Revista de Estudios Andaluces (REA) | Núm. 45, febrero (2023) | e-ISSN: 2340-2776
https://dx.doi.org/10.12795/rea.2023.i45 | © Editorial Universidad de Sevilla 2023
https://editorial.us.es/es/revistas/revista-de-estudios-andaluces
https://revistascientificas.us.es/index.php/REA

SUMMARY OF ARTICLE: HTTPS://DX.DOI.ORG/10.12795/REA.2023.I45.11

Methodology for the incorporation of geographic information in Wikidata

Ángel Obregón-Sierra
angel.obregon@ui1.es 0000-0001-8801-317X
Javier López-Otero
Javier.lopez.otero@ui1.es 0000-0002-6543-2926
Antonio Gavira-Narváez
antonio.gavira@ui1.es 0000-0002-5389-8315
Rafael Vega-Pozuelo
rafaelfernando.vega.pozuelo@ui1.es 0000-0003-4982-9285

Universidad Isabel I. Calle Fernán González, 76. 09003 Burgos, España.

KEYWORDS

Wikidata
Open data
SPARQL
GIS
Spatial analysis
Triangulation station

1. INTRODUCTION

The changes brought about by the present information society and, more specifically, the advance of the new information and communication technologies of the fourth industrial revolution[1], have multiplied the traffic of software, documents, maps and manufacturing systems (Stark et al., 2006). It's worth noting the development of artificial intelligence (AI) algorithms, that consume a large amount of information to train their models (González & Evans, 2019). Likewise, there is a growing development of 2D and 3D modeling of the territory and objects that consume huge amounts of space (Shan & Sun, 2021). In this sense, the existence of information repositories has become increasingly necessary. However, the management and organization of the information repositories has not been homogeneous and, in fact, the data is distributed in a multiplicity of repositories of public and private property in which there is abundant unconnected information.

Against this prevailing model, it is worth noting the Wikidata project. This has the particularity of being a free knowledge base owned by the Wikimedia Foundation. Wikidata was created to support the rest of the foundation's projects, such as Wikipedia or Wikimedia Commons. In addition, as it is a collaborative project, it also serves as a free knowledge base that can be accessed and edited by any user with Internet access (Obregón, 2022). All of this has multiplied the use of Wikidata. However, despite the great growth of the platform, there are few studies on the presence and treatment of geographic information about this repository, among other reasons, because there is no explicit methodology for the inclusion of some data.

1. This concept refers to the most important revolution to date, the one that allows uniting technologies in test or development that help to blur the existing borders between the physical, biological and digital spaces.

DOI: https://dx.doi.org/10.12795/rea.2023.i45.11
Formato de cita / Citation: Obregón-Sierra, A. et al. (2023). Methodology for the incorporation of geographic information in Wikidata. *Revista de Estudios Andaluces,* (45), 207-230. https://dx.doi.org/10.12795/rea.2023.i45.11
Correspondencia autores: Javier.lopez.otero@ui1.es (Javier López-Otero)

Consequently, this paper proposes a specific methodology for geography that allows the incorporation of geographic data into Wikidata, which can later be analyzed using spatial analysis tools, such as a GIS. This methodology and analysis will be used with a database of triangulation stations of Spain.

2. METHODOLOGY FOR UPLOADING THE TRIANGULATION STATIONS OF SPAIN TO WIKIDATA

The proposed methodology includes 6 phases that allow the introduction of geographic information in Wikidata, in large quantities and subsequently have it available for spatial analysis. These phases are outlined below:

2.1. Selection of the origin of information

Prior to entering the data on the platform, it is convenient to identify four essential aspects of it, such as its public nature, its structure, the nature of the geographic data (vector or raster) and the availability of the data in the platform.

Therefore, only data that has a Creative Commons license or is in the public domain is compatible with Wikidata. Likewise, the geographic information to be stored in Wikidata must be vectorial, while the raster information must be entered in the Wikimedia Commons multimedia file store. For this reason, we will work with vector information in this proposed methodology, specifically with elements identified as points which have spherical coordinates.

The application of this methodology requires entering data that does not exist in the knowledge base, so it is convenient to previously review the data available at that time in Wikidata. To do this, a query was made in Wikidata, verifying the existence or not of the geographic information to be uploaded, in order to avoid the generation of duplicates. This query can be made from the Wikidata Query Service (https://query.wikidata.org), that is, the query service of Wikidata that uses a query language called SPARQL, which is a standardized language similar to SQL, although adapted to the query of RDF (Resource Description Framework) graphs.

Once verified that the National Geographic Institute (IGN) page has several independent repositories on triangulation stations and that this information has a CC-BY 4.0 compatible license, all the information was unified in a single table. Specifically, 11,127 triangulation stations were identified.

2.2. Data cleaning

This phase consists of correcting or deleting records, for which it is necessary to know the structure of the destination platform, where the data will be inserted, since it imposes the way in which the information will be stored. This requires knowledge of the formats that Wikidata accepts. Specifically, there are two essential components: the elements (start with a Q followed by a number) and the properties (start with a P followed by a number).

The adaptation of the content of the previous unified table to what Wikidata requires has been done with a spreadsheet and OpenRefine, whose purpose is to clean the data and provide the corresponding identifiers of Wikidata. Finally, the values to upload to Wikidata are shown in the following table 1.

Table 1. Data that was worked with in OpenRefine.

Column in table	Property in Wikidata (WD)	Example (Q115498684)	Element in WD
Num	P528	2355	
Name	Label	A Carba	
Descripción en español	Descripción	vértice geodésico en Vilalba, España	
Descripción en inglés	Descripción	triangulation station in Vilalba, Spain	
Alias	También conocido como	Carba	
Instance of	P31	Vértice geodésico	Q131862
Country	P17	España	Q29
Municipality	P131	Villalba	Q1605437
Province		Ourense	
REGENTE (network)	P361	Regente	Q115497793
Unavailable	P5817		
Coordenates	P625	43.4203692, -7.6586297	
Height (above the sea)	P2044	907,99	
Height of the mark	P2048	1,2	
URL of origin	P973	https://datos-geodesia.ign.es/Red_Geodesica/Hoja0023/002355.pdf	

Source: Authors (2023).

3. INSERTION

The storage of massive information makes it impossible to save thousands of data in a knowledge base one by one, it is necessary to use tools or programming languages that allow it to be done at a large scale. So, once the data has been cleaned with the spreadsheet and OpenRefine, and after adjusting it to the information contained in Wikidata, it is necessary to create a schema, that is, a Wikidata edits template that is applied to each row to be inserted.

Once the schema was completed with all the columns to be entered, some inconsistencies in the data to be entered were identified and corrected. Subsequently, the information was uploaded and exported to QuickStatements.

4. REVIEW

The data entered may contain erroneous data, even if the cleanup was successful. Therefore, queries must be made to the knowledge base, in order to check if there are odd values, duplications or missing information.

QuickStatements may duplicate the data inadvertently. Therefore, once the insertion in Wikidata is finished, queries must be made to the knowledge base, to check that all the elements contain the correct information. The query can be viewed at the following link in the Wikidata Query Service: https://w.wiki/69yR.

5. VISUALIZATION AND ANALYSIS

The storage of the data in itself does not contribute anything, it is only useful when it is available for visualization in tables or in maps. This information can be consulted in the Wikidata Query Service, these queries must be formulated in the SPARQL language.

However, the spatial information to be consulted can also be entered into a GIS, thus multiplying the spatial analysis options that allow going beyond simple visualization. Specifically, in this work the use of the QGIS plugin "SPARQLing Unicorn" is proposed, which allows queries to be made in Wikidata using the SPARQL query language.

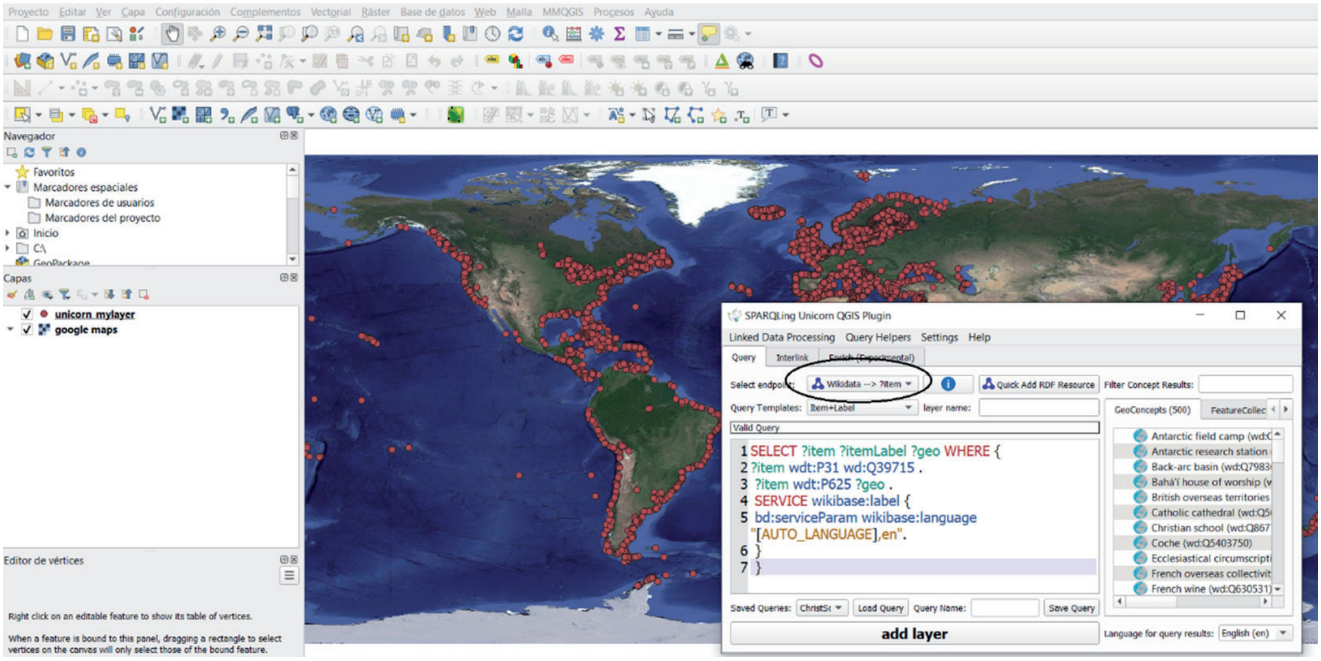

Figure 1. Example query in the SPARQLing Unicorn. Source: Own elaboration from Wikidata (2022).

6. REPLICATION

Once the methodology has been completed it is desirable to review whether the same process can be repeated with similar data. This would be the last step and its purpose is to validate the methodology, verifying that the procedure is equally valid and applicable to other data with similar characteristics.

7. RESULTS

Once the insertion and review were completed, it was verified that 11143 triangulation stations were updated or inserted, which can be seen at the following link: https://w.wiki/64t6. It can be consulted data such as the triangulation station's elevation above sea level (https://w.wiki/64yH) and other similar characteristics.

This information can be displayed in an interactive map using a double method, either through the Wikidata Query Service application https://w.wiki/6Dg8, or through the QGIS SPARQLing Unicorn plugin https://slovenianman.github.io/geodesic_vertices/.

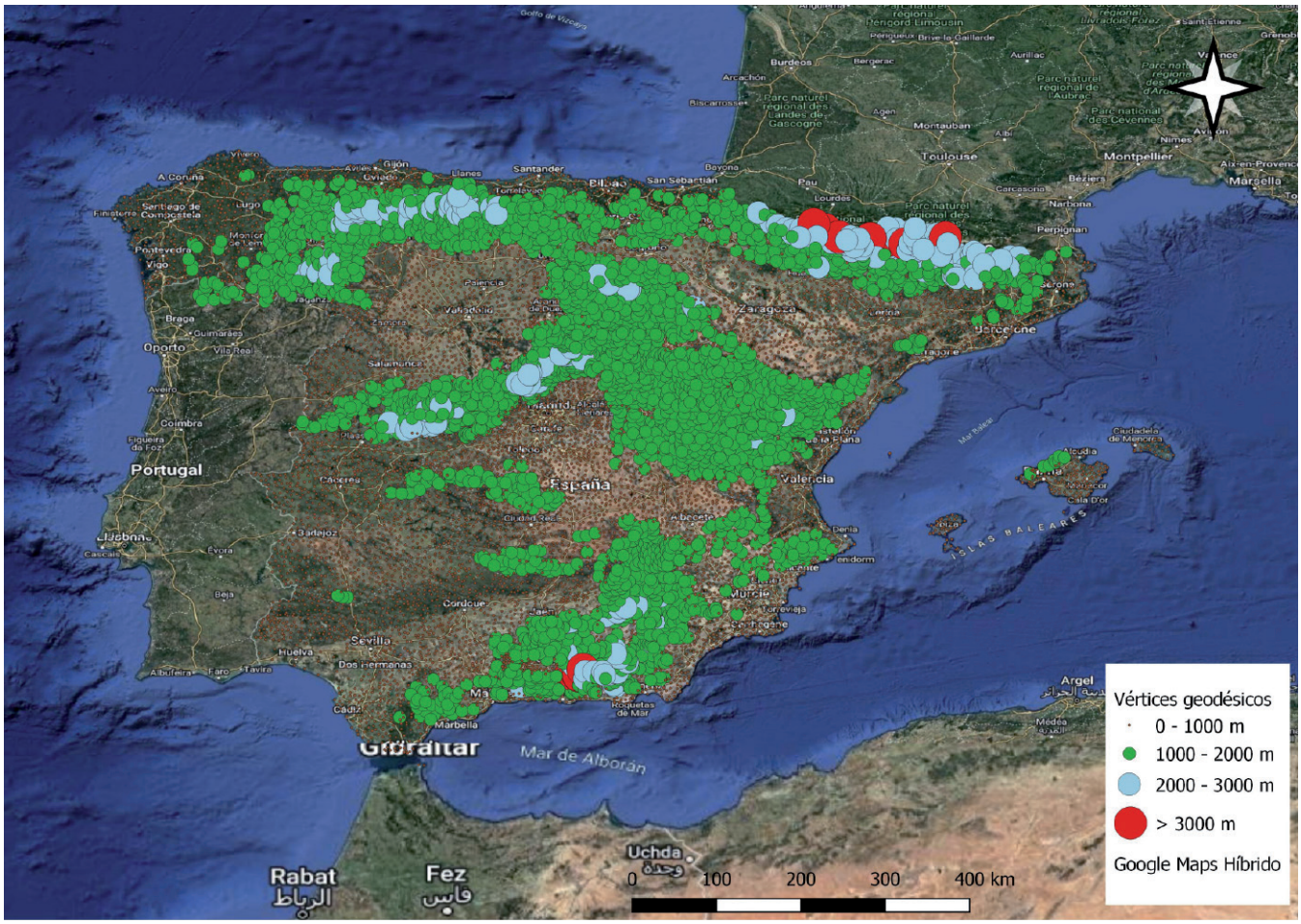

Figure 2. Map obtained with the data uploaded to Wikidata and processed with QGIS. Source: Self made (2023).

8. DISCUSSION AND CONCLUSIONS

Therefore, this methodology is accessible to researchers who are not specialized in programming, making new geographic information available in Wikidata for the scientific community as a whole.

Likewise, this geographic information can be easily entered into a GIS through a QGIS interface plugin.

Another notable aspect of this methodology is that it is capable of introducing large amounts of geographic information into the platform, consistent with the growing need for data by modeling algorithms to make their predictions.

Finally, the possibility of working from QGIS or ARCGIS with massive and public repositories such as Wikidata favors the analysis process by increasing the availability and accessibility of the data. In this regard, the fact that the geographic data used to carry out the analyzes are public and free will make it easier for any researcher or reviewer to replicate the methodology of an author and verify the veracity of the results obtained. Likewise, we will have the possibility of carrying out derived or alternative analyses, which could enrich or alter the conclusions obtained.