information

MDPI

Article

A Proximity-Based Semantic Enrichment Approach of Volunteered Geographic Information: A Study Case of Waste of Water

Liliane Soares da Costa [1,*], Italo Lopes Oliveira [2], Alexandra Moreira [1] and Jugurta Lisboa-Filho [1]

[1] Departamento de Informática, Universidade Federal de Viçosa (UFV), Viçosa, MG 36570-900, Brazil; xandramoreira@yahoo.com.br (A.M.); jugurta@ufv.br (J.L.-F.)

[2] Departamento de Informática e Estatística, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC 88040-900, Brazil; italo.oliveira@posgrad.ufsc.br

* Correspondence: lilianesoaresnc@gmail.com; Tel.: +55-031-3612-6358

Received: 22 May 2019; Accepted: 26 June 2019; Published: 8 July 2019

Abstract: Volunteered geographic information (VGI) refers to geospatial data that is collected and/or shared voluntarily over the Internet. Its use, however, presents many limitations, such as data quality, difficulty in use and recovery. One alternative to improve its use is to use semantic enrichment, which is a process to assign semantic resources to metadata and data. This study proposes a VGI semantic enrichment method using linked data and thesaurus. The method has two stages, one automatic and one manual. The automatic stage links VGI contributions to places that are of interest to users. In the manual stage, a thesaurus in the hydric domain was built based on terms found in VGI. Finally, a process is proposed, which returns semantically similar VGI contributions based on queries made by users. To verify the viability of the proposed method, contributions from the VGI system Gota D'Água, related to water waste prevention, were used.

Keywords: volunteered geographic information; semantic enrichment; linked data; thesaurus

1. Introduction

The development of the web has led to more widespread and visible content. However, the rapid growth in the amount of data available resulted in the need to include more semantics in websites, leading to the creation of the semantic web. In order to enable the implementation of the semantic web, new technologies have been developed (e.g., linked data) and existing technologies were incorporated (e.g., thesauri, ontologies, semantic bases). The addition of semantics and the incorporation of technologies allowed the evolution of search and research mechanisms on the web, which innovates the way data are discovered, accessed, integrated, and used [1].

The semantic web aims to provide a space of shared semantic information, qualitatively changing experiences on the web [2]. With the help of such technologies, user participation on the Internet has become more active and enabled content creation besides the adaptation and development of applications for varied fields. One of the fields that benefit from greater user participation on the internet is geographic information.

Citizens (users) are using mobile devices to collect geographic information using web-based mapping interfaces to tag and annotate geographic characteristics such as adding geotags to photographs. These actions originated the term Volunteered Geographic Information (VGI) [3]. However, discovering and properly using VGI data still face several challenges, such as ambiguity in terms employed by the user and precision of geographic coordinates. One of the ways to mitigate these issues is by adding semantics to VGI using linked data.

The linked data concept arose to aid in the discovery, access, and use of online data. Linked data are semi-structured data that allow specifying semantic relationships among themselves. Using such concept, data on the internet become semi-structured nodes of a semantic network [1]. The semantic relationships expressed by linked data facilitate their discovery across different data repositories, besides allowing semantic searches to the performed [4].

Given the use of semantic relationships to relate data and, therefore, improve sharing, the linked data concept has been suggested as an approach in semantic enrichment of VGI [5,6]. Nonetheless, manually adding semantics to VGI is a costly, tedious, error-prone task [7]. Requiring users to describe the semantics and semantic relationships of previously produced volunteered geographic data is impractical either due to the lack of user knowledge on how to properly specify data semantics or due to the burden of this task, which would prevent users from producing new geographic data. Therefore, an automated process is required to semantically enrich large amounts of VGI in an attempt to meet the needs of a specific application domain and provide a more thorough description of user-generated data. However, existing works focus on the textual data of the VGI contributions to semantically enrich them [8] or proposes a semi-automatic to tackle this task [9].

This paper proposes a method to semantically enrich VGI that present little textual data and it is based on spatial relations of proximity. The VGI semantically enriched by the proposed approach can be used to build a thesaurus, which further contributes to add semantics to the data. From this, semantic analyses can be performed on such data and, consequently, improve their discovery and use in the semantic web. Moreover, adding semantics to VGI allows finding or solving inconsistencies and ambiguities, thus improving the quality of user-generated data. In our case, a VGI system with contributions related to water waste called Gota D'Água (water drop, in free translation) was used as study case to demonstrate analyses with the contributions that are only possible after the semantic enrichment process.

The remainder of the paper is structured as follows. Section 2 makes a review of the main concepts involved in the study. Section 3 presents some related works. Section 4 presents the method proposed. Section 5 describes the study case. Section 6 discusses the results obtained and, finally, Section 7 makes some final considerations.

2. Theoretical Framework

This section presents the bases required to understand the remainder of the paper. Section 2.1 describes the VGI characteristics. Sections 2.2 and 2.3 present, respectively, the concept of linked data and the semantic enrichment based on such concept.

2.1. Volunteered Geographic Information

The expansion in sources of information as a result of networked computers and other interconnected devices led to significant changes in the amount, availability, and nature of geographic information. Among the most significant changes is the growing amount of readily available geographic information produced voluntarily [10]. It is visible that the current media environment has the ability to promote, support, and sustain collective effort among individuals. Through more dynamic web systems, users play the roles of information consumer and information provider [2]. Consequently, this structure is particularly appropriate to collaboration among individuals.

Data with some spatial characteristic generated by some sort of volunteered contribution are considered VGI [11]. The marked increase in VGI contributions leads to several new platforms and projects that use data and technologies in spatial decision-making, participatory planning, and citizen science. One such example is OpenStreetMap (OSM), a collaborative mapping project that aims to create a set of geographic data that is free for users to use and edit [12]. It allows users to download geospatial data free of charge and use them in personal projects [13]. Projects such as OSM feature tools able to capture, display, produce, and spread information on a larger scale, besides enabling experiences for several purposes.

The potential use of VGI has been proven for urban management, damage from floods, fires, and earthquakes, and other important cases of risk, crisis, and natural disaster management [14]. VGI systems may help prevention, carrying out actions during a phenomenon, or recovery of a region after a natural disaster. Therefore, some of the advantages of volunteered information include it being free, being possibly obtained quickly, and generating data that have not yet been made available by official agencies. However, its content lacks quality assurance [15]. In this way, it is advisable to use methods to assess and ensure the quality of VGI content.

There is notable concern with VGI quality given the large amount of data provided by different individuals [10,16]. Those concerns are also related to the lack of quality oversight during the data creation process. However, several authors [13,17–19] have compared volunteered data with data from official agencies and concluded that the quality level of VGI data is closing the gap to the level of data from official agencies, particularly in densely populated urban areas. Moreover, VGI can be enriched for improved quality. One way of improving such data is by using linked data and through the semantic enrichment process [5].

2.2. Linked Data

Linked data refers to a style of publishing and interlinking semantic data on the web. The data are made available online in a way that they are easily processed by machines and their meaning/semantics are defined explicitly. In addition, these data are linked bidirectionally to other external datasets [1]. The goal is to allow people to share data with well-defined semantics on the web as easily as documents are currently shared. Therefore, it targets massively sharing and reusing information in a global data space, besides allowing new data to be discovered [20,21]. The more a piece of data is interlinked to other data, the greater its value and usefulness.

Some linked data principles were introduced by [22], namely: using uniform resource identifiers (URIs) as resource names; using URIs with HTTP (URLs) so that people are able to find those names; when someone searches for a URI, ensure that useful information can be obtained through those URIs, which must be represented in RDF format; and including links to other URIs so that other resources can be discovered.

One of the most commonly employed formats for linked data structuring is the Resource Description Framework (RDF, https://www.w3.org/RDF/) format. RDF is a generic data model based on graph structure that describes things in the world. The RDF model codes data as triplets: subject, predicate, object. The subject and object of a triplet are both URI that identify each resource, while the object may take on the value of a string. The predicate specifies how the subject and object are related and is also represented by a URI of a resource that describes such relationship [1].

One of the most well-known example of an open linked data repository is the DBpedia (http://wiki.dbpedia.org/) project [23], a semantic wiki based on Wikipedia. DBpedia provides its information in RDF format, which allows queries to be made. Such queries may be performed using the SPARQL (http://www.sparql.org/) language.

2.3. Semantic Data Enrichment

As the semantic web advances, systems that allow users to generate web content through real-world experiences (e.g., social media, VGI systems) are seen as systems that will strongly impact informal learning. However, such contents lack a well-defined learning domain and are normally written in natural language. In order for information to be comprehensible and workable by a computer agent, and adequate for inclusion in the semantic web, techniques and tools that provide a minimum understanding of the content must be used [24].

For that to be possible, it is important to resort to semantic enrichment, which can be described as the process of attributing greater meaning to metadata and data by applying auxiliary resources aiming to facilitate understanding, integration, and processing of data by people and machines. Retrieval systems that return more relevant results make users more productive and make the content found

more useful. Thus, semantic enrichment is seen as a way of improving the result of a search. Despite the benefits of adding semantics to data, issues such as cost, precision, and scalability must be taken into account before a semantic enrichment approach or process is adopted. A well-though semantic strategy will minimize the costs of the process [25].

Some resources and techniques are used for semantic enrichment and to obtain additional concepts. An important technique is semantic annotation, which consists in attributing meaning (semantics) to the elements of a scheme of origin, which, according to [25], can be done manually or in an automated manner.

Manually, a person with proper knowledge reads the content and applies the annotation. Manual annotation is ideal when its use requires a high degree of precision. However, its cost can be prohibitive for large volumes of content since it is labor-intensive and hard to scale for large volumes of work.

In automated annotation, the software analyzes the content and adds annotations based on correspondence of concepts, statistical patterns, and linguistic analysis. Most automated systems use algorithms to adequate to a dataset and a specific field of knowledge, thus increasing the level of precision that can be reached through automation. Automated annotation is highly scalable and is sometimes the only option for very large datasets (e.g., social media posts). However, automated approaches may lead to false positives (e.g., tags applied to non-relevant parts of the data), lost concepts, and other such imprecisions [26].

3. Related Works

According to [8], integrating a VGI set to the Linked Open Data (LOD) cloud provides advantages beyond only resources that are directly interlinked with the VGI set. That is because of the great interconnectivity between the set of data published as linked data. In that study, ref. [8] seeks to investigate to what extent the LOD cloud could help semantically enrich VGI in order to obtain better search results in the context of operations and crisis. Based on that, it is said that the use of URIs of an open knowledge database such as LinkedGeoData (LGD) eliminates the possibility of ambiguity when indicating a place. The additional semantics obtained from structured information presented in LGD enables, for example, quick access to basic and useful knowledge on infrastructure objects and public buildings, which could be used to speed up and improve the decision-making process. Therefore, integrating VGI with relevant entities on the LOD cloud makes it possible to semantically enrich unstructured user-generated content with structured information presented in LOD.

Sorrentino et al. [9] propose interlinking semantically enriched data with linked data repositories through semi-automated experimental methodology to aid resource providers in publishing public data on LOD repositories and to help consumers (companies and citizens) efficiently access and query them. The author presents a method for publishing, linking, and semantically enriching open data by performing automated semantic tagging of scheme elements. The method was applied to a dataset provided by the Research Project on Youth Prevention, in Italy, which investigates the precarious situation of young people living in the Modena district.

Semantic enrichment derived from knowledge on the relationships among geospatial data is a potential source of solutions to traditional issues when recovering geographic information [27]. This way, ref. [27] discusses the applicability of linked data concepts to several issues, such as resolving ambiguity and recognizing the spatial context of documents.

This work propose a process to semantically enrich VGI contributions and produce thesauri from these semantic contributions. Unlike the works mentioned, specifically for the semantic enrichment step, this study proposed a fully automated method to links the contributions to places close to them. Moreover, our method does not depend on textual descriptions, lacking in several VGI data (and our study case). To demonstrate that the semantically enriched VGI data can be discovered in the semantic web, we execute several SPARQL queries on the semantic contributions. Lastly, we present the steps necessary to build thesauri from the semantic contributions manually.

4. A Method for VGI Enrichment with Linked Data and Thesaurus Creation

This section presents the process proposed for semantic enrichment of VGI with linked data. The process aims to semantically enrich VGI contributions based on (i) their geographic coordinates while annotating the contributions with possible places of interest (PoI) for the user in the context in which the contribution is found and (ii) the texts of the contributions, attempting to find relevant words and to attribute appropriate meaning to generate a thesaurus.

Figure 1 presents a flowchart of the process proposed. The process inputs are the VGI contributions to be semantically enriched and a database (e.g., knowledge base, LOD repository) containing the PoIs of the region to which the geographic coordinates of the contributions point. Both the VGI contributions and the database containing the PoIs undergo a pre-processing step, where noise (e.g., grammar mistakes, time and space inconsistencies), duplicates, and other issues in the data are corrected.

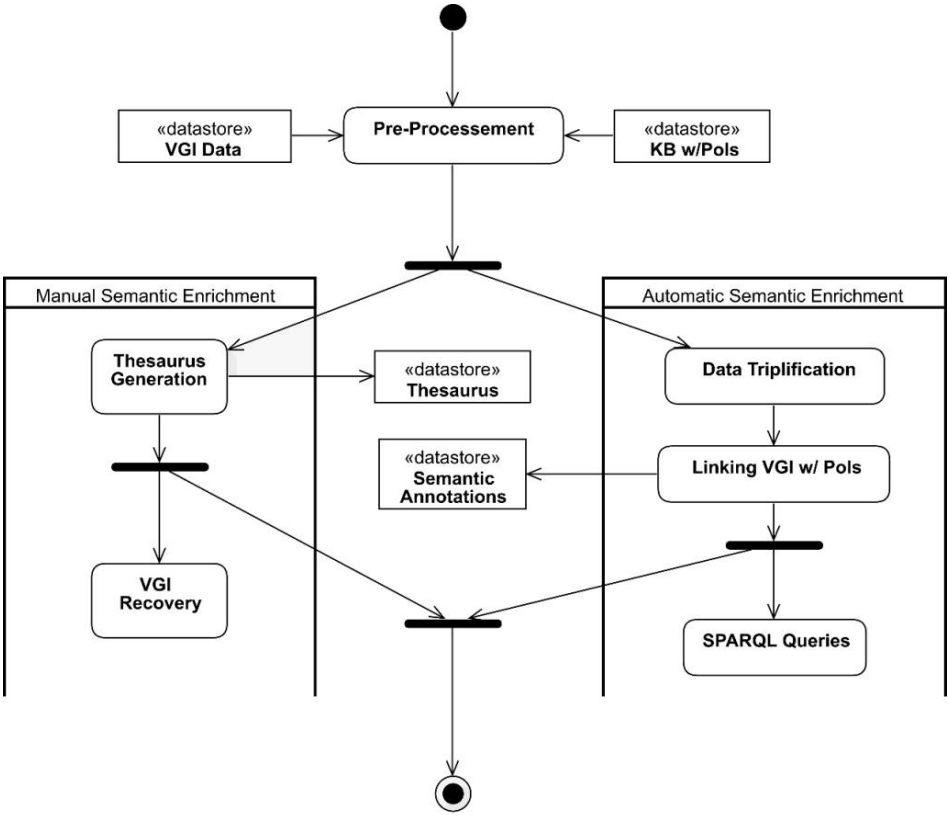

Figure 1. Flowchart of the volunteered geographic information (VGI) semantic enrichment process proposed.

After pre-processing, the VGI contributions are semantically enriched by two distinct tasks that may occur simultaneously, namely automated semantic enrichment and manual semantic enrichment. Manual semantic enrichment, in the scope of this research, consists of generating a thesaurus from the textual elements present in the VGI contributions. A process similar to that used in [28] was adopted to generate the thesaurus.

Thesaurus generation may be semi-automated with tools for disambiguation of relevant words (word sense disambiguation—WSD) and tools that identify mentions to real-world entities and link such mentions to semantically well-described resources that properly describe them in the context in which they are found (entity linking—EL tools) [29]. However, these tools still lack satisfactory results when the text to be annotated has little context, such as social media posts and VGI contributions. Thus, the use of such tools was not considered in the process during thesaurus generation.

Automated semantic enrichment consists in annotating VGI contributions with the PoIs of the user for the contribution context. In order to extract all the potential of semantic enrichment,

i.e., facilitate its discovery and use by applications, VGI contributions were triplified using the RDF standard. An algorithm was developed to perform the triplification. Although there are tools for data triplification, such as Triplify (http://aksw.org/Projects/Triplify.html), a simple algorithm was sufficient to perform the task given the simplified structure of the data used in the study case (Section 5). Moreover, due to its simplicity and as it is not the target of this study, the triplification algorithm is not presented.

After triplification of the VGI contributions, the next step of the process is the automatic semantic enrichment of the contributions, more specifically, the connection of the contributions with PoIs.

Algorithm for Automatic VGI Semantic Enrichment

Algorithm 1 consists in the algorithm proposed for linking VGI contributions of the users with PoIs. The algorithm uses geospatial operations to find the possible places of interest for the user. Therefore, the algorithm works with any library or database management system (DBMS) with support for geospatial operations among different geometries.

Algorithm 1 Algorithm for automated volunteered geographic information (VGI) enrichment with places of interest.

Require:
$V = \{v_0, \ldots, v_n\}$ // VGI set
$P = \{p_0, \ldots, p_m\}$ // Knowledge base that contains places of interest
$T_r \in R$ // Threshold in meters for spatial distance
Output:
SA // Initially empty set of semantic annotations
1: $SVP \leftarrow (\sqcap v \leftarrow V.cod, p \leftarrow P.cod,\ names,\ geoDist(V \bowtie (geoDist \leftarrow dist(v.geom,\ p.geom)) \leq T_r))$
2: **for** each $v \in V$ **do**
3: **for** each $(v,\ p,\ geoDist) \in SVP$ **do**
4: $R.AddAnnoration(v,\ p)$
5: **end for**
6: $SA.add(R)$
7: **end for**
8: **return** SA

The algorithm inputs consists of a set of VGI contributions V, a knowledge base P containing the PoIs to be used to enrich the VGI set V, and a value in meters T_r for the buffer size to be used in the geospatial operations. The output of the algorithm is a set SA of semantic annotations as RDF files and is initially an empty set. For each contribution $v \in V$, a circular buffer with radius T_r is generated around geographic position v. It can be noted that, although it is not present in the algorithm to approximate what was implemented, the buffer to be generated may assume several other geometries (e.g., rectangular, complex geometries) so as to better adapt to the needs of several applications. After the creation of the buffer, it is verified which places $p \in P$ are geographically within the buffer. For each of those places, a semantic annotation $r \in R$ is created, whose subject will be the URI for VGI v and whose object is the URI of PoI p. The predicate value will change depending on the type of issue described in v. For example, if the type of issue of a contribution v is "leak," the predicate value is "influence," whereas if the type of issue is "waste," the predicate value is "notify." After all PoIs p are annotated to v, annotations r_i contained in R are added to the SA set. Finally, set SA is returned.

To exemplify how the algorithm works, a VGI contribution was randomly selected from the database of the Gota D'Água system (used as study case and described in Section 5). When the algorithm is applied, a buffer with 100 m radius is created around the geographic coordinate of the contribution while verifying which of the PoIs in OSM are contained in this buffer. Table 1 shows which places (represented by yellow circles) are contained in the buffer region (represented by a blue circle) in Figure 2 around the place of the contribution. Figures 2 and 3 show the buffer enveloping

the businesses near the contribution in different representations. While Figure 2 is represented in OSM, Figure 3 is represented in Google Maps. Note that the choice of knowledge base containing the PoIs impacts the number of semantic annotations generated. As shown in Figures 2 and 3, due to the difference in the number of registered PoIs between OSM and Google Maps, the result returned by the algorithm significantly varies according to the PoI base.

Table 1. Places found within a 100 m radius from the geographic coordinate of a volunteered geographic information (VGI) contribution.

Place Code	Place Characteristics
2959454656	"name" → Lanchonete do Dênis "amenity" → "fast_food"
2959454657	"name" → "Bar Norte Mineiro" "amenity" → "bar"
2959454658	"name" → "Droga Farma" "amenity" → "pharmacy" "opening_hours" → "24/7"
2959454659	"name" → "Alfa Hotel" "tourism" → "hotel"
2959454660	"name" → "Orquídea Hotel" "tourism" → "hotel"
2966424164	"name" → "Clic Clic" "shop" → "art"
3412042504	"name" → "Subway" "amenity" → "fast_food" "cuisine" → "sandwich"

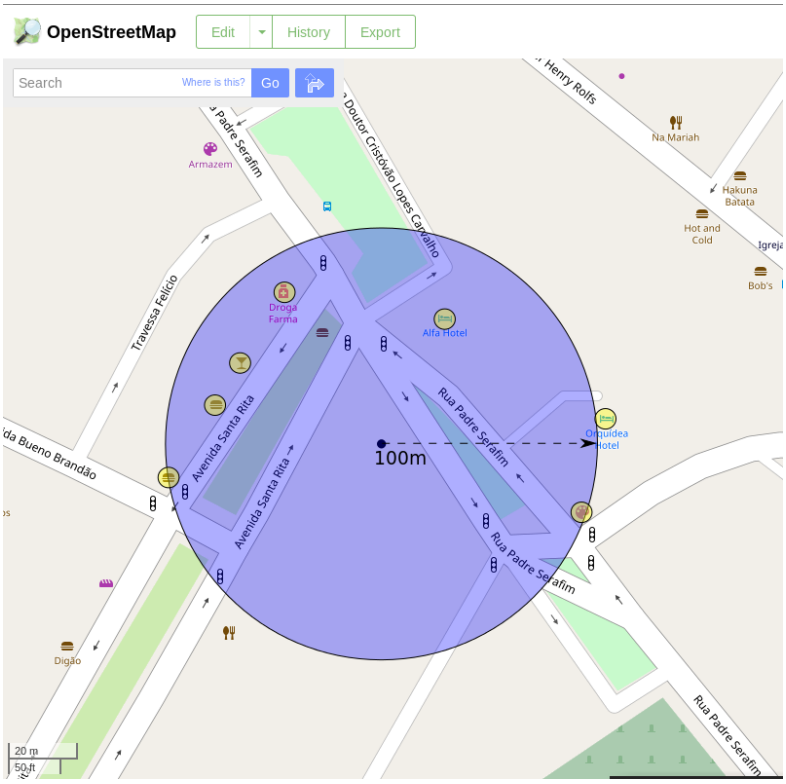

Figure 2. Example of a contribution semantically enriched by the proposed algorithm using OpenStreetMap.

Figure 3. Example of a contribution semantically enriched by the proposed algorithm using Google Maps in the Gota D'Água system interface.

5. Study Case: Gota D'Água System

As a study case, the automated VGI semantic enrichment methods and the semantic analyses in this paper were implemented and tested in a VGI collection system called Gota D'Água (http://www.gotadaguaufv.com.br). This system was developed aiming to collect data through volunteered citizen contributions regarding water waste and shortages, a recurring issue in several regions of Brazil during the dry season.

Gota D'Água allowed users from different parts of Brazil to contribute information on different types of issues related to water shortage and/or waste. When a contribution was made, the place indicated by the user was identified from its geographic coordinates and stays highlighted on the map for visualization, as seen in Figure 4. Moreover, the user can choose the type of issue identified (e.g., leak, waste) and provide more detailed information as free text, image, or video.

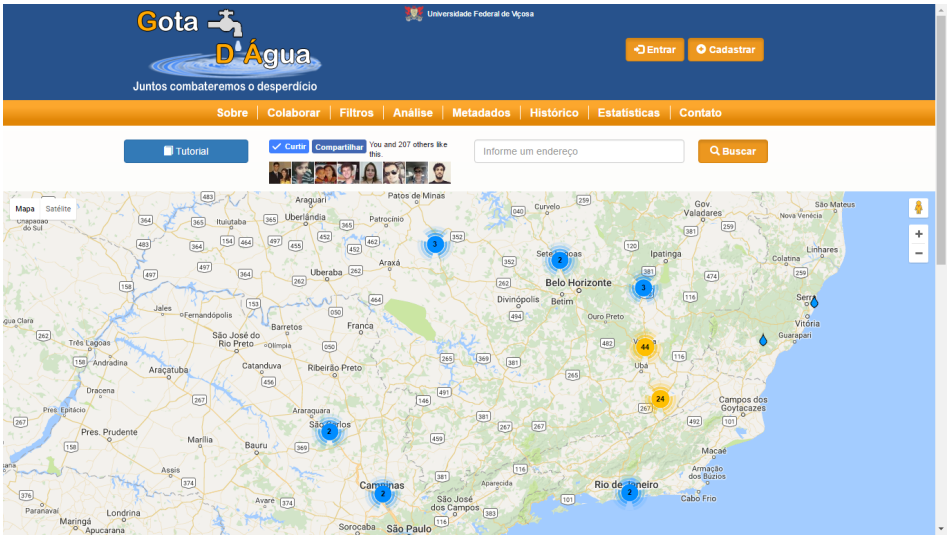

Figure 4. Home page of the Gota D'Água system.

5.1. Tools Employed

In order to carry out the steps shown in Figure 1, several tools and artifacts were used. The contributions come from the Gota D'Água system and pre-processing is applied to the contributions to deal with issues such as privacy (replacing the names of users with aliases) and possible noise generated by the system (e.g., duplicate contributions). The PoIs used as resources for semantic enrichment are obtained from the LinkedGeoData (LGD) repository. That repository used information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the linked data principles. LGD interconnects those data with other knowledge bases of the Linking Open Data (LOD) initiative, such as DBpedia. To carry out the geospatial operations proposed in Algorithm 1, the PoIs were added to a PostgreSQL+PostGIS database using the Osmosis (http://wiki.openstreetmap.org/wiki/Osmosis) tool. Although there are geographic data management systems (GDMS) that support spatial queries on RDF data (Strabon and Oracle, for example), geographic operations on RDF data are still less efficient than geographic queries performed on relational DBMS. Moreover, PostgreSQL is free and open source software. Finally, the Gota D'Água system also used PostgreSQL, which further justified its choice.

The pre-processing of the data was done through several scripts. For grammar mistakes, we used the Java library JOrtho (http://jortho.sourceforge.net/). Cases where JOrtho could not suggest the correct form of the word were manually analyzed. Some cases of time and space inconsistencies were fixed by comparing with the log files of the Gota D'Água that we have access. The VGI contributions were semantically annotated with Algorithm 1, implemented in Java, and the Apache Jena (https://jena.apache.org) library was used to generate the semantic annotations using the RDF standard. The triplified data and the semantic annotations generated by the algorithm (both follow the linked data principle) were stored in the Virtuoso database, thus allowing it to be queried in SPARQL. Finally, the thesaurus developed in this study (detailed in Section 5.4) was based on the Thesaurus of Sanitary and Environmental Engineering (Cepis, 2005).

5.2. Examples of SPARQL Queries

To demonstrate some of the several possibilities and advantages that semantic enrichment provides to VGI, some SPARQL queries were proposed for analysis of the existing data in the Gota D'Água system. Such queries show analyses that are only possible after the semantic enrichment process. All queries were tested on the Virtuoso database, whose instance contains the triplified contributions of the Gota D'Água system and the semantic annotations of the contributions.

Q1. Recover the IDs of the contributions, described keywords in the contributions, and the names of PoIs linked to them through semantic annotations.

This query aims to identify the existing <keyword, PoI> pairs for each VGI contribution. That enables verifying which keywords appear for certain PoIs to verify, for example, whether the PoIs returned are related to a large number of keywords (indicating that different issues occur near it) or whether a keyword is recurring to that PoI (indicating the issue repeats n times or that different user reported the same issue). Algorithm 2 presents Q1 in SPARQL language performed in this study.

Algorithm 2 Q1 in SPARQL language.

```
1: SELECT ?contribution ?key ?place
2: WHERE {?contribution <gd:keywords> ?key.
3: ?a <oa:hasTarget> ?contribution; <oa:hasBody> ?l.
4: ?l rdfs:label ?place}
5: ORDER BY ?place
```

In line 1 of Algorithm 2, all <keyword, PoI> pairs are recovered for all contributions. Property <gd:keyWords> was defined by the authors of the present study, being gd the prefix for <http://www.gotadaguaufv.com.br/>. On line 3, it is specified that, for each semantic annotation (associated with variable ?a), a PoI pointed by the semantic annotation (variable ?l) is desired. Finally, the name of the PoI is associated with the variable ?place (line 4), with the tuples returned sorted in ascending order by the name of the PoIs returned (line 5).

Table 2 presents a sample of the tuples returned by query Q1 performed on the Virtuoso database. It also shows that, for PoI Alfa Hotel, two distinct keywords, each in a different contribution, were associated with the PoI. Both keywords ("leak" and "broken pipe") refer to a water leak issue, possibly of the water or sewage grid. These tuples enable outlining two possible scenarios, which can be verified through the timestamp of the submission of the contribution: (i) in case both contributions were reported in the system within a short period of time, the reliability of such problem having actually happened increases, i.e., the reliability of both contributions increases, and (ii) in case the times of each contribution are far apart, that means the issue is persistent in that PoI for some reason, which should be investigated to reduce the losses to the utility company. However, due the VGI contributions in the Gota D'Água does not spam in a wide period of time (the first contribution and the last contribution considered is this work are only a few weeks apart), this kind of analysis is not present in this work.

Table 2. Sample of tuples returned by query Q1.

Contribution	Key	Place
http://gotadagua.com.br/96	Water	Academia Body Move
http://gotadagua.com.br/111	Waste	Academia Vila Fit
http://gotadagua.com.br/111	Hose	Academia Vila Fit
http://gotadagua.com.br/111	Sidewalk	Academia Vila Fit
http://gotadagua.com.br/2	Water, broken tap	Aconchego Bar e Restaurante
http://gotadagua.com.br/25	Leak	Alfa Hotel
http://gotadagua.com.br/18	Broken pipe	Alfa Hotel
http://gotadagua.com.br/3	Leak	Armazém
http://gotadagua.com.br/3	Waste	Armazém
http://gotadagua.com.br/86	Leak	Armazém
http://gotadagua.com.br/1	Water shortage, linked data, example	Artes Gerais
http://gotadagua.com.br/18	Broken pipe	Bar Norte Mineiro
http://gotadagua.com.br/96	Water	Bar do Gomes

Q2. Recover the number of times a PoI was indicated by a semantic annotation in the semantic enrichment process.

This query aims to return the number of contributions linked to each PoI. The results of query Q2 allows identifying the most recurring PoIs when some issue related to water shortage or waste occurs. Algorithm 3 presents Q2 in SPARQL language.

Algorithm 3 Q2 in SPARQL language.

```
1: SELECT ?place (count (distinct ?contribution))
2: WHERE { ?a <oa:hasTarget> ?contribution; <oa:hasBody> ?l.
3: ?l rdfs:label ?place}
4: ORDER BY ?place
```

On line 1 of Algorithm 3, all PoIs are recovered while the contributions related to each place recovered are counted. Lines 2 and 3 of Algorithm 3 behave exactly the same as lines 3 and 4 of Algorithm 2. Line 4 is responsible for organizing the result by PoI name in descending order.

Table 3 presents a sample of the tuples returned by query Q2 performed on the Virtuoso database. Table 3 shows that only one contribution was associated with PoI Academia Vila Fit, and, finally, that PoIs Cacau Show and Boca Viçosa were associated with three contributions. In other words, there was evidence that a greater number of issues occur in the regions surrounding these businesses, which must be investigated by the proper organs.

Table 3. Sample of tuples returned by query Q2.

Place	Count
Academia Vila Fit	1
Alfa Hotel	2
Armazém	2
Bar Norte Mineiro	1
Bar do Gomes	1
Biblioteca IME-USP	1
Bob's	2
Boca Viçosa	3
Cacau Show	3

Q3. Verify the number of keywords associated by PoI and which keywords they are.

This query returns the same result as Q1, however, in a more user-friendly manner. This query was proposed to demonstrate that SPARQL is able to carry out complex operations such as concatenation of strings to return results that satisfactorily meet several applications, as shown in Algorithm 4. As seen in Table 4, it is easier to verify the keywords associated with each PoI than the result shown on Table 2.

Algorithm 4 Q3 in SPARQL language.

```
1: SELECT ?place (count(?place)) (sql:group_concat(?key, '; ') as ?names)
2: WHERE { ?contribution <gd:keywords> ?key.
3: ?a <oa:hasTarget> ?contribution; <oa:hasBody> ?l.
4: ?l rdfs:label ?place}
5: ORDER BY ?place
```

Table 4. Sample of tuples returned by query Q3.

Place	Count	Keys
Academia Body Move	1	Water
Academia Vila Fit	3	Waste; sidewalk; hose
Aconchego Bar e Restaurante	2	Water; broken tap
Armazém	3	Leak; waste; leak
Biblioteca IME-USP	1	Leak
Bar Norte Mineiro	1	Broken pipe
Bar do Gomes	1	Water
Vitrola Café	3	Waste; sidewalk; hose
Viçosa Outlet	3	Waste; sidewalk; hose

Q4. Verify how many times the keywords in the contributions are mentioned for each PoI type present in the semantic annotations.

This query aimed to verify the number of PoIs associated with each keyword. The types of PoIs, or of any data that follow the RDF standard, are defined by attribute "rdf:type," which can be replaced by expression "a" in SPARQL, as shown in line 4 of Algorithm 5. To facilitate the presentation of query Q4 results, only types expressed in the ontology of LinkedGeoData were recovered. A regular expression was employed to apply this filter (line 4 of Algorithm 5).

As shown in Table 5, many PoIs linked in the semantic enrichment process are of the type Amenity. That occurs because Amenities are places that provide leisure and/or comfort, considered a generic type of place in the LinkedGeoData ontology. Since LinkedGeoData is based on OSM, whose contributions are often made by users with no knowledge or training in geospatial data, the number of PoIs with generic classification is high.

Algorithm 5 Q4 in SPARQL language.

```
1: SELECT ?key ?TypePlace (count(?TypePlace)) AS ?count
2: WHERE { ?contribution <gd:keywords> ?key.
3: ?a <oa:hasTarget> ?contribution; <oa:hasBody> ?l.
4: ?l a ?TypePlace. FILTER regex(?TypePlace, "^http://linkedgeodata.org/ontology/", "i")}
5: ORDER BY ?TypePlace
```

Table 5. Sample of tuples returned by query Q4.

Key	Type Place	Count
Water	http://linkedgeodata.org/ontology/AlcoholShop	1
Leak	http://linkedgeodata.org/ontology/Amenity	9
Waste	http://linkedgeodata.org/ontology/Amenity	7
Water	http://linkedgeodata.org/ontology/Amenity	8

5.3. Thesaurus Creation

Information recovery has become indispensable among the several members of the scientific community and thesauri arose as an answer to this need of recovering information [30]. While summarizing the countless definitions of thesauri found in the literature, ref. [31] describe thesauri that define their characteristics and goals. A thesaurus is defined as a controlled vocabulary formed by semantically related terms/descriptors that acts as an instrument of terminology control. Thesauri are used in knowledge organization/representation [32]. They aim to organize specialized information, coordinate the specialized vocabulary, systematically help the user's query, and potentialize the recovery of information, acting as interfaces between information and consumers.

The relationship among concepts, guided by their characteristics, is hierarchical and may manifest in a superordinate (from the most specific to the most general concept) or subordinate (from the most general to most specific concept) manner.

In this research, a segment was extracted from a thesaurus on water resources based on VGI contributions of the study case. To this task, first a survey was performed of the words used in the descriptions of volunteered contributions stored on the database of the Gota D'Água system. Based on this collection, the entry in the form of terms was determined and two lists were generated: (i) Terms that appear on the title of the contribution and (ii) terms found in the contribution to describe the water-related issue. Moreover, the terms found were organized and classified as related to the region/place, to water, or regarding quality/characteristic. After the survey and division of terms into categories, each one was defined based on reliable references.

The thesaurus obtained is the result of excerpts from the Thesaurus of Sanitary and Environmental Engineering [28], which compiles 2098 descriptors and 3037 non-descriptors. This thesaurus was reviewed and updated mainly based on the bibliographical database of the Pan-American Network of Information on Environmental Health (REPIDISCA) and the different sources of information of the Virtual Library of Environmental Health (BVSA). The General Multilingual Environmental Thesaurus (GEMET), version 2004, prepared by the Consiglio Nazires desconale delle Ricerche (CNR) and Umweltbundesmt (UBA), were also used as references.

The descriptors are presented with their preferential, hierarchical, and associative relations. The abbreviations to represent them in the thesaurus are as follows:

- SN (scope note)—defines, explains, or limits the meaning of the descriptor for indexing purposes;
- UF (used for)—indicates the synonyms or non-descriptors; they are valid for indexing;
- GT (generic term)—marks the broadest term to which the descriptor belongs;
- ST (specific term)—indicates the specific terms (types or classes) associated with the descriptor;
- RT (related term)—indicates the terms semantically associated with the descriptor.

Each base descriptor has a code so it can be located in the semantic or classified section, as seen in Box 1.

Box 1: Example of a descriptor found in the thesaurus [28].

Water resources
SN The volume of surface or underground water available for any use in a specific region.
UF Affluents
Water masses
Water sources
Hydrologic resources
Limnological resources
GT Natural resources [Environmental Health]
ST Underground water
Surface Water
Transborder water resources
RT Water supply
Basins
Water resource planning
Water policy
Marine resources

5.4. Thesaurus-Based Process of Recovering Related VGI

As shown in Section 5.3, a thesaurus is a resource that helps in semantic enrichment by relating and restricting meanings. Therefore, the embedded semantics in a thesaurus can be explored by several applications to obtain more precise and semantically complete results. In order to exemplify how the semantics present in the thesaurus built from VGI contributions may be used, this study proposes a process that returns semantically similar VGI contributions based on a query provided by a user.

Figure 5 presents the process proposed. A user makes a query with terms he or she wishes to feature in the VGI contributions. For example, given the query "leaks close to stores," the user wants all contributions that relate leaks that occurred near PoIs of the type Shop. The query is processed in the query processing module, which applies several natural language processing techniques (e.g., tokenization, part-of-speech, shallow parsing) according to the application need. The processed query is used both in the sub-module, compare w/thesaurus, and in the sub-module, compare w/enriched VGI. Both modules compare the terms identified in the query to the types of data each sub-module refers to. In order to speed up the process, both use indices. Finally, the queries recovered by each sub-module are compared. The contributions present in both sets are considered similar and are returned to the user.

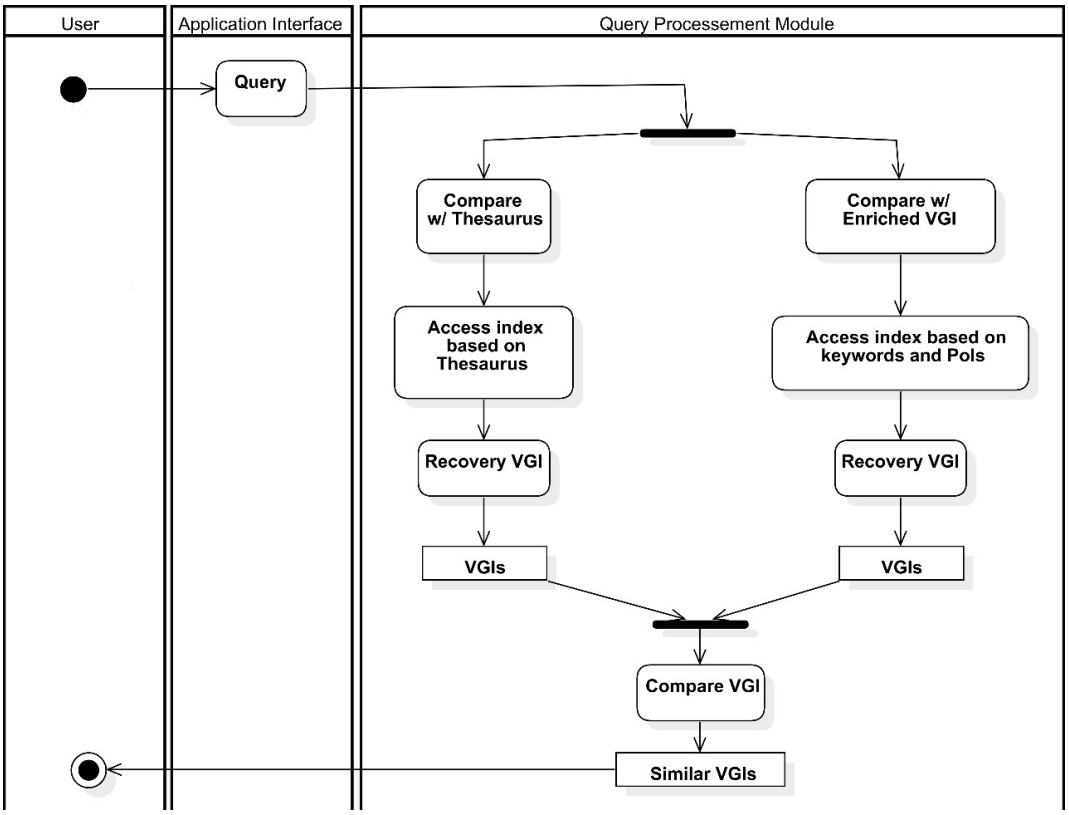

Figure 5. Process to return semantically similar VGI contributions based on a user-provided query.

6. Discussion

The Gota D'Água VGI system, used in this paper as study case, allows users from different parts of Brazil to contribute information on different types of issues related to water shortage and/or waste. VGI normally has little context attached to it since users choose to write short messages for several reasons, such as lack of time, convenience, difficulty in using the system, etc.

According to [33], when working with semantic enrichment with little text or multimedia content such as social media posts, the results have little precision due to the lack of context. It is believed that the same also occurs with VGI, which usually has succinct contributions. Since the descriptions of user contributions in the Gota D'Água system are little detailed, the objective of this research is not to be more precise than existing approaches in the literature, but rather to show the viability and applicability of automated semantic enrichment of VGI.

With the generation of semantic annotations in VGI adopting the RDF standard, queries can be made to be answered by the system using the SPARQL language, which makes the system create logical inferences. The SPARQL language allows a query to be used to formulate questions ranging from a simple graphical standard of correspondence to more complex queries involving several RDF repositories around the web [34]. Although the SPARQL language is equivalent to SQL, i.e., any SPARQL expression can be translated into an SQL expression with no loss of semantics [35,36], expressing a semantic query in SPARQL is friendlier than in SQL. In addition, DBMSs that have SPARQL support have specific query optimization modules for semantic queries.

The simplicity of the method proposed in this study to semantically enrich VGI contributions allows it to be implemented in a variety of applications of several domains and may be adapted according to the needs of each domain. Moreover, the method proposed aims to use geospatial operations that already exist in a database since these databases implement optimized geospatial operations for large datasets, which adds to the efficiency of the method.

In order to assess the method described in Section 4, the following steps must be executed: Formulating questions related to the water distribution subject, whose answers are obtained via SPARQL queries. The SPARQL queries are made based on the triplified data of the Gota D'Água system and on the semantic annotations in RDF generated by the semantic enrichment. A small number of contributions is manually examined and annotated. The method proposed generates similar annotations and a certain level of imprecision is already expected. After validation of the semantic enrichment method, the queries are executed so as to attempt to answer the previously proposed questions. The data obtained are transcribed in the research results in Section 5.2.

The greatest difficulty of the research is dealing with the fact that VGI contributions are succinct. For example, many contributions in the Gota D'Água system have no type of comment attached to them and feature only the type, the geographic location, and timestamp of the contribution. Such lack of details prevents the use of tools such as EL, WSD and other semantic enrichment methods that may generate new semantic annotations and, consequently, contribute to the results of queries to be executed. Furthermore, given the limited number of contributions, messages with little context and small number of attributes per contribution, the possible queries to demonstrate the applicability of semantic enrichment have limited scope. However, the queries presented in this study show the usefulness of semantic enrichment and enable answering useful questions to the organs responsible for water supply.

7. Conclusions and Future Works

Linked data along with the RDF standard is used to integrate data by simplifying the relationship schemas and enriching semantics. Although its use seems advantageous, the technology is relatively new and there are several challenges to be explored and areas of application to be discovered [20]. This paper describes the results of a research project whose goal was to verify the applicability of linked data in semantic enrichment of VGI.

The importance of volunteered data is on the rise; however, such data often lack descriptive information. This research aimed to manually and automatically add information to VGI with the resources used for semantic enrichment coming from LOD repositories. The research resulted in a VGI set with annotations that point to LOD, which enables semantic and/or more complex queries to be performed on the contributions. A thesaurus was also produced, which was used to obtain more precise and semantically more complete results since the semantics embedded in a thesaurus may be explored by several applications.

Future works may include developing the process proposed using the thesaurus, which returns semantically similar VGI contributions based on a user-provided query to show how the semantics present in the thesaurus used from VGI contributions can be used for applications. Another approach is to build an ontology using as starting point the ontology present in LOD resources. Such ontology will be useful for description and relations among VGI contributions since an ontology requires a record of knowledge of the domain in a language that can be processed by computers for inferences.

Author Contributions: Conceptualization, L.S.d.C., I.L.O. and J.L.-F.; Methodology, L.S.d.C., I.L.O. and A.M.; Software, L.S.d.C. and I.L.O.; Validation, L.S.d.C.; Analysis, L.S.d.C. and I.L.O.; Resources, L.S.d.C. and J.L.-F.; Data Curation, L.S.d.C.; Writing-Original Draft Preparation, L.S.d.C., I.L.O., J.L.-F. and A.M.; Writing-Review and Editing, L.S.d.C., I.L.O. and J.L.-F.; Supervision, J.L.-F.; Project Administration, J.L.-F.; Funding Acquisition, J.L.-F.

Funding: This research was funded by Fundação de Amparo à Pesquisa do Estado de Minas Gerais—FAPEMIG grant number APQ 03763-12.

Acknowledgments: The authors would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES for the scholarship.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data: The story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*; IGI Global: Hershey, PA, USA, 2011; pp. 205–227.
2. Goodchild, M.F. Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *Int. J. Spat. Data Infrastruct. Res.* **2012**, *2*, 4–32.
3. Elwood, S.; Goodchild, M.F.; Sui, D.Z. Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Ann. Assoc. Am. Geogr.* **2012**, *102*, 571–590. [CrossRef]
4. Clarke, C. A resource list management tool for undergraduate students based on linked open data principles. In Proceedings of the European Semantic Web Conference, Crete, Greece, 31 May–4 June 2009; pp. 697–707.
5. Schade, S.; Granell, C.; Diaz, L. Augmenting SDI with linked data. In Proceedings of the Workshop on Linked Spatiotemporal Data, in conjunction with the 6th International Conference on Geographic Information Science (GIScience 2010), Zurich, Switzerland, 14 September 2010.
6. Stadler, C.; Lehmann, J.; Höffner, K.; Auer, S. Linkedgeodata: A core for a web of spatial open data. *Semant. Web* **2012**, *3*, 333–354.
7. Bontcheva, K.; Rout, D.P. Making sense of social media streams through semantics: A survey. *Semant. Web* **2014**, *5*, 373–403.
8. Ronzhin, S. Semantic Enrichment of Volunteered Geographic Information Using Linked Data: A Use Case Scenario for Disaster Management. Master's Thesis, University of Twente, Enschede, The Netherlands, 2015.
9. Sorrentino, S.; Bergamaschi, S.; Fusari, E.; Beneventano, D. Semantic annotation and publication of linked open data. In Proceedings of the International Conference on Computational Science and Its Applications, Ho Chi Minh City, Vietnam, 24–27 June 2013; pp. 462–474.
10. Flanagin, A.J.; Metzger, M.J. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148. [CrossRef]
11. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
12. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
13. Zielstra, D.; Zipf, A. A comparative study of proprietary geodata and volunteered geographic information for Germany. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 11–14 May 2010.
14. Neis, P.; Zielstra, D. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet* **2014**, *6*, 76–106. [CrossRef]
15. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120. [CrossRef]
16. Foody, G.M.; See, L.; Fritz, S.; Van der Velde, M.; Perger, C.; Schill, C.; Boyd, D.S. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Trans. GIS* **2013**, *17*, 847–860. [CrossRef]
17. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Des.* **2010**, *37*, 682–703. [CrossRef]
18. Fan, H.; Yang, B.; Zipf, A.; Rousell, A. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 748–764. [CrossRef]
19. Brovelli, M.A.; Minghini, M.; Molinari, M.; Mooney, P. Towards an automated comparison of OpenStreetMap with authoritative road datasets. *Trans. GIS* **2017**, *21*, 191–206. [CrossRef]
20. Azevedo, P.C.N. Uma Proposta Para Visualização de Linked Data Sobre Enchentes na Bacia do Rio Doce. Ph.D. Thesis, Universidade FUMEC, Belo Horizonte, MG, Brazil, 2014.
21. Beneventano, D.; Bergamaschi, S.; Sorrentino, S.; Vincini, M.; Benedetti, F. Semantic annotation of the CEREALAB database by the AGROVOC linked dataset. *Ecol. Inf.* **2015**, *26*, 119–126. [CrossRef]
22. Berners-Lee, T.; Chen, Y.; Chilton, L.; Connolly, D.; Dhanaraj, R.; Hollenbach, J.; Lerer, A.; Sheets, D. Tabulator: Exploring and analyzing linked data on the semantic web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop, Athens, GA, USA, 5–9 November 2006; p. 159.
23. Lehmann, J.; Bizer, C.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. DBpedia—A Crystallization Point for the Web of Data. *J. Web Semant.* **2009**, *7*, 154–165.

24. Moreira, J.D.C.; Neto, F.M.M.; da Costa, A.A.L.; Sombra, E.L.; de Aliança Neto, A.S.; de Medeiros Valentim, R.A. Um sistema de enriquecimento semântico de perfil de usuário baseado em traços digitais para apoio à aprendizagem informal no contexto da saúde. *RENOTE* **2014**, *12*. [CrossRef]
25. Clarke, M.; Harley, P. How smart is your content? Using semantic enrichment to improve your user experience and your bottom line. *Science* **2014**, *37*, 41.
26. Moro, A.; Raganato, A.; Navigli, R. Entity linking meets word sense disambiguation: a unified approach. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 231–244. [CrossRef]
27. de Moura, T.H.V.; Davis, C.A., Jr. Linked Geospatial Data: desafios e oportunidades de pesquisa. In Proceedings of the XIV GEOINFO, Campos do Jordão, Brazil, 24–27 November 2013; p. 13.
28. CEPIS—Centro Pan-Americano da Engenharia Sanitária e Ciencias do Ambiente. Tesauro de Engenharia Sanitária e Ambiental. 2005. Available online: http://www.bvsde.paho.org/bvsair/e/manuales/tesa/tespo.pdf (accessed on 26 June 2019).
29. Gao, H.; Barbier, G.; Goolsby, R. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intell. Syst.* **2011**, *26*, 10–14. [CrossRef]
30. Moreira, M.P.; Moura, M.A. Construindo Tesauros a Partir de Tesauros Existentes: A Experiência do TCI-Tesauro em Ciência da Informação. 2006. Available online: http://www.brapci.inf.br/_repositorio/2010/01/pdf_6c43aff315_0007598.pdf (accessed on 26 June 2019).
31. Moreira, A.; Alvarenga, L.; Oliveira, A.P. Thesaurus and Ontology: A Study of the Definitions Found in the Computer and Information Science Literature, by Means of an Analytical Synthetic Method. *Knowl. Organ.* **2004**, *31*, 231–244.
32. Rosati, I.; Bergami, C.; Stanca, E.; Roselli, L.; Tagliolato, P.; Oggioni, A.; Fiore, N.; Pugnetti, A.; Zingone, A.; Boggero, A.; et al. A thesaurus for phytoplankton trait-based approaches: Development and applicability. *Ecol. Inf.* **2017**, *42*, 129–138. [CrossRef]
33. Guo, W.; Li, H.; Ji, H.; Diab, M. Linking tweets to news: A framework to enrich short text data in social media. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 239–249.
34. World Wide Web Consortium. SPARQL 1.1 Overview. W3C Recommendation. 2013. Available online: http://travesia.mcu.es/portalnb/jspui/handle/10421/7464 (accessed on 26 June 2019).
35. Chebotko, A.; Lu, S.; Fotouhi, F. Semantics preserving SPARQL-to-SQL translation. *Data Knowl. Eng.* **2009**, *68*, 973–1000. [CrossRef]
36. Elliott, B.; Cheng, E.; Thomas-Ogbuji, C.; Ozsoyoglu, Z.M. A complete translation from SPARQL into efficient SQL. In Proceedings of the 2009 International Database Engineering & Applications Symposium, Calabria, Italy, 16–18 September 2009; pp. 31–42.