Analysis of Urban Information of Colombia on Wikidata

John Samuel and Oscar Carrillo

October 23, 2019

Analysis of Urban Information of Colombia on Wikidata

John Samuel[1,2][[0000−0001−8721−7007] and Oscar Carrillo[1,3][0000−0001−5081−1774]

[1] CPE Lyon, Université de Lyon, LIRIS, UMR 5205, Villeurbanne, France
[2] LIRIS, UMR 5205, Villeurbanne, France
[3] CITI Lyon, INSA de Lyon, Lyon, France
{john.samuel,oscar.carrillo}@cpe.fr

Abstract. Wikidata started in 2012 is a free, open, multilingual, collaborative, linked and structured knowledge base. During the last couple of years, Wikipedia communities of different languages have started exploring the use of Wikidata as a central store of information, where the community members can directly add structured data on Wikidata to be later be used by all the language editions of Wikipedia. However, making it a central store has several challenges, especially when key information like administrative heads, population etc. are outdated. In this article, we will first consider the available urban data related to some of the main cities of Colombia on Wikidata and see what type of information are currently contributed by the community members. We will then compare the historic evolution of these contributions with respect to the size of the cities in terms of the area, population, tourists as well as the influence of important moments and events on the recent history of Colombia. We will also focus on the frequency and recency of the contributions. The analysis of these data is important since it will ensure that relevant government agencies, Wikidata community members and tourism organizations work together to keep the information up to date. We will also discuss how our study can be further explored by other cities as well as countries for ensuring timely information on Wikidata and associated Wiki-media projects.

Keywords: Wikidata · Urban Data · Collaboration.

1 Introduction

The world is increasingly referred to as global village [3], thanks to the growing inter-connectivity among different cities, especially with the advances of transport infrastructure as well as the internet usage. It is much easier to know information on remote towns and villages now, than it was a couple of decades ago, thanks to the growing availability of linked open data [1] which encourages autonomous institutions to not only publish their data with open licences but also link with each other. Websites like Wikipedia provide a platform where users across the world can write articles on various topics, including countries, capitals, towns, villages etc. Wikipedia also allows these articles to be written in almost more than 300 multiple languages. These multilingual articles are linked to each other, thereby letting users easily switch and see other languages versions of the information available for a given subject. However, much of the content on Wikipedia is unstructured and several efforts have been made to extract relevant information from Wikipedia [9] including Wikipedia infoboxes [1, 2].

With the advent of Wikidata in 2012 [8], multilingual contributors can make contributions in the form of triples (subject item-property-value statements). Thus to state that Bogotá is the capital city of Colombia, the contributor can state Q739-P36-Q2841, where Q739 and Q2481 are the identifiers of Colombia (subject) and Bogotá (value) on Wikidata and P36 is the identifier of the property 'capital'. Wikidata (https://www.wikidata.org) with its single domain website address, unlike Wikipedia (a multi-subdomain website) has the advantage that users can change the language settings and can see the facts related to a given subject in any supported language (from a list of more than 300 languages) and also contribute in their local languages. Maintaining up-to-date information on all the multilingual Wikipedia sites is challenging and using a central multilingual site like Wikidata for recent information may be a possible way forward. Many language Wikipedias like Catalan, Basque and even English are now exploring to use data from Wikidata to enrich the Wikipedia infoboxes.

Wikidata is a collaborative website like Wikipedia and hence it's extremely important to understand how and what type of facts are entered by the users. Some of these facts evolve during time. Take for example, values like population of countries, cities change over time. It's therefore important to monitor whether articles (called items on Wikidata) on different subjects contains the latest information. Another major problem with collaborative sites is vandalism [4] and Wikidata also suffers from it, especially in the form of label or description changes[7, 6].

Wikidata is also increasingly being used as a knowledge hub [5] for other knowledge organisation systems. With the initial goal of supporting other Wikimedia projects including Wikipedia, it is now linked to other databases, with the use of external identifiers, thereby allowing its users to verify and also have additional information from these systems. However, maintaining a central site for facts means ensuring the latest up-to date information of all the facts.

In this article, our goal is to look at the current information of different cities of Colombia on Wikidata and see what type of information are currently contributed by the community members like the different properties used by urban cities, the use of external identifiers, use of images, edit reverts, article lengths (number of statements) etc.

Section 2 briefly presents the cities that we took into consideration. Analysis of the information of these cities are done in section 3. We discuss our results in section 4 and briefly present how our work can be further extended to other cities. Finally, we conclude our article in 5.

2 Cities of Colombia

We take into consideration the different departmental capitals, intermediary capitals, touristic cities, some villages and towns of interest. Table 2 gives a detailed information about our focus categories and the associated list of cities. We obtained the 20 main cities of Colombia using the Spanish language Wikipedia template *Principales ciudades de Colombia*[4]. We also took into account some of the towns in conflict. In total, we are considering 40 towns, villages and cities.

In the table, we have also given the Wikidata identifiers of each of these towns in parentheses. In the following section, we will explore the statements used to describe them and the associated multilingual articles.

S.No.	Focus category	List of cities
1.	Main cities of Colombia [5]	Bogotá (Q2841), Medellín (Q48278), Cali (Q51103), Barranquilla (Q62823), Cartagena (Q657461), Soledad (Q767071), Cúcuta (Q216847), Soacha (Q1011151), Ibagué (Q222755), Bucaramanga (Q243766), Villavicencio (Q749224), Santa Marta (Q209016), Bello (Q816024), Valledupar (Q376903), Pereira (Q51111), Buenaventura (Q996581), San Juan de Pasto (Q320015), Manizales (Q235190), Montería (Q852725), Neiva (Q638260)
2.	Important departmental capitals	Bogotá (Q2841), Barranquilla (Q62823), Cali (Q51103), Bucaramanga (Q243766), Medellín (Q48278)
3.	Intermediary capitals	Armenia (Q328518), Manizales (Q235190) and Montería (Q852725)
4.	Touristic cities	Cartagena (Q657461), San Andrés (Q134678), Providencia (Q681111), Leticia (Q214913), Puerto Nariño (Q767738), Santa Marta (Q209016) and El Cocuy (Q1655984)
5.	Villages	Villa de Leyva (Q1409503), San Gil (Q1294128), Barichara (Q1576773) and Zipaquira (Q205429)
6.	Small towns	Mocoa (Q579803), Villavicencio (Q749224), Soledad (Q767071), Pasto (Q320015) and Neiva (Q638260)
7.	Towns in Conflict	Cúcuta (Q216847), Medellín, Arauca (Q626543), Paraguachón (Q6060938), Corinto (Q2236398), Piendamó (Q2433349), Santander de Quilichao (Q1093175), Caldono (Q1391900), Jambaló (Q1525285), Puerto Tejada (Q1256377), Toribío (Q1870972)

[4] https://es.wikipedia.org/wiki/Plantilla:Principales_ciudades_de_Colombia

We make use of the Wikidata SPARQL endpoint[6] and Wikidata Mediawiki API[7] to collect the relevant information for our analysis.

3 Urban Information of Colombia on Wikidata

Table 3 gives a list of all the properties (total 96) currently used by the contributors to describe the cities above. It also shows the property identifiers and the associated data types. We see the use of 9 out of 17 supported data types. The table also shows the use of 36 external identifiers to various external data sources including national libraries and encyclopedic entries.

Figure 1 shows the number of statements used by the selected cities. We see the higher number of statements for Bogotá, Medellín, Cali and Cartagena. In Figure 2, we explore the number of distinct properties used by the different cities. We see Bogotá using 75 (out of 96) distinct properties.

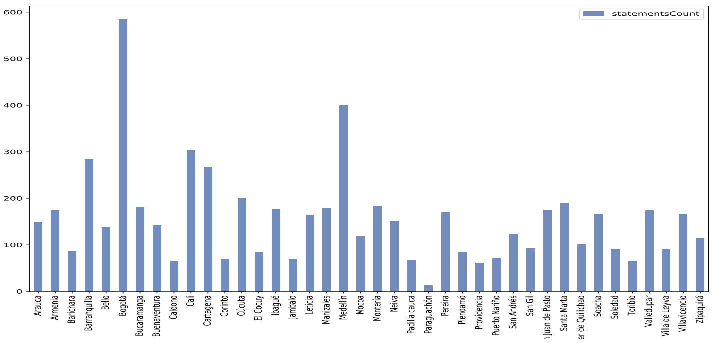

Fig. 1. Number of statements for different cities

Property values can be multi-valued. We compare the number of distinct properties versus the number of statements in Figure 3.

It's very important that all of these statements have associated references. Figure 4 shows the number of references used on the different Wikidata items. Unlike previous figures, Medellín is taking the lead. We compare the number of statements and the number of references in Figure 5. We see that in a majority of the cities, the number of references do not even cross 50%.

Next in Figure 6, we look at the number of Wikidata items that make use of the above cities in their statements, i.e., as a property value in the subject-property-value triplet. Here, Bogotá with 7048 inbound links is significantly ahead of the next city Medellín with 2659 links. Similar observation can be found

[6] https://query.wikidata.org/

[7] https://www.wikidata.org/w/api.php

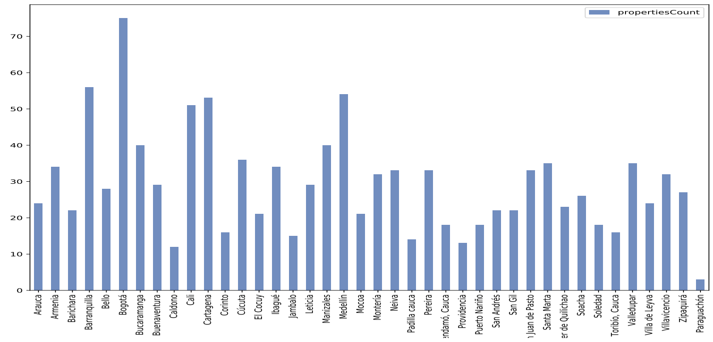

Fig. 2. Number of distinct properties for different cities

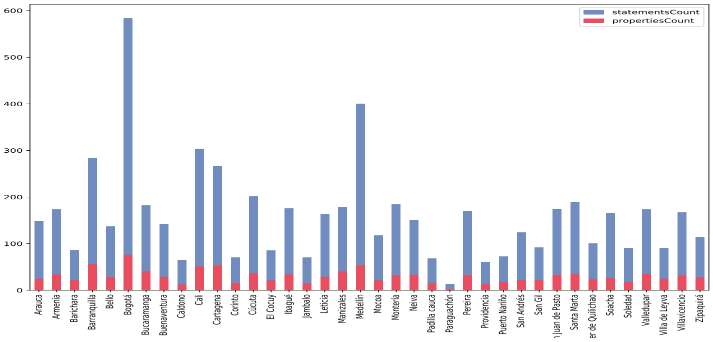

Fig. 3. Number of statements versus number of distinct properties for different cities

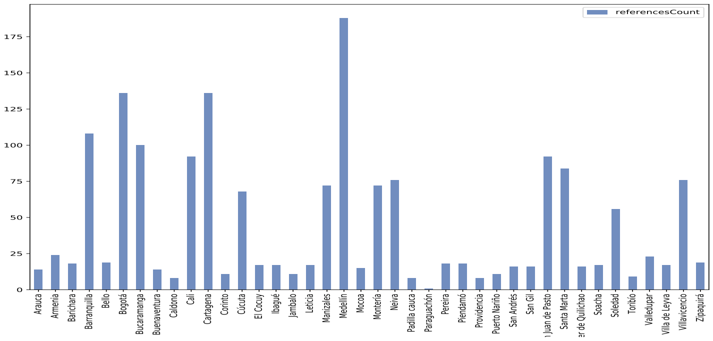

Fig. 4. Number of references for different cities

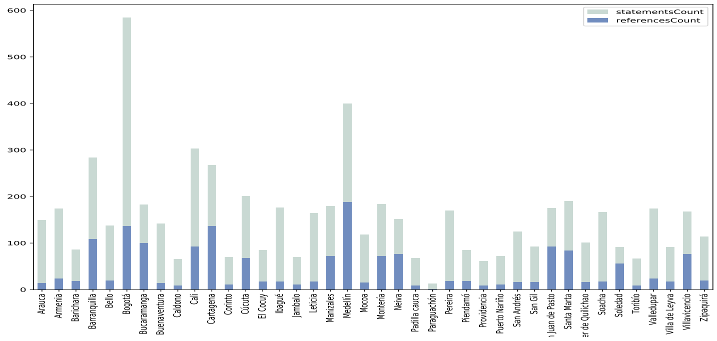

Fig. 5. Number of statements versus number of references for different cities

in the number of multilingual Wikimedia articles (on several projects including Wikipedia, Wikivoyage, Wikispecies etc.) of these cities with Bogotá taking the lead followed by Medellín (See Figure 7 and Figure 8).

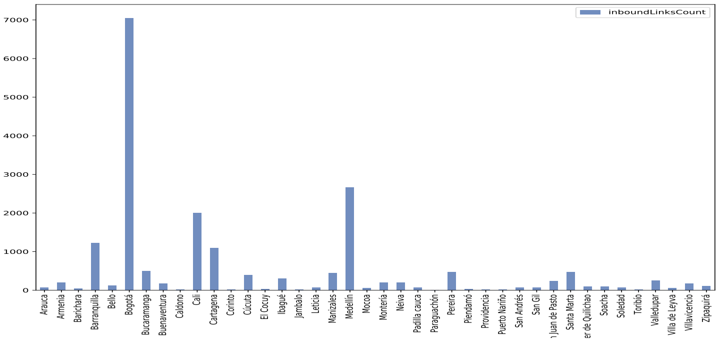

Fig. 6. Number of inbound-links for different cities

We also look at the most commonly used languages by these Wikimedia articles in Figure 9. It may not be surprising to see English taking the lead, followed by Spanish, French, Portuguese etc.

As discussed above, thanks to the use of external identifiers by Wikidata items, users can verify information from other external sources. In Figure 10, we see the number of external identifiers used by the cities.

Finally we take a look at the multilingual labels, descriptions and aliases on Wikidata in Figure 11 first by comparing them and then individually analysing them in Figure 12.

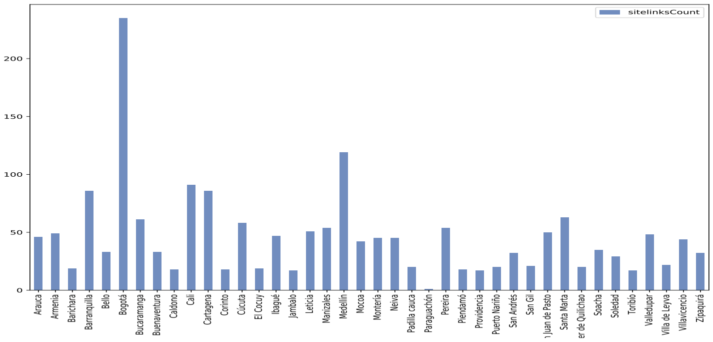

Fig. 7. Number of Wikimedia articles for different cities

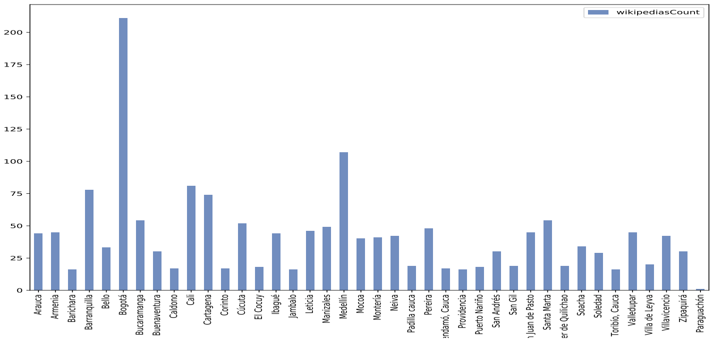

Fig. 8. Number of Wikipedia articles for different cities

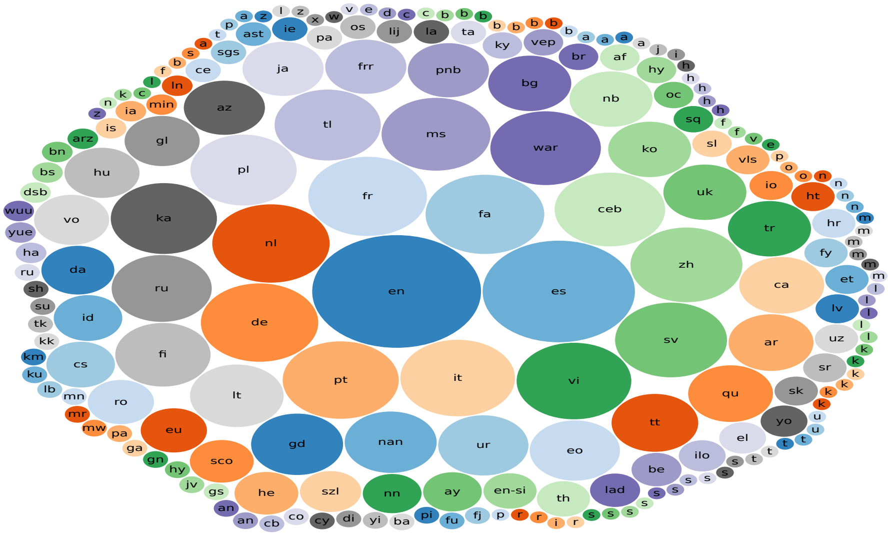

Fig. 9. Bubble graph on the number of articles of the selected cities in different languages

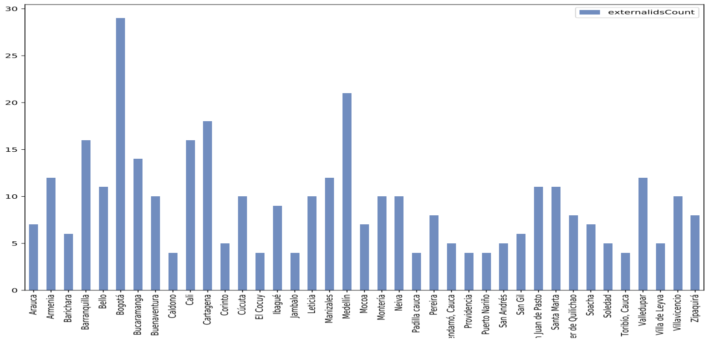

Fig. 10. Number of external identifiers for different cities

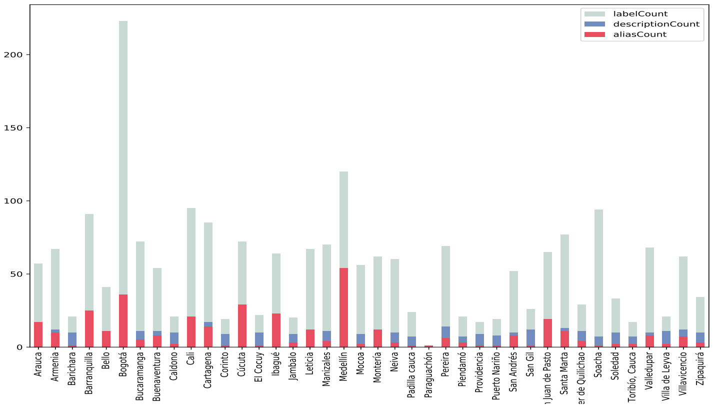

Fig. 11. Number of multilingual labels, descriptions and aliases of different Colombian cities

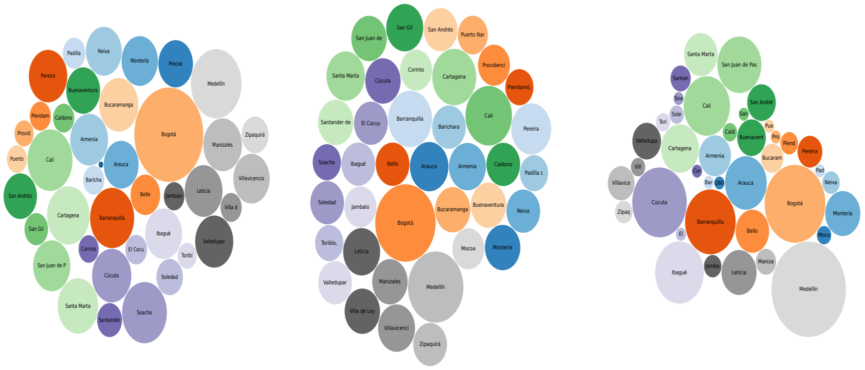

Fig. 12. Bubble graph on the number of multilingual labels, descriptions and aliases respectively for the selected cities

Current usage of images for these cities is also not complete. Only 33 of the above cities have an associated image (i.e., make use of the property P18). Also approximately 300 of the edits were reverted (detected by making use of "undo", "reverted", "Undid") pointing to possible vandalism attempts.

4 Discussion

We see the detailed statistics on translation of labels, description and aliases in Table 4. Statistics related to the number of statements, references, inbound Links and Wikimedia articles are presented in Table 4.

It's interesting to observe that four major cities: Bogotá, Medellín, Cartagena, Cali have significantly good number of statements and references. 96 distinct properties are used by the selected Colombian cities, towns and villages. Bogotá uses 75 distinct properties, compared to Paris (131), London (109), New York (108) and Berlin (129). No property using the Wikidata datatype GeoShape is currently being used. A total of 36 external identifiers to various external data sources including national libraries and encyclopedic entries are currently used. Bogotá has 29 distinct external identifiers, compared to Paris (58), London (57), New York (50) and Berlin (59).

It's interesting to compare the statistics between different cities in an interactive manner so that new contributors can explore how to improve different items. WikiProvenance[8] provides such a limited feature. However, it's unable to compare a large number of cities as explored in this article.

5 Conclusion

The major potential of Wikidata is its multilingual and structured nature that can be used to build personal assistants that can respond to user queries in their local languages. However, in a collaborative environment like Wikidata, it's also very important to have periodic monitoring of statements, translations and references needed in order to obtain accurate information from these assistants.

Another major direction is to have a greater participation from public-private partnerships including tourism agencies. For this purpose, generic and simpler tools to monitor urban data information as well as to understand and detect possible vandalism events need to be explored.

[8] https://doi.org/10.5281/zenodo.1252427

Table 1. Properties used on Wikidata for urban information

Datatype	Properties
CommonsMedia	image (P18), flag image (P41), coat of arms image (P94), seal image (P158), locator map image (P242), pronunciation audio (P443), page banner (P948), spoken text audio (P989), collage image (P2716) nighttime view (P3451)
ExternalId	ISNI (P213), VIAF ID (P214), GND ID (P227), Library of Congress authority ID (P244), Bibliothèque nationale de France ID (P268), SUDOC authorities ID (P269), ISO 3166-2 code (P300), OSM relation ID (P402), Freebase ID (P646), NKCR AUT ID (P691), FIPS 10-4 (countries and regions) (P901), SELIBR ID (P906), Biblioteca Nacional de España ID (P950), MusicBrainz area ID (P982), DMOZ ID (P998), AAT ID (P1014), BAV ID (P1017), US National Archives Identifier (P1225), WOEID (P1281), Gran Enciclopèdia Catalana ID (P1296), Encyclopædia Britannica Online ID (P1417), GeoNames ID (P1566), TGN ID (P1667), Global Anabaptist Mennonite Encyclopedia Online ID (P1842), Facebook Places ID (P1997), GRID ID (P2427), Great Russian Encyclopedia Online ID (P2924), Encyclopædia Universalis ID (P3219), NE.se ID (P3222), Quora topic ID (P3417), archINFORM location ID (P5573), Petit Futé site ID (P5760), Dizionario di Storia Treccani ID (P6404), Who's on First ID (P6766), ROR ID (P6782), DANE code (P7325)
GlobeCoordinate	coordinate location (P625)
Monolingualtext	official name (P1448), demonym (P1549), native label (P1705), short name (P1813)
Quantity	population (P1082), length (P2043), elevation above sea level (P2044), area (P2046), width (P2049)
String	postal code (P281), Commons category (P373), licence plate code (P395), local dialing code (P473), Commons gallery (P935), Commons maps category (P3722)
Time	inception (P571)
Url	official website (P856)
WikibaseItem	head of government (P6), country (P17), instance of (P31), official language (P37), shares border with (P47), founded by (P112), located in the administrative territorial entity (P131), contains administrative territorial entity (P150), flag (P163), twinned administrative body (P190), legislative body (P194), located in or next to body of water (P206), executive body (P208), coat of arms (P237), part of (P361), located in time zone (P421), said to be the same as (P460), opposite of (P461), topic's main category (P910), topic's main Wikimedia portal (P1151), office held by head of government (P1313), described by source (P1343), capital of (P1376), present in work (P1441), category for people born here (P1464), category for people who died here (P1465), category of people buried here (P1791), category of associated people (P1792), owner of (P1830), different from (P1889), history of topic (P2184), on focus list of Wikimedia project (P5008)

Table 2. Translation of labels, description and aliases on Wikidata for urban information of selected Colombian cities

Measure	Labels	Description	Alias
count	40	40	40.000000
mean	54.70000	11.500000	9.375000
std	38.87996	5.223222	11.331118
minimum	1	0	1
25%	21.75	9.75	2.00
50%	56.50	10.50	4.50
75%	69.25	12.00	12.000
maximum	223	32	54

Table 3. Statements, References, Inbound Links, Wikimedia articles on Wikidata for urban information of selected Colombian cities

Measure	Statements	References	Inbound Links	Wikimedia articles
count	40.000000	40.000000	40.000000	40.000000
mean	154.775000	43.450000	480.625000	45.375000
std	102.644142	44.839857	1195.322158	39.173963
min	13.000000	1.000000	1.000000	1.000000
25%	85.750000	14.750000	49.500000	20.000000
50%	145.500000	18.000000	103.000000	38.500000
75%	176.750000	73.000000	332.000000	51.750000
max	584.000000	188.000000	7048.000000	235.000000

Bibliography

[1] Li Ding, Vassilios Peristeras, and Michael Hausenblas. Linked Open Government Data [Guest editors' introduction]. *IEEE Intelligent Systems*, 27(3):11–15, May 2012.

[2] Dustin Lange, Christoph Böhm, and Felix Naumann. Extracting structured information from Wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 1661, Toronto, ON, Canada, 2010. ACM Press.

[3] Marshall McLuhan and Bruce R. Powers. *The global village: transformations in world life and media in the 21th century.* Communication and society. Oxford Univ. Press, New York, 1992. OCLC: 845334715.

[4] Santiago M. Mola-Velasco. Wikipedia vandalism detection. In *Proceedings of the 20th international conference companion on World wide web - WWW '11*, page 391, Hyderabad, India, 2011. ACM Press.

[5] Joachim Neubert. Wikidata as a linking hub for knowledge organization systems? integrating an authority mapping into wikidata and learning lessons for KOS mappings. In *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017.*, pages 14–25, 2017.

[6] John Samuel. Analyzing and visualizing translation patterns of wikidata properties. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, pages 128–134, 2018.

[7] Thomas Pellissier Tanon and Lucie-Aimée Kaffee. Property label stability in wikidata: Evolution and convergence of schemas in collaborative knowledge bases. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1801–1803, 2018.

[8] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[9] Fei Wu and Daniel S. Weld. Automatically refining the wikipedia infobox ontology. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 635, Beijing, China, 2008. ACM Press.