

A user-centered process for the analysis and visualization of open data sets

Marcia Tejada¹ and Diego Torres^{1,2}[0000-0001-7533-0133]

¹ Dto. CyT, UNQ, Roque Saez Peña 352, Bernal, Argentina
tejedamarcia@gmail.com

² LIFIA, CICIPBA-Facultad de Informática, UNLP
La Plata, Argentina
diego.torres@lifia.info.unlp.edu.ar

Abstract. Open data is growing all the time throughout the world. Open government is advancing and more and more open data portals are available to be consulted by anyone. It is assumed that joining and combining two or more data sources can provide new information or knowledge that was not previously available. To be able to combine different *datasets*, the use of statistical and computer science techniques such as data mining and machine learning is suggested. Although this is a practice that is currently carried out by data science professionals, this work invites the community in general to be able to do it through a process centered on the user. This article presents a process and a web tool that implements the analysis and visualization process.

Keywords: Open-datasets · User centered Process · Data visualization.

1 Introduction

Currently a large number of public administrations in the world and non governmental organizations are opening their data so that any person or entity can use them [1, 2]. Projects that combine smart cities and the use of the internet of things present a scenario of proliferation of open and standardized data, with IoT and open data, interoperability and open standards[3, 4]. The way of publishing open data is done through data sets (textit datasets), which can cover different areas: science[5], economy[6], transport[7], education.

Open data[8] is data that anyone can access, use and share. It can come from any source and cover different topics: science, technology, economy, finance, education, among others[2]. However, not having forms of aggregation and visualization makes open data difficult for users to understand and manipulate[9].

Combining different data sources increases the level of information that can be extracted from open data. To a large extent, open data is made for a specific purpose. However, the possibility of combining two open*datasets* that describe events in the same geographic area can generate new interpretations, for example combining cases of disease infections with the description of housing and economic development. Interoperability is one of the goals of open data[4]. Through

the use of standard formats, it is possible to connect computer systems through data sharing. Open data presents great opportunities to be combined with studies other than the ones that originated it. Since this speeds up research times and provides scientific and social advances of great impact. The use of open data during the advance of COVID-19 has demonstrated the importance of the same[10, 11].

Information is a processed data that may have some kind of utility or value. Converting data into information involves a process of knowledge and understanding that was not previously known[12]. Taking advantage of open data and the information that it provides, new knowledge could be generated by studying the relationships between different *datasets*. This can be possible using tools from the field of statistics and computation: data mining [13, 14] and machine learning [14, 15].

It is difficult to analyze and process the large amount of available open data by people who do not have the skills to analyze it. There are some alternatives like the ones described in the[9] work. However, allowing the general public to interpret open data is a constant concern of governments[16].

There are some approaches without requiring programming skills. Tableau enables end users to create visualizations and order information. Other approaches for end users simplify viewing but still require some technologies advanced knowledge[17]. Google Maps is a well known tool for combining georeferenced data, however it has limitations[18].

This work will focus on presenting a process that allows the analysis and visualization of open geo-referenced data[19] from a user-centered perspective. We propose the creation of a tool that allows relating georeferenced *datasets* so that they can be combined and analyzed in a simple way. This approach is aimed at people with no programming skills. The strategy is to accompany the user during the process that involves viewing the *datasets*, analyzing them alone, combining them with other *datasets* available in the application and then proceeding to the analysis and display of the information that was built in the process in maps. It presents a process to manipulate and visualize *datasets* that combine transformation and analysis functionalities that are generally offered as parts of software modules to be manipulated by developers, such as *clustering* algorithms.

This article is organized as follows. Section 2 describes the proposed approach in combining and visualizing the dataset. The whole merge and display process is described in Section 3, which includes implementation details. Finally, Section 4 presents the conclusions and future work.

2 Motivation and approach

Generation of open data allows their free use for analysis, visualization and use, possibly, in other contexts. Open science and citizen science are contexts in which a large number of *datasets* are generated. So do governments with open data policies. It is natural for the generation of *datasets* to be carried out

in a specific context and then released for public use, however their use and combination is complex.

For example, the *Encuesta permanente de hogares in Argentina*³ lists the housing characteristics of Argentina and the citizen science project GeoVin⁴ analyzes appearances of the insect vector of Chagas disease, endemic in Latin America. The combination of both *datasets* could be considered interesting to analyze if there is any relationship between the number of vector insect occurrences and the housing conditions in the region. This, even though neither of the *datasets* was thought in terms of the other. Thus, it is possible to think of combining different *datasets* to analyze an endless number of variables.

The focus of this work is to propose a user-centered tool that is easy to use for the analysis and visualization of open data. The focus is on defining a usable process that articulates strategies for combining *datasets* that were created in isolation from each other, and strategies for displaying the combined data. The following sections describe the basic tools for combining *datasets* and display tools in isolation. And then the user-centric process will be described in the way they abstract from the combination and display algorithms, and turn them into simple utilities of a larger process.

2.1 Combination of *datasets*

Datasets combination has the main goal of adding value to a data set with values from another data set. This allows you to take advantage of different *datasets* together with others to increase their potential and it also could generate more valuable information.

The design of three types of strategies for the combination of *datasets* is proposed. All three follow the philosophy of ontology alignment [3], based on ontology reconciliation, in this case *datasets*. Find relationships between concepts that belong to different sources [20].

Each file merge strategy will take two *datasets* and return only one with the result of the merge strategy.

For this, we define *datasets* as a matrix, or table, where the columns indicate characteristics and the rows have the values of an element. As *datasets* are geolocated, each row contains coordinates and the characteristics of the element that is in those coordinates. For example, if dataset A contains 3 elements that are found in a latitude, a longitude, and the characteristics CarA and CarB, then we can annotate the dataset as A (3X4), since A has 3 rows (one for each element) and 4 columns (latitude, longitude, CarA and CarB). More generally, a D (mXn) dataset is a dataset that has m rows and n columns.

The proposed strategies consist of aligning each row of a *dataset* with one or more rows of another *dataset*. If a row is aligned, then a new row is generated in the result dataset where the characteristics of the first row are concatenated

³ <https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos>, accessed on July 29, 2020,

⁴ <http://geovin.com.ar> accessed on July 29, 2020.

with those of the aligned row. This gives an idea of augmenting the original row with information from the lined up. In other words, if you want to combine the data set A ($n \times m$) with the data set B ($n' \times m'$). The result would be a new data set C ($p \times q$) where $n \leq p \leq n * p$ y $m \leq q \leq m + q$.

Here are three strategies for aligning, although the list could be longer.

Closest point In this type of combination for each element located at a point of the *dataset* to which you want to add information -the 'base' set-, find the element that is at the closest point (at a maximum parameterizable distance) from the data set to be added. Once the *datasets* have been combined, this information will result in a single line with the data from both points. When applying this combination what happens is that two nearby points become one, containing the information of both. In other words, the information about the elements found at that point has been enriched.

If the base **dataset** is B ($n \times m$), and the *dataset* to be added is S ($j \times k$), the resulting dataset will have a maximum of R ($n \times m + k + 1$), since it will have the number of rows of the base *dataset*, and the columns of the sum *dataset* will be concatenated, and an additional column with the value of the maximum distance. If there are no nearby points, the information from the base file is removed from the result. It is necessary that both files have georeferenced information. This type of combination has a parameter that is the maximum distance to which the closest point can be, which allows to equalize the way in which the information of the points is increased. If this were not parameterizable, all the points would have a closer point even though they were separated by many kilometers apart and could cause unwanted information. Information on the distance to the closest point of each point is also saved in the resulting dataset.

Radial distance In this case, for each point on the map of a data set, the points in another data set that are at a certain parameterizable distance are searched, thus forming a circle of determined radius around each point. The size of the circumference around a point is defined by a distance parameter. It is also necessary that both *datasets* have georeferenced information. If the base **dataset** is B ($n \times m$), and the *dataset* that will be adding is S ($j \times k$), the resulting dataset will have a maximum of R ($n * j \times m + k + 1$), since it will have a maximum that each row of the base *dataset* relates to all the rows of the dataset to be added. The number of columns maintains the logic of the previous combination.

Equal characteristic The combination of *datasets* seeks to add the information of a data set (not necessarily georeferenced) to a georeferenced data set based on some similarity between its columns other than the location. The final result will be at most like that of the radius combination around the point.

2.2 Configurations and visualizations

In addition to the combination of *datasets* presented in the previous section, the possibility of displaying the *datasets* is presented, in their original version or the combined one. We present three ways of displaying maps.

- **Simple map:** It simply shows on a map the detail of information contained in the rows located in the position that indicates the latitude and longitude.
- **Layered map:** This type of visualization displays the information from two *datasets* on a map at the same time. Each set will be displayed on a different layer of the map. Layers can be viewed individually or together.
- **Clustered map:** This visualization shows for a single data set the information of the points that it contains but grouped in clusters. To do this, you can choose to analyze the data with different clustering algorithms: One of them is KMeans[15], which is an unsupervised classification algorithm that works with a K parameter that defines the number of clusters (groups) in which you want to group the information. Based on this parameter, the algorithm will divide the information into K groups according to their characteristics (See Fig. 1). The other algorithm that is being applied in this work is called Meanshift [21]. It also works by grouping information, but unlike KMeans it does not receive any parameters and decides based on the information that it is classifying how many clusters it should form.

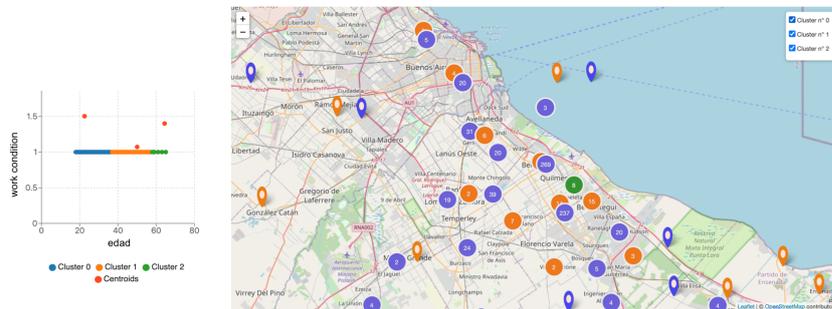


Fig. 1. Clustered map

3 Combining and displaying open *datasets* process

All the activity of manipulating open *datasets*, their analysis and visualization are simplified by thinking from the user's perspective through a sequence of processes. It is illustrated in Fig. 2. There you can see the sequence and communication between processes and sub-processes with specific tasks, for example

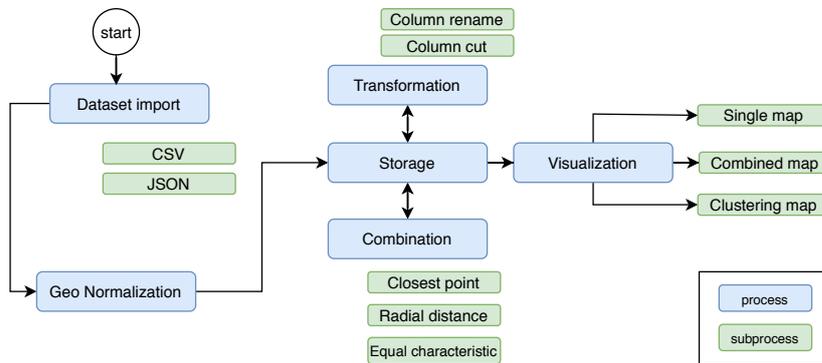


Fig. 2. Combining and displaying open *datasets* process

those dedicated to importing files in CSV (comma-separated values) or JSON format.

As told before, the general process is geared towards simplifying *datasets* manipulation tasks so that a person without programming skills can analyze *datasets*. The processes and sub-processes generate a high level of abstraction and increase the simplicity of the activity as black boxes. The processes are described below in order:

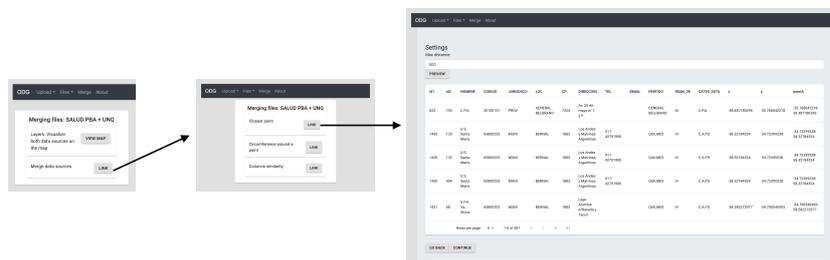


Fig. 3. Combining process screening.

- **Dataset import:** Allows you to add a dataset file in any format to the system. The user should not worry about understanding the format of the data set. This process includes sub-processes to decode different formats, for example CSV or JSON. These sub-processes can be extended to support more formats.
- **Geo normalization:** Once the *datasets* have been imported, the columns representing the geolocation information must be detected. In case it cannot be detected automatically, the user will be asked to indicate the column (s) with the latitude and longitude information.

- **Storage:** The normalized *datasets* are stored in the system database. There they will be available to apply functionalities of the following processes.

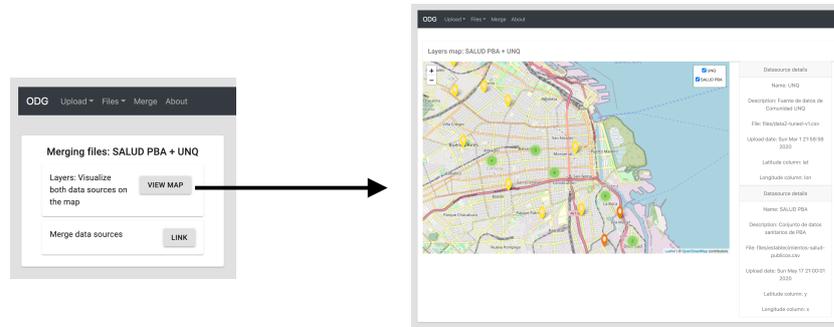


Fig. 4. Layering maps screening process.

- **Transformation:** This process modifies the general structure of a data set. Sub-processes include renaming columns, transforming values or formats of a column, deleting a set of columns.
 - **Combination:** This process encompasses those operations that allow the combination of *datasets*. The sub-processes that are included are those that have been described as Closest Point, Distance Radius and Equal Characteristic. New ones can also be added. Fig. 3 shows the sequence in the processes to combine two datasets through closest point. The final screen shows a field where the maximum distance is indicated and below the preview before saving the changes. Preview is a functionality that was considered relevant, since much of the analysis work requires a trial and error stage.
- **Visualization:** This is the final process. At this point, the ways to visualize the work done with the *datasets* are decided, whether simple or combined. The threads seen in the figure correspond to those previously described and in the same way with the combination ones, they can be extended. As an example, Fig. 4 shows the steps required to, after selecting two datasets, visually combine them into a map combining the layers. On the left of the figure you can see the option to visualize, and on the right the result of the visualization on a map that includes the points and can be selected (in the upper right corner) to see both datasets simultaneously or one at a time.

3.1 Prototype

Both the process and the prototype have been developed through proofs of concept and interviews with open data users and professionals in disciplines other than software engineering. In particular, the people with whom the evaluations

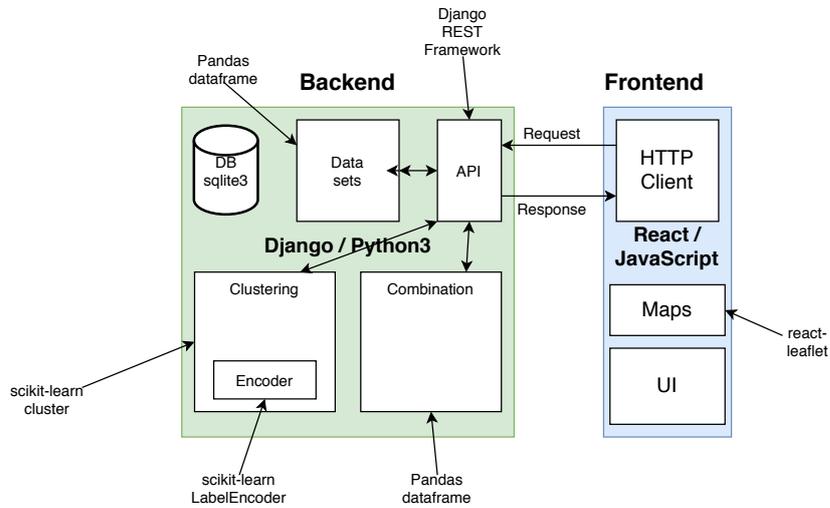


Fig. 5. Architecture

and interviews have been carried out are dedicated to sociology, economics and different studies related to student mobility in universities.

The developed prototype is implemented as a Web application. It is defined in a *backend* and *frontend* client-server architecture, which communicate through a restful API. It was decided to use React as a frontend tool, a JavaScript library developed and maintained by Facebook. Leaflet and React-Leaflet Material-UI. The Fig. 5 details the organization of the architecture.

4 Conclusions and future work

Open data sets have proliferated in recent times so that any citizen can consume them, however the volume of data in these sets requires easy-to-use tools. This work presents a user-centered process model to be able to analyze, combine and visualize open data sets. It is modularized in threads, which can be extended.

The approach is bundled with an implementation of basic merge, display, and clustering capabilities.

As future work, the need to carry out usability evaluations with a significant number of users is highlighted, since the present work includes conceptual tests at the moment. It is also desirable to incorporate more merge and display threads.

References

1. Davies, T., Perini, F., Alanso, J.: Researching the emerging impacts of open data (2013)
2. Kitchin, R.: The Data Revolution. Sage Publications Ltd (2014)

3. Ahlgren, B., Hidell, M., Ngai, E.C.H.: Internet of things for smart cities: Interoperability and open data. *IEEE Internet Computing* **20**(6) (2016) 52–56
4. Domingo, A., Bellalta, B., Palacin, M., Oliver, M., Almirall, E.: Public open sensor data: Revolutionizing smart cities. *IEEE Technology and Society Magazine* **32**(4) (2013) 50–56
5. Arza, V., Fressoli, M., López, E.: Ciencia abierta en argentina: un mapa de experiencias actuales. *Ciencia, docencia y tecnología* **28**(55) (2017)
6. Mourmoutsev, D., d’Aquin, M.: *Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning*. Springer Verlag (2016)
7. Kujala, R., Weckström, C., Darst, R.K., Mladenović, M.N., Saramäki, J.: A collection of public transport network data sets for 25 cities. *Scientific data* **5** (2018) 180089
8. : El manual de open data
9. Saddiqa, M., Larsen, B., Magnussen, R., Rasmussen, L.L., Pedersen, J.M.: Open data visualization in danish schools: a case study. (2019)
10. Amaro, R.E., Mulholland, A.J.: A community letter regarding sharing biomolecular simulation data for covid-19. *Journal of Chemical Information and Modeling* (2020)
11. Moorthy, V., Restrepo, A.M.H., Preziosi, M.P., Swaminathan, S.: Data sharing for novel coronavirus (covid-19). *Bulletin of the World Health Organization* **98**(3) (2020) 150
12. Engvall, E.B.T.: *Open data? Data, information, document or record?* Emerald Group Publishing Limited (2014)
13. Maimon, O., Rokach, L.: *Data Mining and Knowledge Discovery Handbook (Second Edition)*. Springer Verlag (2010)
14. Sammut, C., Webb, G.I.: *Encyclopedia of Machine Learning and Data Mining*. Springer Verlag (2017)
15. Witten, I.H., Frank, E.: *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufman (2005)
16. Sieber, R.E., Johnson, P.A.: Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly* **32**(3) (2015) 308 – 315. <https://doi.org/https://doi.org/10.1016/j.giq.2015.05.003>
<http://www.sciencedirect.com/science/article/pii/S0740624X15000611>
17. Ahrens, J., Geveci, B., Law, C.: *Paraview: An end-user tool for large data visualization*. *The visualization handbook* **717** (2005)
18. McQuire, S.: One map to rule them all? google maps as digital technical object. *Communication and the Public* **4**(2) (2019) 150–165. <https://doi.org/10.1177/2057047319850192>
<https://doi.org/10.1177/2057047319850192>
19. Hill, L.L.: *Georeferencing - The Geographic Associations of Information*. Cambridge, MA: The MIT Press (2006)
20. Euzenat, J.: An api for ontology alignment. In: *International Semantic Web Conference*, Springer (2004) 698–712
21. Anand, S., Mittal, S., Tuzel, O., Meer, P.: Semi-supervised kernel mean shift clustering. *IEEE transactions on pattern analysis and machine intelligence* **36**(6) (2013) 1201–1215