energy production plants. With this information, the line charts presented in Figs. 7 and 8 were generated, showing that NYC produced, in the period analyzed, more energy than all of Paraguay.

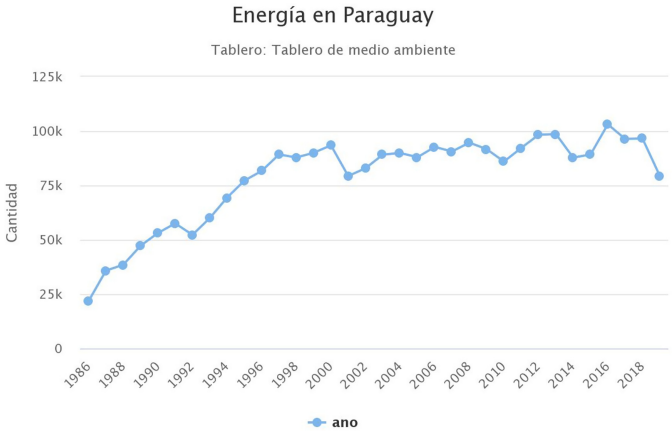

Fig. 7. Energy production per year in Paraguay

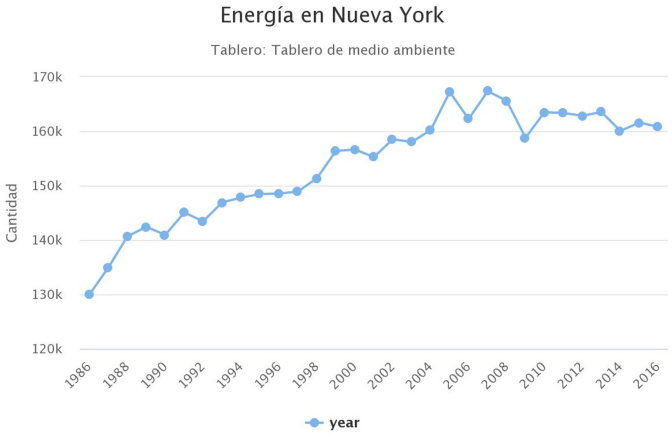

Fig. 8. Energy production per year in Paraguay

6 Conclusions and Future Work

Throughout the article, the basic concepts of Smart Sustainable Cities, Open Data, Data Sources, Datasets and Dashboards were introduced. Emphasis was placed on how these concepts can be combined to generate a higher level of information for society, which is achieved by linking already available data using the IndiMaker tool.

The tool was briefly described and a five-stage framework was generated to link the data made available by various organizations and compare them easily.

As a case study for the application of the framework, open data sources on water quality, air quality and energy production from 10 countries were downloaded, obtaining information in 8 of these data portals.

The information obtained was processed using the framework, standardizing the information where necessary, and then IndiMaker was used. The tool allowed performing

operations on the data and easily generating indicators. The differences found in the information published on the same topic were notable; however, with the application of the proposed framework, the tool was able to process the information, generating indicators in a user-friendly way and various charts that facilitated data analysis. It should be noted that the model can be extended to any area of interest.

In the future, we expect to connect the tool with APIs from various data portals from several organizations and expand the range of operations offered by the tool, including the possibility of displaying various indicators on the same chart.

Acknowledgments. Project co-funded by the Erasmus+ Programme of the European Union. Grant no: 598273-EPP-1-2018-1-AT-EPPKA2-CBHE-JP.

References

1. Estevez, E., Lopes, N.V., Janowski, T.: Smart Sustainable Cities - Reconnaissance Study. Operating Unit ON Policy-Driven. Electronic Governance. United Nations University, Canada (2017)
2. Chun, S.A., Shulman, S., Sandoval, R., Hovy, E.: Government 2.0: making connections between citizens, data and government. Inf. Polity **15**(1–2), 1–9 (2010). https://doi.org/10.3233/IP-2010-0205
3. Concha, G., Naser, A.: Datos abiertos: Un nuevo desafío para los gobiernos de la region. Instituto Latinoamericano y del Caribe de Planificación Económica y Social (2012)
4. Preisegger, J.S., Greco, A., Pasini, A., Boracchia, M., Pesado, P.: Marco de vinculación de datos abiertos aplicado al contexto de datos medioambientales, pp. 684–694 (2020). https://sedici.unlp.edu.ar/handle/10915/113243
5. Calderón, C., Lorenzo, S.: Open Government. Gobierno Abierto (2010)
6. Naser, A., Ramírez-Alujas, Á., Rosales, D. (eds.): Desde el gobierno abierto al Estado abierto en America Latina y el Caribe: Planificación para el Desarrollo (2017)
7. Gil-García, J.R., Criado, J.I.: Las Tecnologías de Información y Comunicación en las Administraciones Públicas Contemporáneas (2017)
8. Pasini, A., Preisegger, J.S., Pesado, P.: Modelos de evaluación de gobiernos abiertos, aplicado a los municipios de la provincia de Buenos Aires. In: XXIV Congreso Argentino de Ciencias de la Computación, vol. XXIV, p. 10 (2018)
9. Pasini, A., Preisegger, J., Pesado, P.: Open Government assessment models applied to province's capital cities in Argentina and municipalities in the province of Buenos Aires. In: Pesado, P., Aciti, C. (eds.) CACIC 2018. CCIS, vol. 995, pp. 355–366. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20787-8_25
10. Noy, N., Burgess, M., Brickley, D.: Google dataset search: building a search engine for datasets in an open web ecosystem. In: The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, pp. 1365–1375 (2019). https://doi.org/10.1145/3308558.3313685
11. Naser, A., Ramirez, A.: Plan de gobierno abierto. Una hoja de ruta para los Gobiernos de la Región. CEPAL - Manuales **81**, 80 (2017)