International Journal of Geo-Information

MDPI

Article

User Generated Spatial Content-Integrator: Conceptual Model to Integrate Data from Diverse Sources of User Generated Spatial Content

Jacinto Estima * and Marco Painho

NOVA IMS, Universidade Nova de Lisboa (UNL), Lisboa 1070-312, Portugal; painho@novaims.unl.pt
* Correspondence: jacinto.estima@gmail.com

Academic Editor: Wolfgang Kainz
Received: 7 September 2016; Accepted: 26 September 2016; Published: 9 October 2016

Abstract: Geographic information has been traditionally produced by mapping agencies and corporations, using highly skilled professionals as well as expensive precision equipment and procedures, in a very costly approach. The production of land use and land cover databases is just one example of such traditional approaches. At the same time, the amount of Geographic Information created and shared by citizens through the web has been increasing exponentially during the last decade as a result of the emergence and popularization of technologies such as the Web 2.0, cloud computing, global positioning systems (GPS), smart phones, among others. This vast amount of free geographic data might have valuable information to extract. Combining data from several initiatives might further increase the value of such data. We propose a conceptual model to integrate data from suitable user generated spatial content initiatives. A prototype to demonstrate the ability of the model to perform such integration, based on two identified use cases, was also developed.

Keywords: land use/land cover; geographic information systems; user generated spatial content; spatial data integration; VGI; volunteered geographic information

1. Introduction

Official Geographic Information (GI) has been produced by mapping agencies and corporations and sold to users as paper maps or atlases [1], following a very expensive approach that requires expert people as well as expensive equipment and precise procedures. Consequently, priority has been given to the most important and unchanging geographic themes and those with multiple applications, thereby relegating the others to a secondary role [2].

Recently, we have witnessed the emergence of a new phenomenon in which citizens have been creating and sharing GI through the web. The development and popularization of technologies such as the Web 2.0, cloud computing, global positioning systems (GPS), smart phones, among others, has transformed, and continues to transform, the way that geographic data are produced, stored, and used [3]. Research has already been conducted exploring the enormous potential that this type of data seems to be hiding and find possibilities of applying it to real world problems, such as Land Use/Cover mapping [4,5], disaster response [1,6,7], representation of natural features [8], exploration of vernacular language [9], or the enhancement of cultural heritage [10].

On this matter, Land Use/Cover (LULC) databases represent one of the interesting areas in which these data sources could be very helpful. Their production is very costly and time consuming, as it is mainly based on interpretation and classification of remote sensing data made by highly trained and skilled people [11]. Moreover, the process includes a validation phase that is extremely important to provide quality indicators to the final product. This validation is done by confronting the produced database with reference data assumed to be true, which includes, among other sources, "ground truth"

collected directly from the field in pre-selected sites [12]. GI shared by citizens through the web can be used to help the production process, for instance, by decreasing the need to collect data in the field. Various authors have already explored for this purpose diverse sources of GI produced by citizens and proposed methods to overcome some of the particularities of this type of data ([4,13–25]), with most of them reflecting the possible benefits from the integration of data from various sources such as, for instance, the increase of spatial and temporal resolutions.

This study develops a data model able to integrate diverse sources of GI produced by citizens to help in the production of LULC databases. As different initiatives have different goals, interests, and audiences, and different types of data are produced, stored with different structures, and made available by different types of access, such an integration represents additional challenges to retrieve, analyze, extract, and visualize useful information from various sources. This requires the development of integration models to overcome their dissimilarities.

This paper is organized as follows. First we look at the different initiatives of GI produced by citizens, establish a list of minimum requirements for an initiative to be included in the integration model, and select the initiatives that meet these requirements. Then we discuss the most important dissimilarities among the sources selected and propose a conceptual integration model. Finally, we assess the model through a set of pre-defined use cases and make some final remarks.

2. User Generated Spatial Content

In 2007, Goodchild coined the term Volunteered Geographic Information (VGI) to describe "the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies" [26]. One year before, in 2006, Neogeography was introduced by Turner as a term to describe the phenomenon of "people using and creating their own maps, on their own terms and by combining elements of an existing toolset, sharing location information with friends and visitors, helping shape context, and conveying understanding through knowledge of place" [27]. Crowdsourcing geospatial data is another term used to describe the phenomenon of large unorganized groups of users generating content (spatial in this case) that is shared [28].

Despite some differences between these terminologies [29], they are all related to a type of User Generated Content (UGC) that deals directly or indirectly with spatial content and refers to volunteers and large groups of people, sometimes acting like a crowd, often without expertise or formal qualifications, contributing with spatial data to the "community".

More recently, Stefanidis et al. [30] came up with what they defined as a "deviation from Goodchild's notion of volunteered geography" (p. 319). They argue that the information disseminated through some social media initiatives is not geographic information per se (i.e., geography is not their main purpose, unlike other initiatives such as OpenStreetMap), although they provide a geographic context since they have associated information about location. They called it Ambient Geospatial Information (AGI). Fischer [31] argued that in some cases, when VGI is used for purposes other than those for which volunteers have contributed, it can be seen as a not-so-Volunteered Geographic Information, and termed this as involuntary geographic information (iVGI).

We adopt the wider term User Generated spatial Content (UGsC) to unite all these diverse definitions [32]. This term is a particular case of UGC that deals with spatial content, and is intended to encompass all the initiatives containing data with spatial characteristics provided by citizens with or without the purpose of contributing data for spatial purposes, such as VGI, iVGI, neogeography, crowdsourcing geospatial data, and AGI.

3. User Generated Spatial Content—Integrator Model

Different types of data, stored with different structures and made available by different types of access, represent additional challenges in dealing with data from various sources. This requires the development of integration models to overcome the dissimilarities and extract useful information. For this it is important to define the minimum requirements that an UGsC data source must have in order to be included in the model, identify those initiatives that follow these minimum requirements, look at their similarities/dissimilarities, and finally develop the integration model.

3.1. Minimum Requirements and Relevant Initiatives

Following the inventory made by Elwood et al. [29], 99 initiatives were identified in 2009 and the most recent version of the list, available online, counts 100 initiatives, but no update date is mentioned [33]. Each initiative was checked for availability, resulting in 61% of initiatives still active without changes, 3% having changed their name, and 36% no longer active. The most well-known initiatives such as OpenStreetMap (OSM), Flickr, Panoramio, Wikimapia, among others, are still active. The inventory classifies the initiatives according to their purpose in three groups: geovisualization, geoinformation, and geosocial. Geovisualization is oriented to mapping user-contributed information. Geoinformation is concerned with capturing, compiling, and integrating geotagged content (data generated through location-based services) and geolocational information for place names. Geosocial is more focused on users sharing geolocated media with others in their professional or social networks.

Given the purpose of this study, we are more interested in UGsC projects that acquire and store data related with physical aspects of the Earth rather than data about users' locations or being a platform for the aggregation of all types of data. We start by analyzing the active initiatives identified in the inventory to establish a list of essential requirements that any source needs to meet to be included in the UGsC Integration model. From this analysis some important characteristics were identified, and need to be discussed prior to the requirements definition:

1. **Type of spatial context**: In this matter we found two main types of spatial resolution: places and coordinates (latitude and longitude). Places are not accurate and sometimes can be very vague in terms of spatial location [9]. For instance, when one mentions the name of a city, there is no accurate position in that city. Coordinates refer to a location with much more accuracy and therefore are of more interest for this study.
2. **Type of spatial phenomena**: landscape, user position, highly dynamic phenomena (natural, such as fires, tornados, etc., or artificial, such as cars, animals, people, etc.), and static entities (buildings, roads, farms). User position and highly dynamic phenomena are not of interest for this study because they do not represent physical aspects of the earth.
3. **Type of data**: text, photos, and geometries. Text events, when georeferenced by latitude and longitude coordinates or similar, can be very precise and rich in terms of geographical information, but more research that is outside the scope of this study is needed to extract meaningful information from messages/descriptions. Photos, when georeferenced by latitude and longitude coordinates, are very useful as they provide an image of the location. Photos georeferenced by places, as mentioned in the previous point, can have a very imprecise location. Geometries are usually georeferenced by their coordinates representing precise geographic data.
4. **Type of access**: no public access, access using public APIs, access using private API, and access using direct URLs to the photos. Some initiatives, usually held by private companies, do not provide public access to stored data or require users to pay a fee to use their private API. Public APIs are available free of charge and manage privacy issues internally, so by using them only publically available content will be accessed. In this model only public APIs are considered.
5. **Type of data license**: Open Data Commons Open Database License (ODbL), license to public use, and license that belongs to the contributor, among others, are some of the types of data licenses

used. It is important to note that our model will use only publically available data and will not store or commercially exploit the data used.

6. **Type of coverage**: local, regional, or global. Local coverage is more related with a small portion of the Earth, like a country or a region inside a country. Regional coverage is more connected with areas covering groups of countries or continents. Global coverage is associated with the entire globe. Depending on the type of coverage of the LULC being produced and the area of the Earth being classified, some initiatives can be more interesting than others (e.g., if the working area is Portugal, UGsC data covering Ireland will not be of interest).

Useful information can be extracted from this discussion. Spatial context is of extreme importance to have precise locations of UGsC data. This does not mean that the information is accurate but rather that when a location is referred to we know exactly where it is with regard to the reference system used. It was consequently decided to eliminate all the initiatives that do not store data with spatial coordinates such as latitude and longitude or georeferenced geographical objects. Initiatives that do not provide a public API, free of charge, or do not allow access to stored data through Internet open protocols in any way, were also excluded from the study. In the same sense, for legal reasons, all the data without a free type of license were excluded. Consequently, a list of essential requirements that any initiative should follow to be included in the model was developed (Table 1). Table 2 shows UGsC initiatives identified that follow the defined requirements and were subsequently used in the development of this study.

Table 1. List of essential requirements that any initiative must have.

Type of Requirement	Requirement
Spatial context	Data have to be georeferenced by coordinates
Spatial phenomena	Data have to represent, at least partially, physical aspects of the Earth
Data type	Photos and geometries are preferred but text can also be valuable if text mining tools are available and implemented
Access type	Data must be publically accessible through the Internet using open protocols
Data license	Data must be available free of charge for the purpose of land use/cover classification
Coverage	Depends on the type of coverage of the Land Use/Cover (LULC) being produced and the area of the Earth being classified

All the initiatives have the data referenced by coordinates, representing physical aspects of the Earth, and are publically available. Except for the GeographUK (regional dataset covering Great Britain and Ireland) all the datasets have a global coverage. In terms of access type, all the initiatives provide public APIs to access their data, except the Degrees Confluence project, in which the access has to be made using photo specific URLs. Finally, concerning the type of data, two initiatives have vector data, five are based on photos, and seven have textual descriptions incorporated.

Table 2. Selected User Generated spatial Content (UGsC) initiatives.

Name	Since	Spatial Context (Data Georeferenced by Coordinates)	Spatial Phenomena	Coverage		Data Type			Access Type	Availability
				Global	Regional	Vector Data	Photos	Descriptions		
Degrees Confluence	1996	X	X	X			X	X	URL	Public
Flickr	2004	X	X	X			X	X	API	Public
OpenStreetMap	2004	X	X	X		X			API	Public
GeographUK	2005	X	X		X				API	Public
Panoramio	2005	X	X	X			X	X	API	Public
Wikimapia	2006	X	X	X		X		X	API	Public
Twitter	2006	X	X	X				X	API	Public
Instagram	2010	X	X	X			X	X	API	Public

3.2. Structural Similarities and Dissimilarities among the Initiatives Selected

As stated above, different UGsC initiatives have different goals, interests, and audiences, and produce different types of data, and consequently, different structures are adopted. In this section we explore the UGsC initiatives selected to find structural similarities and dissimilarities among them, in order to identify solutions for their integration.

Only one characteristic in common across all the initiatives was identified. All of them have a geographical location expressed in terms of latitude and longitude coordinates associated with the data. In this sense we identified two types of geographical representation: points, and multiple geometries. Most of the initiatives fall into the first and use points to represent their data. Photo based initiatives, such as Flickr and Panoramio, and message based initiatives, such as Twitter, associate, respectively, photos and messages with a point location. Some other initiatives are more related with the second type. OSM and Wikimapia are two examples of initiatives that use a multiple geometry approach by representing their data through points, lines, and polygons.

In terms of dissimilarities, two were quickly recognized. The first difference is related to the type of access. Two different types of access were identified: (1) accessing by using a direct URL; and (2) accessing through a public API. The former does not provide a search mechanism and needs a tailored development to retrieve information for very particular locations: the intersections of meridians with parallels. The latter provides a specific interface, publically available, with known operations to retrieve the desired information from the source. Although the majority of the initiatives provide a public API to access their data, it should be noted that the operations implemented by their interfaces are different from each other. Figure 1 provides a general overview of this common characteristic, also describing the type of access for each of the initiatives selected.

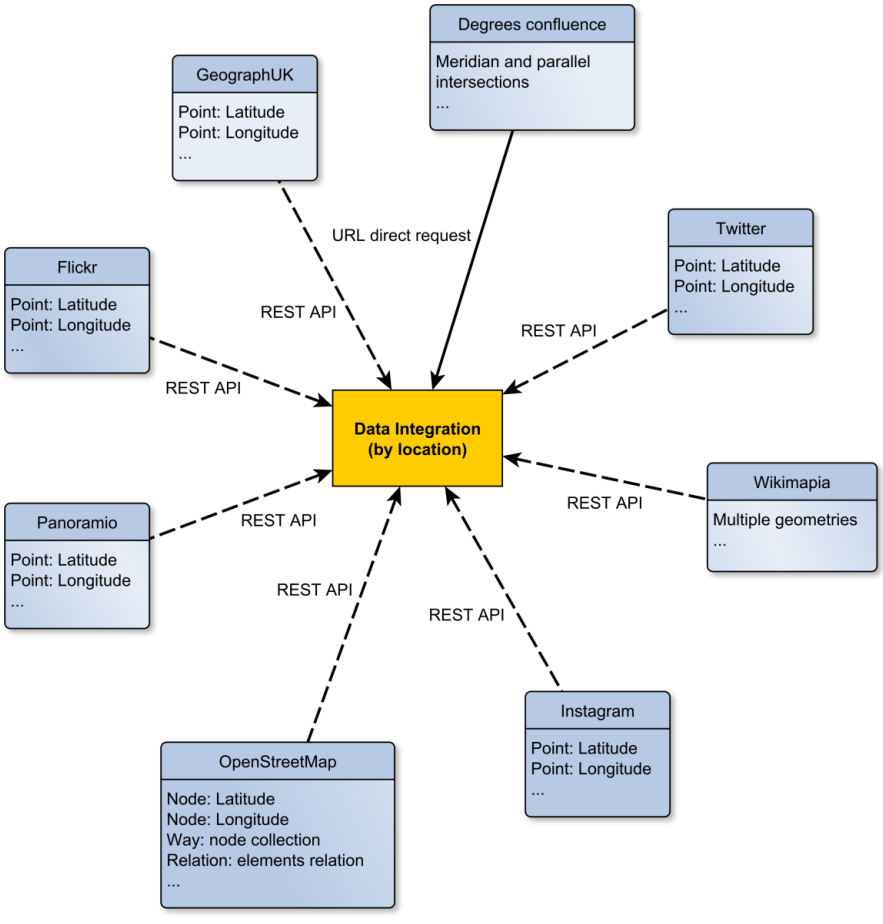

Figure 1. Data integration by location.

Another important difference that has to be pointed out is the schema of the response from each initiatives' API. Although there are some overlaps, the response schema of each initiative is, in general, different, which raises integration issues. Therefore, a common schema needs to be defined so that information aside from the location can also be integrated and used.

3.3. Model Architecture

There are three approaches to integrate several and diverse sources of data: (1) the virtual approach, in which the information is queried and retrieved from the source on-the-fly; (2) the materialized approach, in which a centralized database is developed to store data previously queried to the data sources; and (3) the hybrid approach, which is a mixture of the first two approaches [34]. According to these authors, the virtual approach fits better when the information sources are changing frequently, whereas the materialized approach would be preferable when the changes occur with lower frequency.

As mentioned above, UGsC data are of the type that change/update frequently. Therefore, the data integration model based on a virtual approach better fits the type of data we are dealing with, with the advantage of always accessing the most recent data available.

The data integration model will follow a virtual approach with the data from the different sources being queried and retrieved on-the-fly using an interactive online platform. Given also the nature of these diverse sources, having different structures and types of access, the integration is based on a mediator [35] that resides between the application tier and the UGsC sources. Speaking broadly, the aim of this architecture is to ensure that the query made by the user on the application tier is properly translated to the different UGsC sources automatically, without the user having to know the structure or access type of the sources.

This architecture is based on three tiers or levels: the application, the mediator, and the UGsC sources. As shown in Figure 1, the integration is made by overlaying the different data using their location parameters. Figure 2 presents a developed version of the architecture of the data integration model at the three levels, detailing the mediator tier.

The client tier establishes the interface between the user and the core application. It comprises mainly a Web Graphical User Interface (GUI) that displays all the information and allows user interaction. The user can easily query all the available UGsC sources for a specific location, visualize the response, and interact with the data.

The mediator tier is the core of the data integration model. As shown in Figure 2, it is composed of the integration component, including search settings defined by the user, and a set of readers. The integration component receives the query from the client tier, calculates the bounding box according to the defined settings, and dispatches it to the different available readers. Each reader is then responsible to formulate a specific query to the respective UGsC source, interpret the response, and send it back to the integration component. The integration component will then harmonize all the responses and send the result back to the client, to be displayed by the Web GUI.

One of the main advantages of the approach used in this architecture is the possibility to integrate new UGsC sources at any time, as long as they fulfil the minimum requirements defined, by developing a specific reader for each source and adding it to the integration component configuration settings. The integration component can also evolve in the future to incorporate tools to help in the decision making process. Descriptive statistics, data conflation, data fusion, text and data mining, or even machine learning techniques might be incorporated, and applied at the geographical and semantic levels, to provide better insights about the quality of the classification or, ultimately, to make the decision in a fully automated way.

This tier is composed of the data sources themselves. As mentioned in the previous section, as long as the minimum requirements are met, any new source can be added to the model by developing a reader that knows how to communicate and query the data to the source, as well as to interpret and format the response.

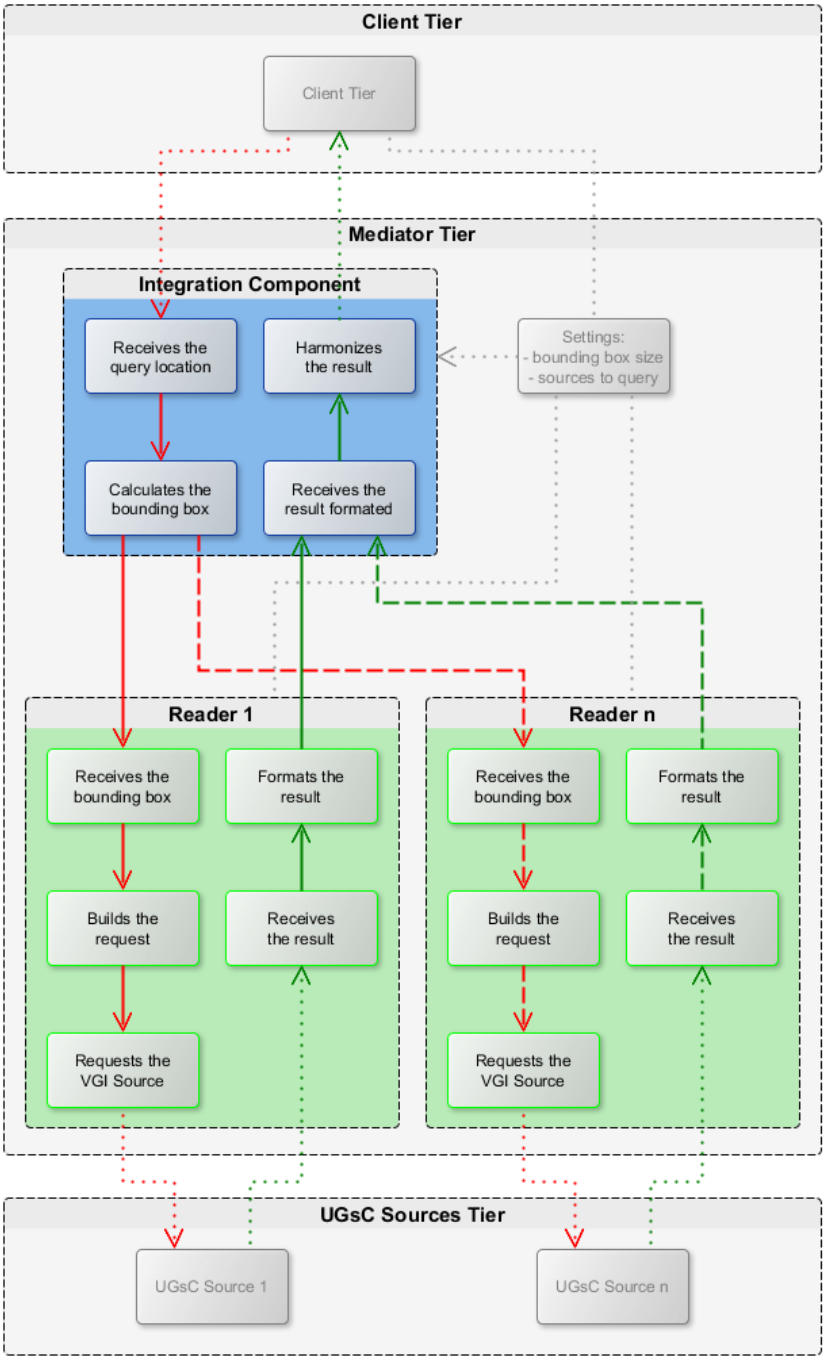

Figure 2. Detailed architecture of the data integration model (note: input, output, and settings' workflows respectively in red, green, and grey colors). VGI, volunteered geographic information.

4. Prototype Development and Implementation

To validate the UGsC-Integrator model proposed in the previous chapter, a prototype was developed and implemented. The first step was to define a set of important use cases in order to understand which features should be included. Use cases are a valuable and widely used tool to capture system requirements [36], and very helpful in designing systems. The architecture and implementation was then achieved based on those requirements.

4.1. Definition of Use Cases

To identify the requirements for the development of the prototype, two use cases were defined. Related with remotely sensed products, the first is about supporting the process of classification, e.g., to help a photo-interpreter to investigate areas of unclear classification, and the second is related to supporting the validation process.

Figure 3 shows an integrated view of the main operations required by these use cases. Basic operations, such as defining location, selecting initiatives to query, and visualizing the retrieved features in an integrated map, are needed by both use cases. These operations are enough to retrieve and visually analyze photos from initiatives providing this kind of data, thus helping in the photo-interpretation process. Advanced operations are of more interest for the validation use case. In this case, tools to analyze tag statistics, to select features by tag, and to export data to be integrated in external applications, are very important.

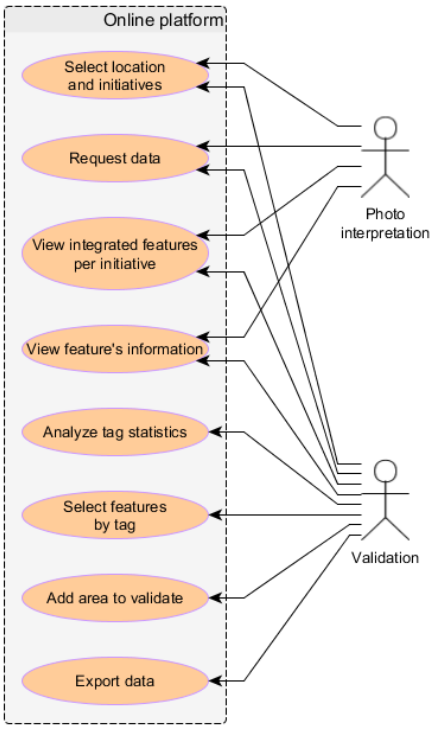

Figure 3. Integrated view of the use cases identified.

4.2. Architecture and Implementation

The prototype implementation started with the selection of the most appropriate technology. Given the fact that: (1) the crowd is continuously sharing geographic information through the initiatives identified; (2) internet access is required to access data; and (3) applications are running more and more in the cloud using the World Wide Web (WWW) to provide online tools for different purposes, it was decided to develop this prototype oriented to work in real-time and using the WWW as the platform of operation.

In terms of technology, and as the objective is not related with any evaluation of software or benchmark measurement, open source options with the necessary flexibility to implement interactive and user friendly solutions were selected. Thus, two main structures were required: (1) a web-based framework and (2) a mapping framework. For the first case the framework Sencha Ext JS, version 4.2.2 [37] was selected. This framework is a JavaScript framework for building feature-rich cross-platform web applications allowing developments with rich User Interface (UI) components. For the second case we selected Open Layers, version 3.1.1 [38]. This library is very well known for its

Web GIS development capability for high performance mapping. To serve the application, the Apache HTTP Server, version 2.4.10, was used [39]. This stack responds to all the defined requirements and has been used in several Web GIS implementations [40–45]. Based on the purpose of the model, we included in the prototype two completely different sources of UGsC: (1) a photo sharing initiative—Panoramio; and (2) a vector-based mapping initiative—OSM.

The next step was to design the main UI for the application. Based on the use cases it was clear that a two-step approach was needed. First the user would need to select the location to analyze along with the input parameters followed by the request itself, and second the resulting data would be displayed in an integrated way allowing a certain level of interaction between the user and the features displayed, such as feature selection, among others. Consequently, the final layout was divided into two main parts: (1) the initial map and input parameters definition; and (2) the features dashboard.

5. Results and Discussion

In this section we use the prototype to demonstrate the model in action by performing the different activities of each use case. We also discuss some challenges and limitations as well as the current status and future developments.

5.1. The Model in Action

The prototype is used here to demonstrate the ability of the model to integrate UGsC data and perform the different activities of the use cases.

To use the prototype, the user initiates the process by defining a set of input parameters to query the UGsC sources. First the location of interest needs to be captured using one of three possibilities: (1) by inputting the latitude and longitude in the respective fields; (2) by searching on the map using the available zoom and pan tools and clicking on the location; or (3) by dragging a KML (Keyhole Markup Language) file containing locations to validate and using it as a reference to select. The third option is of more interest for the validation use case, in which a sample of locations is usually created by other applications and can thus be imported here for validation. The next step is to define the size of the square bounding box by entering the side length (e.g., 200 m) upon which the respective box is drawn on the map. The initiatives to query are also defined here, allowing the user to select one or more UGsC initiatives, followed by the request of the data. Figure 4 shows an example of the initial interface with a location already defined as well as the other input parameters.

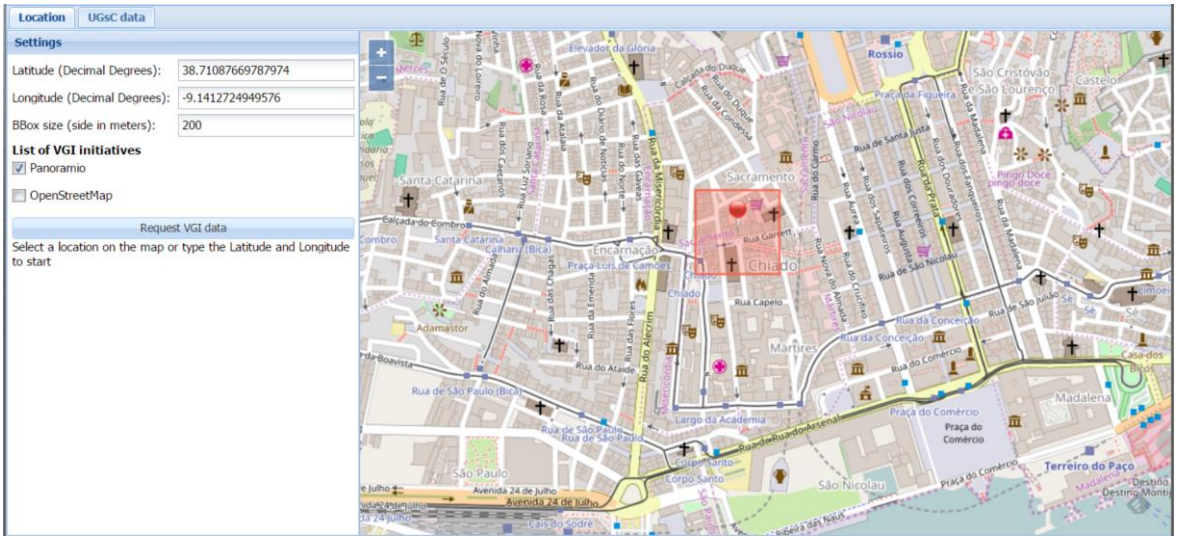

Figure 4. Initial interface for the photo interpretation use case (note: the pin and square represent, respectively, the selected location and the bounding box used in requesting data from the initiatives).

After obtaining the data, the features dashboard UI tab becomes available, allowing access to the data in a map format as well as additional information such as metadata on individual features, as shown in Figure 5. This example is more oriented to the classification use case, in which only the Panoramio source was queried. Each photo is represented by a point on the map and the user is able to access the respective photo and metadata by selecting individual features, including the photo URL giving access to the full size. The user can use all these available data as ancillary information and make the decision on which class best fits the unclear location.

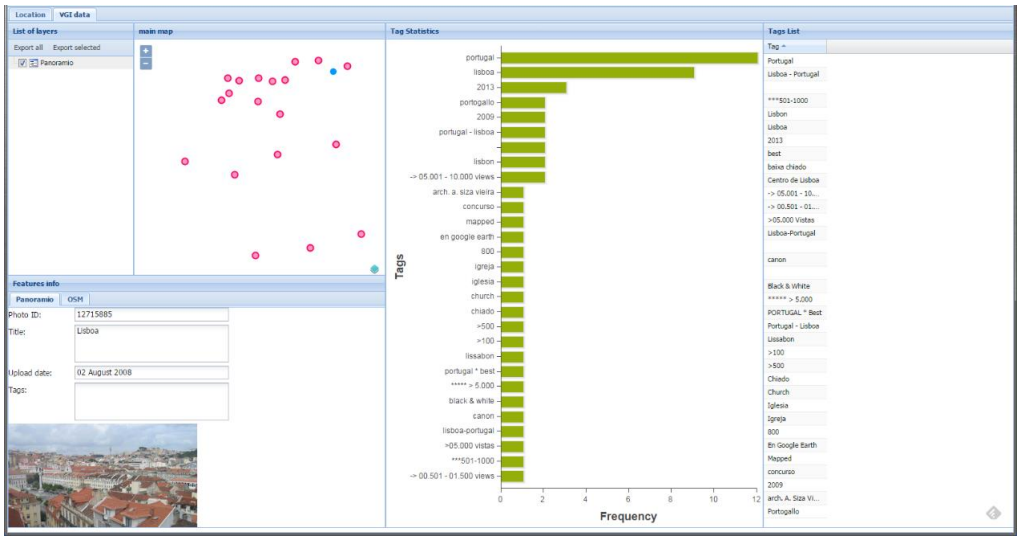

Figure 5. Features dashboard for the photo interpretation use case (note: red dots represent Panoramio photo locations, selected dots are highlighted in blue, and the green bars show the frequency of each tag).

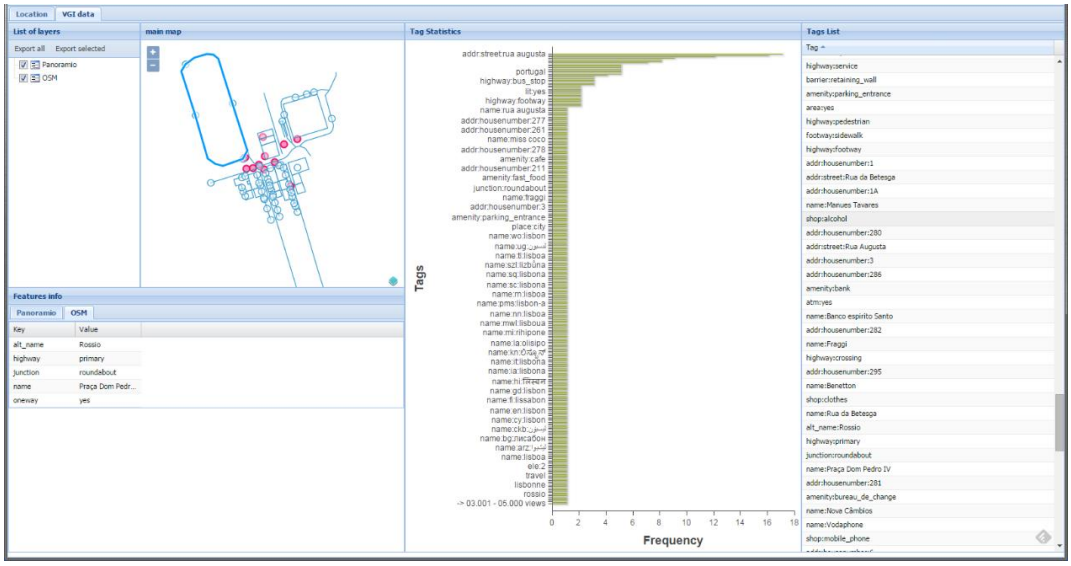

Figure 6. Features dashboard for the cartography validation use case (note: the light blue features depict OpenStreetMap (OSM) features, the red features represent Panoramio photos, and the highlighted blue features represent the OSM selected feature).

The validation use case requires more information to support the decision maker in validating the classification of a specific location. Here the user is probably interested in mixing data from different

sources regardless of the type of data. In this case we selected the Panoramio and OSM sources together as input parameters.

Figure 6 shows the features dashboard UI with all the features and respective metadata added to the different views. The main map is now showing features from both initiatives spatially integrated. The statistics chart is displaying the frequency of each tag (e.g., the number of features per tag), and the list of tags allows multiple selection of tags and features in both directions.

Looking at each box helps us to understand how these pieces of information can support the decision maker of the validation use case. By selecting features on the main map, their attributes are shown in the features information box. Figure 7 shows an example of the attributes of a selected OSM feature.

Figure 7. Detail of the Features info view for an OSM selected feature.

A list of tags, listing all the tags of the features that have been downloaded in a specific request, is also available. This list also gives the possibility of performing multiple selection of tags, seeing their respective features also selected on the map. Figure 8 shows this functionality when a multiple selection of tags is executed and all the features containing at least one of those tags is automatically highlighted on the main map.

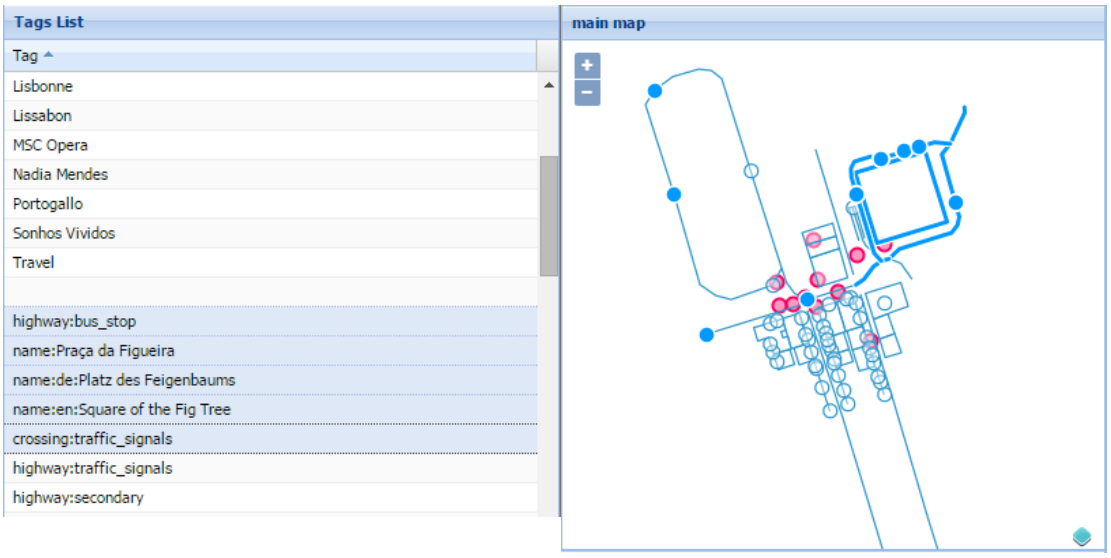

Figure 8. Selecting features by tag with multiple tags selected (note: the light blue features depict OSM features, the red features represent Panoramio photos, and the highlighted blue features symbolize features that have been selected).

A tag statistics box that shows the frequency of tags within the downloaded features is also available. Looking at this box in Figure 6 one can see the name of a street with the highest frequency and also a few tags with house numbers, indicating that this might be a residential area.

Another interesting operation is the possibility to drop a polygon onto the main map. This is particularly useful in this use case since LULC products are usually constituted by classified areas, or polygons, and gives the validator the ability to overlay the polygon containing the location being validated. Figure 9 depicts such a feature, showing the polygon overlaying the other features.

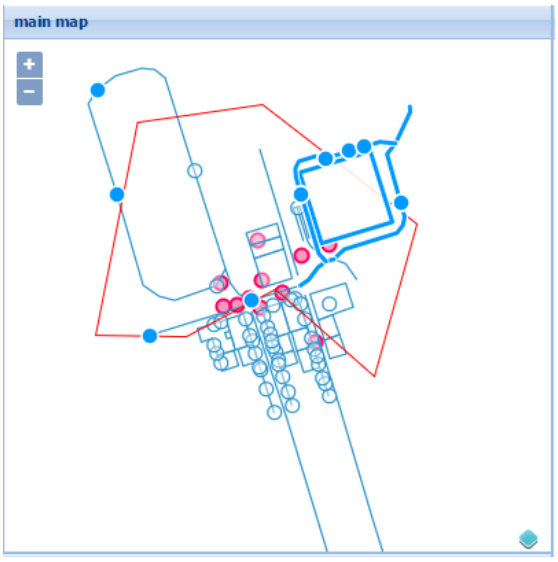

Figure 9. Main map view with a dropped overlaying polygon (note: the light blue features depict OSM features, the red circles represent Panoramio photos, the highlighted blue features represent the features that have been selected, and the red feature depicts the dragged polygon).

Finally, the user can export to KML either all the features present in the map or only the features that have been selected for further analysis in a desktop software by using the appropriate buttons on the list of layers box. The downloaded file can then be opened in any desktop GIS software that supports this format (e.g., QGIS).

Based on all these analyses the user is able to decide if the information provided is enough to support a decision and, if so, decide to validate the location positively or negatively.

5.2. Challenges and Limitations

The greatest concern in using UGsC resides in data quality. Several studies have been undertaken to understand the quality of this type of data as well as measures, indicators, and methods to evaluate that quality [15,46,47]. One characteristic is their heterogeneous nature with a spatial bias in the information. Rural areas have many fewer data than urban areas [15,48,49], and even inside urban areas a spatial bias exists with touristic and popular areas having more data than other less known locations [15]. We believe that the integration of different data sources helps to reduce the impact of this issue, but such investigation is outside the scope of this study. Additional measures specific to UGsC data have been proposed by different authors. Antoniou and Skopelity [50] provided indicators classified in four main categories: (i) data; (ii) demographics; (iii) socio-economic situation; and (iv) contributors. These indicators are of special importance when no authoritative data are available to use as reference. The integration of methods to measure such indicators in the prototype would be a valuable future improvement.

Regarding data access, most of the public APIs of the UGsC initiatives have restrictions in terms of number of requests a user can make, or the quantity of data that can be downloaded within a certain amount of time. This represents a constraint on using the prototype for larger areas or with very high frequency. Another important limitation is related to the semantics of tags. One of the advantages of some UGsC initiatives is to give enough freedom to citizens to classify uploaded data

with non-structured tags. On the other hand, these non-structured tags represent a key challenge when it comes to integration. Tags are related to the language, the region, or even the user environment. To overcome this limitation, ontologies would need to be properly developed and integrated, which is outside the scope of this study. The exponential availability of data produced by citizens is closely related to the introduction of the Web 2.0, the increasing availability of positioning equipment at a lower cost, and better and free imagery of the world. Such technologies are not available in all the locations of the world and consequently UGsC initiatives will present fewer available data, or even no data, for these locations. This phenomenon is identified as the Digital Divide [3] and represents a major limitation of the UGsC-Integrator and prototype for locations where such technologies are not used and data are scarce or non-existent as a result.

5.3. Current Status and Future Developments

This prototype used two initiatives to demonstrate the implementation of the integration model. In the future it can integrate new initiatives at any time by developing and implementing the respective reader and parser to contact, query, download, and integrate their features in the application, taking into account their specificities.

In terms of future research, we foresee the development of more use cases and the integration of more and different UGsC initiatives to increase the reliability and comprehensiveness of the platform. Although data conflation and fusion processes might reduce the level of detail of the information obtained by the integration of different initiatives to a certain extent, such tools might be available optionally on the platform, but further investigation is needed to determine their advantages. Analytical tools such as image processing to automatically remove useless photos, such as photos mostly covered by peoples' faces, and detect the predominant LULC class either for each photo or for a collection of photos with a certain area are an important upgrade. Finally, the development of a web service is planned. The main advantage would be related to the possibility of using the data resulting from the UGsC-Integrator directly in different and independent applications.

6. Conclusions

In this study we developed the architecture of a data integration model that combines diverse sources of UGsC in a common platform; this data integration model is to be used in the process of LULC databases production, more specifically, to help in the validation phase. From a comprehensive list of UGsC initiatives already identified by Elwood et al. [29], we identified and discussed the important characteristics and defined a set of minimum requirements that any UGsC source must meet to be included. A list of the current UGsC initiatives satisfying such requirements was also developed, and the similarities and dissimilarities identified were taken into account in the design of the model. The architecture defined was structured to allow the future evolution of the model by enabling the incorporation of new sources of UGsC as well as techniques that might already give some preliminary quality indicators and, ultimately, automate the decision making process by providing final quality indicators about the LULC database under evaluation.

A prototype application was used to demonstrate the implementation of the model in which the integration of data coming from different sources with different structures was verified using a common map. Additional information, such as tags and attributes, were also analyzed in an integrated approach to calculate statistics and allow the selection of features by tag. Two use cases were used to illustrate the model in action proving that the integration of data from different initiatives is possible. Other use cases not implemented here can also be identified: a landscape architect interested in studying a specific area from the landscape point of view might use available photos; a data/big data analyst interested in analyzing all of the available data for a given location can use the prototype to access and download raw data; an urban planner might use the prototype to access ancillary information to support the planning process, etc.

We hope to have demonstrated how diverse UGsC data can be integrated and how useful information to support decision making can be extracted.

Author Contributions: Jacinto Estima and Marco Painho conceived and designed the experiments. Jacinto Estima developed the software, performed the analysis and wrote the paper. Marco Painho contributed to the analysis and reviewed the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodchild, M.; Glennon, J.A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Digit. Earth* **2010**, *3*, 231–241. [CrossRef]
2. Goodchild, M. Commentary: Whither VGI? *GeoJournal* **2008**, *72*, 239–244. [CrossRef]
3. Sui, D.; Goodchild, M.; Elwood, S. Volunteered geographic information, the exa flood, and the growing digital divide. In *Crowdsourcing Geographic Knowledge*; Sui, D., Elwood, S., Goodchild, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 1–12.
4. Estima, J.; Painho, M. Exploratory analysis of OpenStreetMap for land use classification. In Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Orlando, FL, USA, 5 November 2013.
5. See, L.; Comber, A.; Salk, C.; Fritz, S.; Velde, M.V.D.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; Obersteiner, M. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE* **2013**. [CrossRef] [PubMed]
6. Pultar, E.; Raubal, M.; Cova, T.J.; Goodchild, M.F. Dynamic GIS case studies: Wildfire evacuation and volunteered geographic information. *Trans. GIS* **2009**, *13*, 85–104. [CrossRef]
7. Zook, M.; Graham, M.; Shelton, T.; Gorman, S. Volunteered geographic information and crowdsourcing disaster relief: A case study of the haitian earthquake. *World Med. Health Policy* **2010**, *2*, 6–32. [CrossRef]
8. Mooney, P.; Corcoran, P.; Winstanley, A. A study of data representation of natural features in OpenStreetMap. In Proceedings of the 6th GIScience International Conference on Geographic Information Science, Florence, Italy, 5–9 July 2010.
9. Hollenstein, L.; Purves, R. Exploring place through user-generated content: Using Flickr to describe city cores. *J. Spat. Inf. Sci.* **2010**, *1*, 21–48.
10. Loconte, P.; Rotondo, F. VGI to enhance minor historic centres and their territorial cultural heritage. In *Computational Science and Its Applications—ICCSA 2014*; Springer: Berlin, Germany, 2014; pp. 315–329.
11. Assessment of the Status of the Development of the Standards for the Terrestrial Essential Climate Variables. Available online: http://www.fao.org/gtos/doc/ECVs/T09/T09.pdf (accessed on 27 September 2016).
12. Caetano, M.; Mata, F.; Freire, S. Accuracy assessment of the Portuguese CORINE Land Cover map. *Glob. Dev. Environ. Earth Obs. Space* **2006**, *1*, 459–467.
13. Arsanjani, J.J.; Helbich, M.; Bakillah, M. Exploiting volunteered geographic information to ease land use mapping of an urban landscape. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, London, UK, 29–31 May 2013.
14. Arsanjani, J.J.; Helbich, M.; Bakillah, M.; Hagenauer, J.; Zipf, A. Toward mapping land-use patterns from volunteered geographic information. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2264–2278. [CrossRef]
15. Estima, J.; Fonte, C.C.; Painho, M. Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database. In Proceedings of the AGILE 2014 International Conference on Geographic Information Science, Castellón, Spain, 3–6 June 2014.
16. Estima, J.; Painho, M. Flickr geotagged and publicly available photos: Preliminary study of its adequacy for helping quality control of corine land cover. *Comput. Sci. Appl.* **2013**, *7974*, 205–220.
17. Estima, J.; Painho, M. Photo based volunteered geographic information initiatives. *Int. J. Agric. Environ. Inf. Syst.* **2014**, *5*, 73–89. [CrossRef]
18. Estima, J.; Painho, M. Investigating the potential of OpenStreetMap for land use/land cover production: A case study for continental portugal. In *OpenStreetMap in GIScience: Experiences, Research, Applications*; Arsanjani, J.J., Zipf, A., Mooney, P., Helbich, M., Eds.; Springer: Berlin, Germany, 2015; pp. 273–293.
19. Fonte, C.C.; Bastin, L.; See, L.; Foody, G.; Lupia, F. Usability of VGI for validation of land cover maps. *Int. J. Geogr. Inf. Sci.* **2015**, *4*, 1–23. [CrossRef]

20. Foody, G.M.; Boyd, D.S. Using volunteered data in land cover map validation: Mapping West African forests. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1305–1312. [CrossRef]
21. Foody, G.M. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sens. Environ.* **2010**, *14*, 2271–2285. [CrossRef]
22. Fritz, S.; McCallum, I.; Schill, C.; Perger, C.; See, L.; Schepaschenko, D.; Velde, M.V.D.; Kraxner, F.; Obersteiner, M. Geo-Wiki: An online platform for improving global land cover. *Environ. Model. Softw.* **2012**, *31*, 110–123. [CrossRef]
23. Hagenauer, J.; Helbich, M. Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 963–982. [CrossRef]
24. Arsanjani, J.J.; Vaz, E. An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *35*, 329–337. [CrossRef]
25. Perger, C.; Fritz, S.; See, L.; Schill, C.; Velde, M.V.D.; Mccallum, I.; Obersteiner, M. A campaign to collect volunteered geographic information on land cover and human impact. In *GI Forum 2012: Geovizualisation, Society and Learning*; Herbert Wichmann Verlag: Berlin, Germany, 2012; pp. 83–91.
26. Goodchild, M. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
27. Turner, A.J. *Introduction to Neogeography*; O'Reilly Media: Sebastopol, CA, USA, 2006.
28. Hudson-Smith, A.; Batty, M.; Crooks, A.; Milton, R. Mapping for the masses: Accessing web 2.0 through crowdsourcing. *Soc. Sci. Comput. Rev.* **2009**, *27*, 524–538. [CrossRef]
29. 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]
30. Stefanidis, A.; Crooks, A.; Radzikowski, J. Harvesting ambient geospatial information from social media feeds. *GeoJournal* **2013**, *78*, 319–338. [CrossRef]
31. Fischer, F. VGI as big data: A new but delicate geographic data-source. *Geoinformatics* **2012**, *5*, 46–47.
32. Brando, C.; Bucher, B. Quality in user generated spatial content: A matter of specifications. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 11–14 May 2010.
33. VGI-Net: A Collaborative Research Project. Available online: http://vgi.spatial.ucsb.edu/ (accessed on 27 September 2016).
34. Hull, R.; Zhou, G. A framework for supporting data integration using the materialized and virtual approaches. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, QC, Canada, 4–6 June 1996.
35. Wiederhold, G. Mediators in the architecture of future information systems. *Comput. Long. Beach. Calif.* **1992**, *25*, 38–49. [CrossRef]
36. Neill, C.J.; Laplante, P.A. Requirements engineering: The state of the practice. *IEEE Softw.* **2003**, *20*, 39–45. [CrossRef]
37. Sencha Ext JS (Version 4.2.2). Available online: https://www.sencha.com/products/extjs/#overview (accessed on 27 September 2016).
38. OpenLayers (Version 3.1.1). Available online: http://openlayers.org/ (accessed on 27 September 2016).
39. Apache HTTP Server Project. Available online: https://httpd.apache.org/ (accessed on 27 September 2016).
40. Brovelli, M.A.; Minghini, M.; Zamboni, G. Public participation GIS: A FOSS architecture enabling field-data collection. *Int. J. Digit. Earth* **2014**, *7*, 1–19.
41. Horanont, T.; Basa, M.; Shibasaki, R. Towards thematic Web services for generic data visualization amd analysis. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *I-4*, 147–150. [CrossRef]
42. Cozannet, G.L.; Bagni, M.; Thierry, P.; Aragno, C.; Kouokam, E. WebGIS as boundary tools between scientific geoinformation and disaster risk reduction action in volcanic areas. *Nat. Hazards Earth Syst. Sci.* **2014**, *14*, 1591–1598. [CrossRef]
43. Okladnikov, I.; Gordov, E.; Titov, A.; Bogomolov, V.; Martynova, Y. Application of web-GIS approach for climate change study. *EGU Gen. Assem.* **2013**, *15*, 6682–6692.
44. Simeoni, L.; Zatelli, P.; Floretta, C. Field measurements in river embankments: Validation and management with spatial database and webGIS. *Nat. Hazard.* **2014**, *71*, 1453–1473. [CrossRef]
45. Burdziej, J. A Web-based spatial decision support system for accessibility analysis-concepts and methods. *Appl. Geomat.* **2012**, *4*, 75–84. [CrossRef]

46. Haklay, M. How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Des.* **2008**, *37*, 682–703. [CrossRef]
47. Fonte, C.C.; Bastin, L.; Foody, G.; Kellenberger, T.; Kerle, N.; Mooney, P.; Olteanu-Raimond, A.M.; See, L. VGI quality control. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *3*, 317–324. [CrossRef]
48. Ma, D.; Sandberg, M.; Jiang, B. Characterizing the heterogeneity of the OpenStreetMap data and community. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 535–550. [CrossRef]
49. Neis, P.; Zielstra, D. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Int.* **2014**, *6*, 76–106. [CrossRef]
50. Antoniou, V.; Skopeliti, A. Measures and indicators of VGI quality: An overview. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *3*, 345–351. [CrossRef]