NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
Lisboa, Portugal

User Generated spatial Content sources for Land Use/Land Cover validation purposes: suitability analysis and integration model

A thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Systems
Specialization in Geographic Information Systems
by
Jacinto Paulo Simões Estima

Supervisor
Marco Painho, PhD

Lisbon, June 2015

Abstract

Traditional geographic information has been produced by mapping agencies and corporations, using high skilled people as well as expensive precision equipment and procedures, in a very costly approach. The production of land use and land cover databases are just one example of such traditional approach. On the other side, The amount of Geographic Information created and shared by citizens through the Web has been increasing exponentially during the last decade, resulting from the emergence and popularization of technologies such as the Web 2.0, cloud computing, GPS, smart phones, among others. Such comprehensive amount of free geographic data might have valuable information to extract and thus opening great possibilities to improve significantly the production of land use and land cover databases.

In this thesis we explored the feasibility of using geographic data from different user generated spatial content initiatives in the process of land use and land cover database production. Data from Panoramio, Flickr and OpenStreetMap were explored in terms of their spatial and temporal distribution, and their distribution over the different land use and land cover classes. We then proposed a conceptual model to integrate data from suitable user generated spatial content initiatives based on identified dissimilarities among a comprehensive list of initiatives. Finally we developed a prototype implementing the proposed integration model, which was then validated by using the prototype to solve four identified use cases.

We concluded that data from user generated spatial content initiatives has great value but should be integrated to increase their potential. The possibility of integrating data from such initiatives in an integration model was proved. Using the developed prototype, the relevance of the integration model was also demonstrated for different use cases.

Keywords

Land Use / Land Cover, Geographic Information Systems, User Generated Spatial Content, Integration Model, Spatial Data Integration

Resumo

Informação geográfica tem sido tradicionalmente produzida por agências de mapeamento e corporações, através de pessoas altamente qualificadas, bem como equipamentos de precisão e procedimentos dispendiosos, numa abordagem bastante onerosa. A produção de bases de dados de uso e cobertura do solo são apenas um exemplo da referida abordagem. Por outro lado, a quantidade de informação geográfica criada e partilhada pelos cidadãos através da Web tem vindo a aumentar exponencialmente durante a última década, resultante do surgimento e popularização de tecnologias como a Web 2.0, computação na nuvem, GPS, telefones inteligentes, entre outros. Esta quantidade de dados geográficos livres pode ter informações valiosas para extrair e assim abrir a possibilidade de melhorar significativamente a produção de bases de dados de uso e cobertura do solo.

Nesta tese explorou-se a viabilidade da utilização de dados geográficos, de diferentes iniciativas de conteúdo espacial gerado por utilizadores, no processo de produção de bases de dados de uso e cobertura do solo. Dados das iniciativas Panoramio, Flickr e OpenStreetMap foram explorados em termos de sua distribuição temporal e espacial, e da sua distribuição pelas diferentes classes de uso e cobertura do solo. Foi de seguida proposto um modelo conceptual para integrar dados de iniciativas de conteúdo espacial gerado por utilizadores baseado nas diferenças identificadas de entre uma lista abrangente de iniciativas. Finalmente, desenvolveu-se um protótipo de implementação do modelo proposto, o qual foi

então validado usando o protótipo para resolver quatro casos de uso previamente identificados.

Concluiu-se que os dados de iniciativas de conteúdo espacial gerado por utilizadores tem um grande valor, mas devem ser integrados para aumentar o seu potencial. A possibilidade de integração de dados de diferentes iniciativas num modelo de integração foi provada. Através do protótipo desenvolvido, foi também demonstrada a relevância do modelo de integração em diferentes casos de uso.

Palavras-chave

Uso e Cobertura do Solo, Sistemas de Informação Geográfica, Conteúdo Espacial Gerado por Utilizadores, Modelo de Integração, Integração de Dados Espaciais

List of Publications

List of peer-reviewed publications resulting from this thesis so far:

Jacinto Estima and Marco Painho (2015) Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal. In: Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., OpenStreetMap in GIScience: experiences, research, applications. ISBN:978-3-319-14279-1, PP. 273-293, Springer Press.

Jacinto Estima and Marco Painho (2014) Photo Based Volunteered Geographic Information Initiatives: A Comparative Study of their Suitability for Helping Quality Control of Corine Land Cover. International Journal of Agricultural and Environmental Information Systems 5(3): 75-92. doi: 10.4018/ijaeis.2014070105.

Jacinto Estima, Cidália Fonte and Marco Painho (2014) Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3.

Jacinto Estima and Marco Painho (2013) Exploratory analysis of OpenStreetMap for land use classification. 2nd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD) 2013. 5 – 8 November 2013. Orlando, Florida, USA. doi: 10.1145/2534732.2534734.

Jacinto Estima and Marco Painho (2013) Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. In Computational Science and Its Applications–ICCSA 2013 (pp. 205-220). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-39649-6_15.

Acknowledgements

Despite having the final responsibility of the work presented in this thesis, it would not be possible without the contribution of many persons to whom I would like to express my gratitude:

First and foremost I would like to express my gratitude to my adviser Professor Marco Painho. I am grateful for the guidance, support, patience and friendship. The work presented in this thesis would not be possible without his availability, continuous support and encouragement. I hope that our collaboration may continue for many years. I am also grateful for giving me the possibility to participate in the COST Action TD1202 – Mapping and the citizen sensor – funded by the European Union.

A special thanks to my friend and PhD colleague António José Silva for our discussions. They helped me a lot in developing the initial research idea. A big thanks to my friend Luis Calisto for the fruitful discussions we had especially during the phase of the prototype development. A special thanks to Professor Hosni Ghedira, Director of the Research Center for Renewable Energy Mapping and Assessment at the Masdar Institute of Science and Technology, Abu Dhabi, UAE, for his continuous encouragement and also for allowing me to continue this study in parallel with my main job there for the last three years.

Finally to my family and friends for their continuous support. To my wife Ana Cristina Estima for all the patience, love and encouragement throughout this journey. To my

sun João Estima that is my inspiration and the reason to move forward. To my parents for all the opportunities and belief that allowed me to reach this stage.

List of abbreviations

AA	Accuracy Assessment
AGI	Ambient Geographic Information
API	Application Programming Interface
CASA	Centre for Advanced Spatial Analysis
GI	Geographic Information
GE	Google Earth
GPS	Global Positioning System
LC	Land Cover
LU	Land Use
LULC	Land Use / Land Cover
NCGIA	National Center for Geographic Information and Analysis
OSM	OpenStreetMap
PFM	Public Facility Management
SOA	Service-Oriented Architecture
SOA-SDI	Service-Oriented Architecture for Spatial Data Integration
SR2S	Safe Route-to-School
UGC	User Generated Content
UGsC	User Generated spatial Content

USGS	United States Geological Survey
VGI	Volunteered Geographic Information
VIEW-IT	Virtual Interpretation of Earth Web-Interface Tool

Short index

Index

List of Figures

List of tables

1. Introduction

In the last years, the amount of Geographic Information (GI) created and shared by citizens through the Web has been increasing exponentially. The emergence and popularization of some technologies – Web 2.0, cloud computing, GPS, smart phones, among others – have transformed, and still are, the way how geographic data are produced, stored and used (Sui, Goodchild, & Elwood, 2013). The literature shows that research has been conducted trying to explore the enormous potential that this type of data seems to be hiding and find possibilities of using it in the solution of real world problems (e.g.: Estima & Painho, 2013a; Goodchild & Glennon, 2010; Hollenstein & Purves, 2010; Mooney, Corcoran, & Winstanley, 2010; Pultar, Raubal, Cova, & Goodchild, 2009; See et al., 2013; Zook, Graham, Shelton, & Gorman, 2010).

One important area where this data sources could be very helpful is in the Land Use/Cover (LULC) database production. In this matter, interesting results have already been accomplished (J Jokar Arsanjani, Helbich, & Bakillah, 2013; Jamal Jokar Arsanjani, Helbich, Bakillah, Hagenauer, & Zipf, 2013; Estima, Fonte, & Painho, 2014; Estima & Painho, 2013a, 2013b, 2014, 2015; Fonte, Bastin, See, Foody, & Lupia, 2015; Foody & Boyd, 2013a, 2013b; Foody, 2010; Fritz et al., 2012, 2009; Hagenauer & Helbich, 2012; Jamal Jokar Arsanjani, Helbich, Bakillah, Hagenauer, & Zipf, 2013; Jamal Jokar Arsanjani & Vaz, 2015; Perger et al., 2012). Nevertheless, the literature also shows important gaps providing us with an excellent opportunity to contribute to this interesting topic. Some particularities of this type of data, described later in this document, make their use very challenging and therefore this study is designed to explore different sources of User Generated spatial Content (UGsC) and develop a data model able to integrate them so they can be used to help in the production of LULC databases.

1.1. Identification and contextualization of the problem

GI has been produced by mapping agencies and corporations and sold to users as paper maps or atlases (Goodchild & Glennon, 2010). This approach is very expensive since it requires expert people as well as expensive precision equipment and procedures. Consequently priority is given to the most important and unchanging geographic themes and those with multiple applications relegating the other ones for a second plan (Goodchild, 2008).

One of those examples are the LULC databases that play a very important role in a vast number of research fields (Caetano, Mata, & Freire, 2006; Fritz et al., 2009; Herold, 2009). Its production is mainly based on interpretation and classification of remote

sensing data, made by highly trained and skilled people (Herold, 2009) and goes through a phase process since the planning and data acquisition, pre-processing, analysis/classification, to the final product and documentation (J. Cihlar, 2000). Although all the phases are very important, the validation phase has a particular and very important goal: to provide the final product with quality indicators to those who want to use it. This validation is made by confronting the produced cartography with reference information assumed as true, that includes, among other sources, “ground truth” collected directly from the filed in pre-selected sites (Caetano et al., 2006). This in situ ground measurements acquisition represents a major limitation caused by its high cost, both in terms of money and time (Strahler et al., 2006).

On the other hand, since 2005, with the introduction of the Web 2.0, the spatial data produced by citizens became exponentially available over the Web. This is due not only to the increasing availability of positioning equipment’s at a lower cost, better and free imagery of the world, among others, but also to the willingness of private citizens to contribute for several reasons (Elwood, Goodchild, & Sui, 2012; Heipke, 2010).

The amount of produced data is of very different nature and one of the most important characteristics is the local knowledge of its contributors that know their surroundings better than any outsider (Heipke, 2010). The availability of this quantity of data provides us with a great opportunity to explore new ways to use it for helping LULC production. While the major advantages are associated with its quantity, temporal coverage and size (Leung & Newsam, 2010), this big quantity of data is very heterogeneous and scattered over different projects with completely different data structures, making its integration consequently very difficult.

1.2. Research objectives

Considering the problem stated in the previous section, the aim of this work is to **explore the suitability of data from different UGsC initiatives with different formats and structures to be used in the production of LULC databases, and propose a data model to integrate these data from different sources and structures**. The motivation for this main objective is related with the following research questions:

1. Are the data from UGsC initiatives feasible to help in the production of LULC databases?
2. Which types of geographic data from UGsC initiatives are more suitable to use in the production of LULC databases?
3. Is it possible to integrate these data in a common data model/platform?

1.3. Importance and relevance of research

The exponential availability of geographic data from diverse UGsC initiatives in the last few years has increased the motivation of the research community to explore their potential and usefulness in the solution of real world problems.

Two main strategies have been followed and examples are provided in chapter 2: 1) to ask volunteers to explicitly contribute to specific projects or 2) to explore data already available in different UGsC initiatives. The first strategy needs volunteers to be available and willing to contribute while the second explores existing data already contributed to other initiatives for different purposes. Experiments using the first approach are described in section 2.4 (Clark & Aide, 2011; Fritz et al., 2009).

Focusing on the second strategy, more connected with this study, the literature shows that research already conducted only uses data from one or two different UGsC initiatives (Arsanjani et al., 2013; Hollenstein & Purves, 2010; Kisilevich, Krstajic, Keim, Andrienko, & Andrienko, 2010; Leung & Newsam, 2010; Zielstra & Hochmair, 2013; Zook et al., 2010). Therefore there is no research related with the integration of data from different UGsC sources with diverse structures for the purpose of helping the LULC databases production process.

As already stated before in this document, the major advantages of contributed data are associated with its quantity, temporal coverage and size (Leung & Newsam, 2010) but the fact that it is scattered over several different projects represents a major limitation. This study attempts to bridge the gap and contribute to the scientific community by exploring the suitability of different UGsC initiatives for LULC database production and proposing a data model for the integration of these data from different sources and structures. This data model will indirectly contribute to the cost and time reduction of the LULC production and will also increase even further the value of this type of data.

1.4. Methodology

The approach followed in this thesis is shown in Figure 1.

The first part was to perform an in depth study of the literature in terms of UGsC initiatives. As previously mentioned, UGsC initiatives have been growing in the last years and so the number of research projects trying to explore them in the solution real world problems. We looked at the literature and an in depth review, with a particular focus on the use of UGsC for the specific application on LULC databases production, is

presented. This review allowed us to understand what has been already studied and identify existing gaps that could be explored.

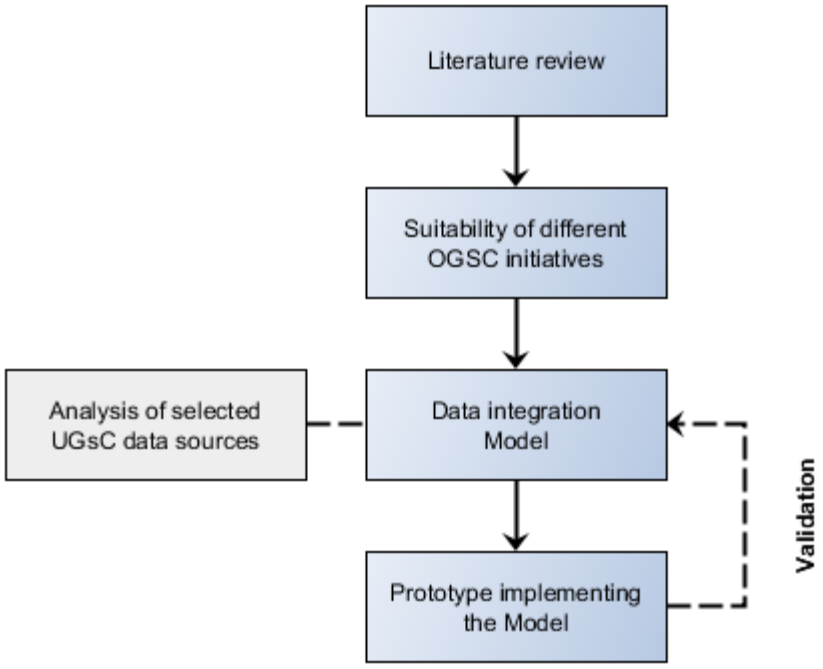

Figure 1 - Thesis methodology

We then conducted a preliminary analysis of the feasibility of some of the most well-known UGsC initiatives to be used in the process of LULC databases production. We explored their spatial and temporal distribution as well as their matching rate against an official and validated LULC database. These analyzes allowed us to understand how suitable these sources of spatial data are and identify already advantages and challenges of using them for LULC database production.

The following part was to develop an integration model to integrate data from different UGsC initiatives. To develop such a model as comprehensive as possible, it was very important to take into account the LULC needs and also the characteristics of available

and suitable UGsC initiatives. Therefore, taking into account these LULC needs, the first step was to discuss the requirements that any dataset would need to be used for the purpose of helping in LULC database production, leading to the definition of a set of minimum requirements. Then a comprehensive list of UGsC initiatives was developed and a subset following these minimum requirements was extracted and their structural dissimilarities analyzed. The integration model was then developed taking into account these dissimilarities.

To assess the developed model, a prototype was then developed and implemented, in this last part of the methodology. We started the development of the prototype by defining its requirements. To do so we identified and analyzed a set of four use cases reflecting four potential users of the model and prototype. The next step was to develop the prototype based on the defined requirements and use it to solve the defined use cases and thus validate the model. This validation was of extreme importance to test not only the model but also to demonstrate its usefulness.

1.5. Thesis organization

This thesis is organized as follows.

Chapter 1 introduces the research topic of this study. It starts by identifying and contextualizing the research problem followed by the definition of the research objectives and a discussion on the importance and relevance of the study. The chapter ends with the description of the methodology and the organization of the thesis.

Chapter 2 presents an in depth review of the literature related with UGsC. First a definition is provided followed by the description of UGsC initiatives already explored for

different applications. A revision of the LULC databases production process is also provided followed by the description of studies already conducted to use data from UGsC initiatives for LULC related matters.

Chapter 3 describes a set of preliminary studies on the use of data from different UGsC sources to help in the process of LULC database production. Different sources of UGsC were used to explore their suitability to be used in the production of LULC databases.

Chapter 4 presents a data model that integrates different sources of UGsC to help in the process of LULC databases production. A list of UGsC sources that follow a set of defined minimum requirements is provided and their structural dissimilarities discussed. The model is then developed based on these dissimilarities.

Chapter 5 describes the development of a prototype implementing the proposed integration model. It starts by identifying four use cases to define the system requirements based on which the prototype was developed. The prototype was then used to solved those use cases and thus validate the model.

Chapter 6 summarizes the conclusion of this research. The drawn hypotheses are discussed and the main contributions to the scientific community presented, followed by the discussion of some limitations of this study as well as the identification of future research directions.

2. State of art

2.1. Introduction

In this chapter we review the different definitions associated with GI produced by citizens and explore some of the most well-known related initiatives reported and studied in the literature. We then analyze the production of LULC databases and discuss the work already conducted using this type of UGsC data in their context. We finish debating spatial data integration as well as the concepts and methods involved.

2.2. User Generated spatial Content

2.2.1. Definitions

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.*" (Goodchild, 2007). 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.*" (Turner, 2006). 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 (Hudson-Smith, Batty, Crooks, & Milton, 2009).

Despite some differences between these terminologies (Elwood et al., 2012), they are all related with 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", a function that for centuries has been reserved exclusively to official agencies (Goodchild, 2007).

More recently, Stefanidis et. al (2011) 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, e.g. 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).

Also Fischer (2012) argued that, in some cases, when VGI is used for different purposes than those for which volunteers have contributed, it can be seen as a not-so-Volunteered Geographic Information and had termed this as involuntary geographic information (iVGI).

We are introducing here a new term, called User Generated spatial Content (UGsC) to integrate all the previous definitions. Moreover, 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.

2.2.2. Relevance

The relevance of the topic has been proved by the growing number of meetings and workshops held in recent years. The first of its kind happened in December 2007 held in Santa Barbara, CA, organized by the National Center for Geographic Information and Analysis (NCGIA), Los Alamos National Laboratory, the Army Research Office and The

Vespucci Initiative where some important topics were discussed and some position papers published[1].

The United States Geological Survey (USGS) organized, in 2010, a workshop on Volunteered Geographic Information, held in Herndon, VA, resulting in a set of publically available presentations and breakout session minutes[2] documenting the activities.

VGI workshops have been also offered by several conferences on GIS Science. The International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL), the AGILE Conference on Geographic Information Science or the International Conference on Geographic Information Science (GISience) are just a few examples.

More recently, an initial training network, called "geocrowd", funded under an FP7 - People Marie Curie Actions by the European Commission, was launched aiming at "*establishing a fertile research environment by means of a training network that will promote the GeoWeb 2.0 vision and advance the state of the art in collecting, storing, analyzing, processing, reconciling, and making large amounts of semantically rich user-generated geospatial information available on the Web*"[3]. Other two projects under the European Cooperation in Science and Technology (COST) framework were funded by the European Union: 1) the COST action IC1203[4] - European Network Exploring Research into Geospatial Information Crowdsourcing: software and methodologies for

[1] The position papers are available at http://ncgia.ucsb.edu/projects/vgi/participants.html

[2] Available at http://cegis.usgs.gov/vgi/results.html

[3] Extracted from http://www.geocrowd.eu/

[4] http://www.cost.eu/COST_Actions/ict/Actions/IC1203

harnessing geographic information from the crowd (ENERGIC) and 2) the COST action TD1202[5] - Mapping and the citizen sensor.

All these research initiatives and activities demonstrate not only the interest of the research community but also how relevant this topic is to the research agenda.

2.2.3. Historical overview

The participation and contribution of citizens in this filed is not new. Various examples are documented like teachers and school children contributed to land use surveys of Britain in the 1930s, or the urban residents involved in the Bunge's "Geographical Expeditions" in 1971 (Elwood et al., 2012). Another interesting initiative, that started around 1999 and is still currently active, is the portal established by the USGS Earthquake Hazards Program for earthquake mapping called ''Did you feel it?'' (http://earthquake.usgs.gov/earthquakes/dyfi/) where people affected by earthquakes could provide information about their experiences regarding its position in the geographical space (Heipke, 2010).

The turning point for an exponential growth of volunteer's participation occurred in 2005 with the development and introduction of Google Maps and its Application Programming Interface (API). This aligned with the Web 2.0 technology have made a revolution providing users with the possibility of embedding their own varieties of Google Map's in their web pages (Batty, Hudson-Smith, Milton, & Crooks, 2010), and along with the availability of cheaper positioning devices combined with camera and mobile or smart phones, fine resolution-imagery, broad band communications, among other

[5] http://www.cost.eu/COST_Actions/ict/Actions/TD1202

improvements, empowered citizens to produce and share their own maps (Elwood et al., 2012; Heipke, 2010).

After Google Maps and the Web 2.0, several VGI projects started and have been contributing since then to the increasingly amount of available spatial data over the Web that exists nowadays. In 2009, an inventory made by Elwood et al. (Elwood et al., 2012) counted ninety-nine VGI initiatives, 70 percent of them started after 2005 against 20 percent that took place before that (Table 1). One of the first initiatives still active, based on Google Maps is Wikimapia (http://wikimapia.org), where people with an Internet connection can select any place in the world map and provide its description along with its boundaries, under the motto "Let's describe the whole world". Its philosophy is adapted from the successful Wikipedia project, where anyone can contribute with content, and a group of volunteers monitor the results checking for accuracy and significance (Goodchild, 2007).

Table 1. Inventory of VGI Initiatives in 2009[6]

Date initiated	Percentage
Pre - 2000	6%
2000 - 2004	14%
2005 - 2009	73%
Unable to identify	7%

OpenStreetMap (OSM, http://www.openstreetmap.org/) is another well-known VGI project developed by the OpenStreetMap Foundation that aims at providing free geographic data, such as street maps, to anyone. Users collect data (including topographic data) mostly with GPS or GPS enabled equipment, upload it to the OSM Web page, and complete it with descriptions, names and other attributes. The data is

[6] Adapted from Elwood et al. (Elwood et al., 2012)

then available to anyone in the form of render maps and other services, including the possibility to download it in vector format (Elwood et al., 2012).

On 6th March 2005, the Geograph initiative launched the Geograph website that aimed to collect and publish online, at least one representative photograph (geograph) per grid square for Great Britain and Ireland. By the end of March, 1 thousand photos had already been uploaded. Since then, the number of submitted photography's has been increasing significantly. One million images by October 2008, two million by August 2010, three million by the end of June 2012 and four million in early June 2014 (Geograph, 2012). Data from this initiative as well as data from the Flickr initiative were used by Leung and Newsam (2010) to derive maps from what-is-where from large collections of georeferenced photo collections. According to the authors, photos from the Geograph initiative were more accurate than those from the Flickr initiative because their contributors were contributing with the specific intention of geo-visually annotate Great Britain and Ireland.

In 2007, Google launched MyMaps, allowing users to create lines and shapes, embedding text, photos and videos with a simple drag and drop interface, based on Google Maps. Hudson-Smith et al. (Hudson-Smith, Crooks, Gibin, Milton, & Batty, 2009) argue that this was probably one of the most important innovations in mapping since the development of GIS. The Centre for Advanced Spatial Analysis (CASA) has also developed a set of tools allowing the non-professional user to integrate their data. Google Map Creator (GMapCreator) enables users to simplify thematic mapping in Google Maps. The London Profiler website (http://www.londonprofiler.org/) is a resource where public data from the public domain can be displayed over Google Maps and GmapCreator and it plays an important role in preparing the thematic maps for

displaying. Another tool developed by CASA is a ”place to put maps” called Map Tube (http://www.maptube.org/), where users can share information in the way of thematic maps produced with the GmapCreator, and based on the generic idea of YouTube (Hudson-Smith, Crooks, et al., 2009).

However, there are initiatives of a different kind with citizens playing a more passive role in terms of contributing with geographic information. In these initiatives, although data has not been contributed with the specific purpose of extracting geographic information, certain spatial characteristics, such as the geographic location of features or assets, are present. We classify these initiatives as a type of UGsC initiatives.

The name Flickr it is today well known in the Internet world. It is an initiative started in 2004 described as an online application that aims at "*…help people make their photos available to the people who matter to them*" and "*…enable new ways of organizing photos and video*"[7]. Flickr photos are stored in databases along with some additional information in the form of tags. Some information is automatically saved (e.g., contributing user, image metadata, and time of upload) and some other is introduced optionally by the user (e.g., title, caption, user restrictions, and a set of textual tags that best describes the photo). Spatial references can also be saved with the photo in the form of a special geotag that stores latitude and longitude (Hollenstein & Purves, 2010). Figure 2 shows that the number publically available Flickr photos has been increasing over the years and in 2012 where uploaded about 40 million of photos per month.

Some of these photos have latitude and longitude tag values which means that they are geo-referenced or "geotagged". There is no available information on how many of these photos are geotagged but in 2010, Kisilevich et al. (Kisilevich et al., 2010) downloaded a

[7] From the Flickr project website: http://www.flickr.com

total of 86,314,466 entries of geotagged Flickr photos to study peoples' activities using geotagged photo collections. This number is a very good demonstrator of the potential these resources may hide.

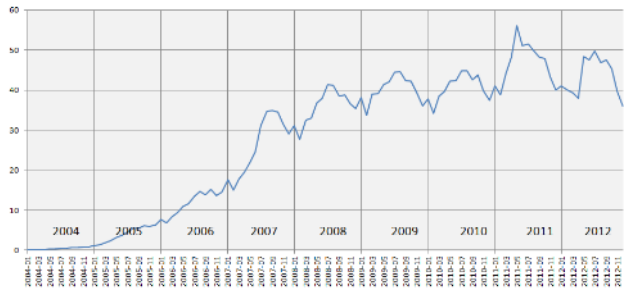

Figure 2 - Millions of photos uploaded per month – Jan. 2004 to Dec. 2012[8]

Started in 2006, Twitter is a very well-known online initiative that allows users to create and share ideas and information instantly using short messages with a maximum of 140 characters (Twitter, 2014). The added value of such short messages is related with the possibility of carrying location information and also the real-time nature of each tweet (message). This has been especially important for disaster detection, communication and response (Adam & Muraki, 2011; Funayama, Yamamoto, Tomita, Uchida, & Kajita, 2014; Mills & Chen, 2009; Sakaki, Okazaki, & Matsuo, 2010; Spence, Lachlan, Lin, & del Greco, 2015), but also for other applications such as political elections (Reips & Garaizar, 2011; Tsou et al., 2013), crime (Gerber, 2014), health (Signorini, Segre, & Polgreen, 2011), just to name a few.

In the private sector domain, the HD TrafficTM initiative from TomTom (http://www.tomtom.com/en_gb/services/live/hd-traffic/) aims at providing instant information about traffic to its customers, based on data collected from car drivers phones and can be regarded as a crowdsourcing system, where the crowd is part of the

[8] Souce: http://www.flickr.com/photos/franckmichel/6855169886/)

group of passive mappers (Heipke, 2010). Another interesting initiative, a real-world outdoor treasure hunting game, is the Geocaching game (http://www.geocaching.com/), where Players try to locate hidden containers, called geocaches, using GPS-enabled devices and then share their experiences and photographs in the website with other geocachers. The location of the geocachers is presented in a map based on Google Maps.

2.2.4. UGsC challenges and issues

The interesting and important initiatives presented in the previous section are only a part of the most well-known VGI initiatives and prove the exponential growth of spatial data availability over the web, but further research in the GIScience domain is needed to maximize their potential. How can we integrate this kind of data with authoritative data to fill gaps in spatial data infrastructure augmenting, updating, or completing it (Elwood, 2008a; Goodchild & Glennon, 2010; Heipke, 2010; Sui & Goodchild, 2011). Are the existing structures and practices for spatial data collection, retrieval, validation, and dissemination appropriate in this new context (Elwood, 2008b)? What types of geographic information are the most suited for acquisition through the efforts of volunteers (Goodchild & Glennon, 2010)?

Cowen, in the position paper presented at the VGI Workshop introduced some existing initiatives involving private citizens contributing to national mapping mostly from private agencies (Cowen, 2007). Google Inc. has enlisted private citizens in India to create content for Google Map products, and has also formed a business relationship with states in Australia to provide parcel level geocoding across the country. Governmental agencies should conduct such a practice by themselves.

These questions are a small set of issues related with the acquisition, integration and management of spatial data, but there are much more already formulated by the research community regarding data quality, legal and ethical issues, the digital divide, social impacts, among many other (Elwood et al., 2012; Elwood, 2008a; Goodchild & Glennon, 2010; Goodchild, 2007; Heipke, 2010; Kuhn, 2007; Sui, 2007, 2008).

2.3. Land Use/Cover data

Cihlar and Jansen (Josef Cihlar & Jansen, 2001) referred Baudiles and Szejwach in their paper to describe that Land Cover (LC) and Land Use (LU) are two key elements that represents respectively natural and human-related environments. LC attempts to characterize the biophysical features while LU is more related with the human interaction with these natural features. Despite some differences, both are related with characterization of land that plays a very important role in a vast number of research fields, such as LULC monitoring and modeling, monitoring of tropical deforestation, climate changes, among others, at both global, regional and local scales (Caetano et al., 2006; Fritz et al., 2009). Its production is mainly based on interpretation and classification of remote sensing data, made by highly trained and skilled people.

2.3.1. Land use/cover production

According to Cihlar (J. Cihlar, 2000) LULC production from satellite data consists of four main steps: data acquisition, pre-processing, analysis/classification, and product generation and documentation. Data acquisition is related with the acquisition of remote sensing data used as the base for the classification process. The pre-processing phase deals with a way to present the data in a proper format to extract information.

Analysis/classification is related with the extraction and classification process while the product generation and documentation deals with the final steps in the conclusion of the final product as well as its appropriate documentation.

The analysis/classification phase, beyond the analysis/classification itself must end with the validation, so called Accuracy Assessment (AA). This is a very important task and it aims to offer map quality indicators in order to provide the cartography with a degree of confidence to those who wants to use it. This AA is made by confronting the produced cartography with reference information assumed as true such as aerial photography; satellite imagery with better resolution than those used in production; and field work (M. Caetano et al., 2006). Magnussen, referred by Cihlar (J. Cihlar, 2000), states that the AA needs to contain "ground truth" as part of the sampling design. This field work increases the cost and the time consumption of the LULC production and can easily become unfeasible.

Due to these cost and time constrains, LULC databases are usually more focused on the most important themes and those with multiple applications, leaving behind those considered "less important". The time between updates or new productions is also a critical factor, but once more, as a consequence of production costs, it is stretched and the databases become outdated quickly (Goodchild, 2008).

Cihlar (J. Cihlar, 2000) concluded in his paper that "*The research agenda needs to address the best ways of taking advantage of the new capabilities and, importantly, the ways of resolving problems identified during the production of the land cover maps over large areas*". Therefore, VGI initiatives should be investigated to evaluate their adequacy in the LULC production processes.

2.4. UGsC and Land Cover mapping

As stated before, VGI has been increasingly used to research novel applications for different areas, including LULC database production. In this particular domain two different approaches have been used so far: 1) asking volunteers to actively contribute to a specific project such as the validation of global land cover datasets (Fritz et al., 2009; Perger et al., 2012) , and 2) using data contributed for other purposes/projects to extract valuable information and develop new ways to use it in this domain (Estima & Painho, 2013a, 2013b, 2014).

Geo-Wiki.Org (Fritz et al., 2009) is a project that fits in the first approach, described as a global network of volunteers who wish to help improve the quality of global land cover maps. “GLC-2000”, “MODIS”, and “GlobCover” global land cover databases are overlaid on a platform based on Google Earth (GE) and their areas of divergence highlighted. Then, a network of registered volunteers helps to solve these discrepancies using their local knowledge along with available GE satellite imagery and other ancillary data coming from other VGI projects such as pictures from Panoramio (http://www.panoramio.com/) and Degrees of Confluence Project (http://www.confluence.org/).

Another example is the Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) initiative based on GE high-resolution imagery to collect LULC reference data (Clark & Aide, 2011). It was tested with a small group of selected users acting as volunteers and not yet in a real crowdsourcing environment. Nevertheless they found important issues with using GE and its satellite imagery, e.g. the legal restrictions in the free use of the

Google Maps/Earth APIs and that some classes that cannot be discriminated with the available imagery (e.g. different annual crops).

In these examples, volunteers need to be available to contribute to these specific projects and they also need to have some familiarity with these tools, which might be discouraging for some groups of participants. To overcome this difficulty, some projects occasionally use contests and a mechanism of rewards to increase contributions and participation (Fritz et al., 2012; Perger et al., 2012).

Using the second aforementioned approach, some experiments were conducted by Leung and Newsam (2010) to derive maps of what-is-where from large collections of georeferenced photos in an automated way. In this initial work the authors derived LC classifications from georeferenced image collections for locations where ground-truth was available. The aim was to evaluate the quality of the results obtained from the automatic classification by comparing them with the available ground truth. They achieved a classification accuracy of approximately 75%.

Another interesting work was conducted by Estima and Painho (2013b) to explore the possibility of using Flickr photos as a source of ground-truth data to help in the accuracy assessment phase of LULC production. Using continental Portugal as the study area and CORINE (coordination of information on the environment) Land Cover (CLC) as a reference LULC database, the authors explored all the publically available and geotagged Flickr photos in terms of their temporal and spatial distributions and their distribution over the different CLC classes. The number of photos and their temporal resolution were the most positive aspects whereas their asymmetry and irregular distribution over different CLC classes the most negative. They concluded stating that

this could be a valuable source of ground truth data if combined with other sources but could not be used alone.

Foody and Boyd (2013) used two sources of volunteered data to illustrate the potential of amateur or neogeographical activity in map validation. They used photographs acquired from an internet-based collaborative project and interpreted by other volunteers to evaluate the Globcover map's representation of tropical forests in West Africa. They confirmed the potential value of VGI projects, such as the Degrees of Confluence project, for the provision of useful, spatially extensive, data to support map evaluation.

As already mentioned, the OSM initiative is one of the most well-known and studied VGI initiatives in the literature. The research that has been conducted to use data from the OSM initiative for LULC mapping purposes is quite extensive.

The possibility of using VGI data to replace training data acquired from in-site visits in the process of LULC classification was investigated by Arsanjani, Helbich and Bakillah (2013). Using the city of Koblenz, Germany, as the study area, they applied a supervised classification approach to classify data from the RapidEye sensor, and they used data downloaded from the OSM project as field measurements to select the most optimal training sites. They performed a comparison of the resultant LU map with the Global Monitoring for Environment and Security Urban Atlas (GMESUA) map achieving a Kappa index of 89%, which proves that OSM data is suitable to use as a source for training site definition. They also stress that the quality of VGI is heterogeneous and location-dependent, and they recommend checking the amount of contributions and also considering other VGI data, such as Flickr photos.

Another study investigated a new approach to generating land-use patterns from VGI without applying remote-sensing techniques and/or engaging official data (Jamal Jokar Arsanjani et al., 2013). Using OSM datasets and Vienna, Austria, as the study area, the authors applied a Hierarchical GIS-based decision tree approach to classify and segment parcels. The results were evaluated by conducting a texture-variability analysis of the LU maps generated using each dataset, and producing a confusion matrix to compare each LU class in the two datasets. Results of the texture analysis showed that the LU patterns derived from OSM data are richer than those derived from GMESUA. The confusion matrix showed a high level of agreement between the two classifications but this decreased when we move from level 1 towards the more detailed level 3. Although they conclude that VGI can be a potential data source for mapping LU patterns, they only used one source of VGI, OSM, and they did not test any other sources. Nevertheless, they pointed out as advantages of such an approach that no inputs from remote-sensing or any other administrative data were used, no financial cost exists as the OSM data is freely available and no field work was required, a number of incorrectly labeled features in the GMESUA were identified when OSM was incorporated, and the process of updating LU maps is facilitated due to the updating rate of OSM while GMESUA requires time and high financial costs to be updated by authorities.

A different approach was previously proposed by Hagenauer and Helbich (2012). They applied Artificial Neural Networks (ANNs) and Genetic Algorithms (GAs) as a machine learning methodology to delineate continuous urban areas using all the information diversity of OSM, where a large set of potential OSM at-tributes was derived for inductive learning. Using OSM and GMESUA data, they applied this methodology to 42 randomly selected GMESUA urban regions and analyzed the significance of the

attributes used and the performance of the mod-el. The model performed comparatively well for most regions, with a few remark-able exceptions. The study shows that if enough OSM data for reasoning is pre-sent, urban patterns can be predicted to a large extent. This approach could be very useful to help map continuous fabric classes, from OSM data, for LULC databases.

The representation of natural features in OSM was also explored by Mooney, Corcoran and Winstanley (2010), who examined the level of detail present in the representation of such polygon features. They tried to verify if there was enough detail in the representation of those features to provide a high-quality spatial representation. They used data for Austria, Estonia, Switzerland, Bretagne, Lower Saxony, Iceland, Ire-land, and Scotland to calculate the statistical distribution of the mean distance between connected vertices of polygons. They found that many of the features are under-represented, with a small number of vertices used to delineate them, while some of them might be considered over-represented (e.g. small urban green spaces and golf courses). Some OSM data collection characteristics, such as the different GIS skill levels of OSM volunteers or the differences in accuracy of equipment and methods used, influence the under-representation of some features. These under-represented features have a serious impact on using OSM data in certain Earth science applications, mainly those that use OSM as ground-truth data. They recommend that the quality of the OSM representation of “natural” polygons and other features should be established against a recognized ground-truth dataset.

In this sense, other authors have been exploring the quality of OSM data that are of interest for LULC database production. Barron, Neis and Zipf (2014) developed a comprehensive framework for intrinsic OSM quality analysis that included the logical

consistency of “natural” and “landuse” polygons. They developed a tool to generate information about OSM data quality for a selectable area without a reference dataset but using only OSM's data history. This tool intends to help users to assess the OSM data quality of a given area for a specific application. As an example, for map applications such as LULC database production, the tool automatically identifies erroneously overlapping land use polygons and analyzes not only the equidistance between the polygons’ adjacent vertices, which is a good way to determine the quality of those polygons, but also the evolution of their equidistance over time.

Methods to analyze the completeness of building footprints over space and time were described and analyzed by Hecht, Kunze and Hahmann (2013) for the German states of North Rhine-Westphalia and Saxony. They used unit-based and object-based methods to analyze the level of completeness of building footprints contained within OSM by always comparing them with a reference dataset regarded as complete. They conclude that unit-based methods require less computation but have limitations in their level of detail when compared with object-based methods. Their results in applying these methods to the mentioned areas of Germany showed that OSM building footprints, as of November 2012, are characterized by a low degree of completeness, below 30%, and a strong geometrical heterogeneity, and the level of completeness is higher in urban than in rural areas.

A similar study for the German city of Munich was developed by Fan, Zipf, Fu and Neis (2014). In this study the authors developed a quality assessment of building foot-print data, after they found that the number of buildings in OSM was over 77 mil-lion on 5 May 2013. Building footprints were assessed using four criteria: 1) completeness, 2) semantic accuracy, 3) position accuracy, and 4) shape accuracy, where OSM data were

compared with the reference data from the German Amtliches Topographisch-kartographisches Informatiosystem – Authorative Topographic-Cartographic Information System (ATKIS) to perform a quantitative assessment. They concluded that, for the case study of Munich (Germany), a high level of completeness was found but OSM building footprints still lack important attributes such as name, type, and height, among others. They found, however, more than 1200 newly constructed buildings which were not documented in the ATKIS data.

On the other side, although OSM building footprints are very similar in terms of shape, they have on average a 4 m offset to their corresponding ones in ATKIS in terms of position accuracy. Building footprints might be an important source of information to help in the classification or validation of urban areas, and these results are a very good indicator. Jokar Arsanjani and Vaz (2015) analyzed the completeness and thematic accuracy of seven European metropolises and thanks to the promising accuracy values concluded that these parameters greatly vary from location to location, which confirms the heterogeneity of contributions.

Building a hybrid land cover map with crowdsourcing and geographically weighted regression was the purpose of a recent study developed by See et al. (2014). The authors used medium resolution land cover products with crowdsourced data from the Geo-Wiki project combined by a geographically weighted regression approach to produce a hybrid global land cover map. They argued that the results serve to demonstrate that medium resolution global land cover information can be improved with existing products using spatial analysis methods.

Fonte, Bastin, See, Foody, & Lupia (2015) studied recently the usability of VGI for validation of land cover maps. They discuss potential and challenges of such type of data for land cover map validation based on a revision of cases where VGI data was used as an additional source of data to assist in map validation and also where only VGI data was used.

2.5. Spatial data integration

The debate about the diverse sources of geographic information we have been discussing in the previous sections drives us to discuss about another inevitable topic: the integration of such different sources and the benefits that might be obtained from their integration by exploiting the merits of each data source (Gösseln & Sester, 2004). Integrated analysis, geometric reference, mutual correction and refinements, semantic and geometric properties enrichment are among the benefits that might be obtained from the combination of different data sources (Butenuth et al., 2007).

However, the integration of data from heterogeneous sources brings up challenges that need to be overcome. Mohammadi, Rajabifard, & Williamson (2010) have identified technical and nontechnical issues related with the effective spatial integration using case studies from countries of the Asia-Pacific region. They propose an open web-based tool for the effective spatial data integration that facilitates data harmonization through the assessment of multisource spatial data sets against a set of defined rules where items of incompliancy are highlighted in a final report.

A 4-layered service-oriented architecture for spatial data integration (SOA-SDI) was proposed by Sha & Xie (2010) to build WebGIS applications. They argue that this 4-

layered SOA-SDI shows more flexibility than the traditional service-oriented architecture (SOA) for building new GIS applications. For demonstration purposes and based on this infrastructure they have developed two WebGIS applications, Safe Route-to-School (SR2S) and Public Facility Management (PFM), based on four categories of data providers services: Google map, WMS services, ArcIMS services and Pictometry image service.

Although nontechnical aspects were identified in the paper of Mohammadi et al. (2010) more related with institutional, policy, legal and social aspects and therefore more connected with authoritative data sources, some technical issues were also acknowledged: inconsistent data specification; multiple raster and vector formats; variety of spatial resolution; different scales; differences in datum, projections, coordinate systems; data models; currency and accuracy; and logical inconsistency. These issues are part of what is called interoperability, a very important concept especially in distributed systems dealing with heterogeneous sources of data. Interoperability is seen as a solution to overcome syntactic, structural, and semantic differences among heterogeneous data sources at both spatial and temporal levels (Bishr, 1997; Brodeur, Bedard, Edwards, & Moulin, 2003; Laurini, 1998).

In the next sections we will describe and discuss the interoperability in 2.5.1 and discuss the integration problem in two dimensions: 1) the communicational dimension, and 2) the compatibility dimension. The first dimension refers to the communication and sharing of data among different sources of information and application interfaces, and is explored in section 2.5.2. The second dimension refers to the compatibility among data from heterogeneous sources and is described in section 2.5.3.

2.5.1. Interoperability

Interoperability refers in general to the ability of a system or systems to communicate and interchange information collaboratively (Bishr, 1998; Kottman, 1999; Vckovski, 1998).

In this particular area, the Open Geospacial Consortium (OGC) plays a very important role in promoting interoperability by developing standards to overcome the challenges related with the exchange of heterogeneous data (Kottman, 1999). The OGC, founded 20 years ago in 1994, is an international industry consortium of 508 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards (OGC, 2014b).

Standards are the main deliverables of the OGC. They are technical documents detailing interfaces or encodings developed to address specific interoperability challenges, and used by the software developers to build open interfaces and encodings into their products and services. According to the OGC standards web page, "*Ideally, when OGC standards are implemented in products or online services by two different software engineers working independently, the resulting components plug and play, that is, they work together without further debugging*" (OGC, 2014a).

The general concept of interoperability might be split into different levels. Mohammadi et al. (2010) acknowledged three levels of interoperability that have been identified by the research community: 1) the syntactical interoperability to overcome the challenge of information reuse; 2) the structural interoperability to help in the conversion among schemas; and 3) the semantic interoperability that deals with the meaning of heterogeneous data from diverse systems.

These concepts have been gaining more and more importance due to a shift that we have been testifying throughout the last decades. The GIS technology has been evolving from mainframe GIS to desktop GIS and more recently to distributed GIS. In fact, most of the todays' systems use the Internet to share data and information which makes this concept of interoperability even more important. Internet GIS refers to a certain type of GIS that uses the Internet as the primary way to exchange data, conduct spatial analysis and disseminate results (Peng & Tsou, 2003). In such systems, interoperability assures the ability of different systems to communicate and exchange information among them. The next section provides an overview of distributed GIS where interoperability is a mandatory concept.

2.5.2. Distributed GIS

Distributed GIS, refers to distributed systems of spatial data based on the standards and software of the Internet (Tait, 2005). Such systems are based on information technology standards, such as the Transmission Control Protocol/Internet Protocol (TCP/IP), Hyper Text Transport Protocol (HTTP), Hyper Text Markup Language (HTML), and eXtensible Markup Language (XML), and other infrastructure related components, such as software, hardware and communications network.

As mentioned above, in the basis of Distributed GIS is Internet GIS, a term that refers to GIS functions and geospatial data sharing over the Internet. The problem of Internet GIS was that most of the Internet GIS applications kept resources and elements centralized as one single application, with their specific logics. This brought up two interconnected major problems of Internet GIS: interoperability and integration (Chang & Park, 2006).

Many Internet GIS applications, given their heterogeneous environments, are not interoperable and therefore cannot be shared.

Distributed GIS applications try to solve these issues by using programmatic interfaces to share resources over the Internet. Such programmatic interfaces are known as Web Services (WS) and provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks (W3C, 2013).

According to Mazzetti, Nativi, & Caron (2009) there are two main architectural approaches to the development of WS: Service-Oriented-Architectures (SOA) and Resource- Oriented Architectures (ROA). In SOA, the central concept is the service, handled by the service provider, which allow the execution of tasks involving resources that, in this case, are not exposed to the user. This approach is powerful but its complexity is one of the main disadvantages. In opposition, the resource is the key in ROA. In this approach, resources are exposed to the user allowing the direct interaction, making the interaction easier.

Web services such as the Simple Object Access Protocol (SOAP) and more recently the Representational State Transfer (REST), the most well-known and widely used web services, were designed to implement respectively SOA and ROA. Due to its simplicity and lightweight, the RESTful approach, the REST web service implementation, is emerging as a popular alternative over SOAP (Sun, 2009).

It is important to mention that the above-mentioned architectures use an approach independent of specific programming languages or operating systems (Fielding, 2000; Papazoglou, Traverso, Dustdar, & Leymann, 2008).

2.5.3. Data harmonization, conflation and fusion

To overcome the identified problems related with the integration of multiple heterogeneous data sources, several techniques and concepts have been developed. Terms such as data harmonization, data conflation and data fusion are widely used by the research community. Despite some differences, they all address the integration of heterogeneous data sources in a common model or platform.

Harmonization can be seen as a general term aimed at minimizing systematic differences between different sources (Bartholomeus, Witte, van Bodegom, & Aerts, 2008; Keune, Murray, & Benking, 1991). In the same way, Herold et al. (2006) frames harmonization in the context of land cover characterization as the "*process whereby the similarities between existing definitions of land cover are emphasized, and inconsistencies reduced*". They argue that harmonization does not necessarily eliminate all differences, but should eliminate major discrepancies, so they become compatible and comparable.

According to Ruiz, Ariza, Ureña and Blázquez (2011), the general term conflation, in the context of heterogeneous sources covering the same geographical area and describing the same reality in different forms, density and accuracy, describes the same procedure as data integration of such heterogeneous sources defined by several other authors (e.g. Butenuth et al., 2007; Olteanu, Mustière, & Ruas, 2006). Cobb et al. (1998) use the term conflation to refer a process similar to what is known as data fusion (Stankut & Asche, 2009), i.e. the integration of two different sources to obtain one new and more richer product.

Data conflation or automated map compilation, coined in the early 1980s by Saalfeld, was first implemented in 1985 by the United States Geological Survey (USGS) and the Bureau of the Census, in a joint project to consolidate the maps of the metropolitan areas of the United States of both entities, a task that justified a major investment given the necessary effort to combine around 5700 pairs of map files (Saalfeld, 1988).

The concept behind data fusion refers to the extraction of the best-fit geometry data and most suitable semantic data and further amalgamation into a new dataset (Stankut & Asche, 2009). Wald (1999), regarding the remote sensing domain, found several different definitions for data fusion and sometimes the same term applied to slightly different concepts. Accordingly, a new definition was proposed stating that "*data fusion is a formal framework in which are expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of 'greater quality' will depend upon the application*".

Summarizing, data conflation refers to processes that identify matching features based on geometrical, topological and semantic attributes and data fusion use those identified features to fuse or combine them in a new and enriched dataset (Wiemann & Bernard, 2010).

2.6. Conceptual Framework

The present study involves the integration of different VGI data sources in order to use them in the process of LULC production. Figure 3 shows the conceptual framework based on the literature review, where the different processes involved and identified are presented. This conceptual diagram gives a more visual insight of this study where one

can easily realize that we intent to bridge the existent gap between different VGI sources and LULC production by proposing a data integration model.

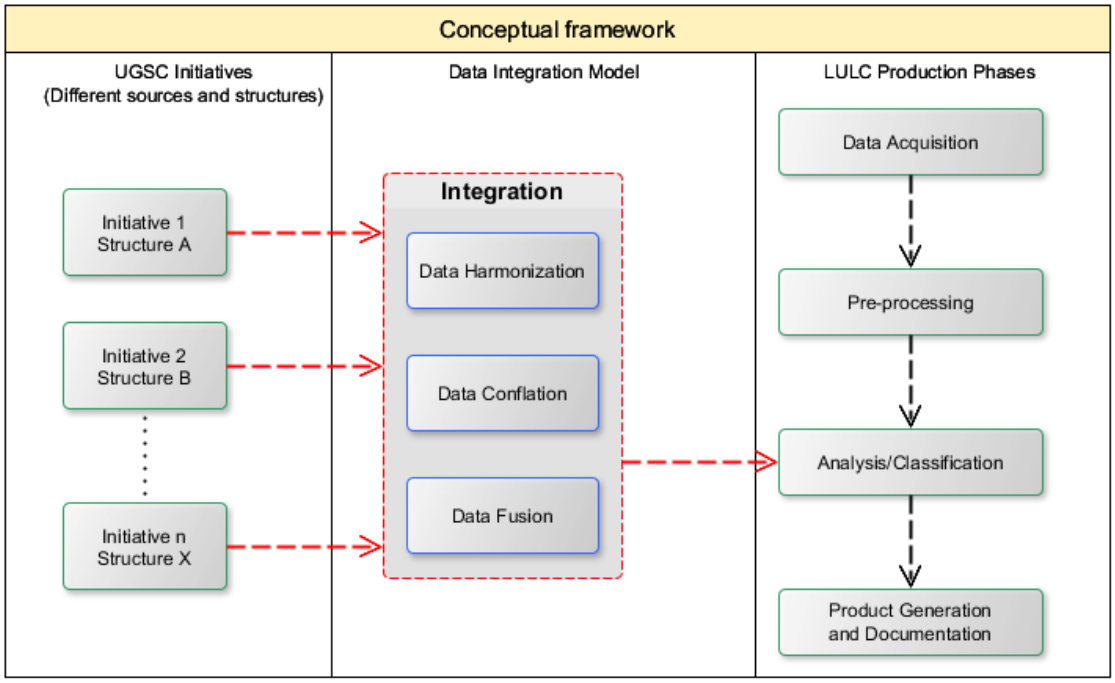

Figure 3 - Conceptual framework

2.7. Discussion and conclusions

In this chapter we discussed concepts and techniques found in the literature in respect to the integration of heterogeneous data sources into a common platform to help in the process of LULC databases production.

We started by debating the new trend of Geographic Information produced and shared by volunteers and explored several VGI initiatives available over the Web to conclude that such an enormous amount of data needs to be exploited to extract meaningful information that helps the society overcoming real world problems.

We then looked at the concept of data integration and all the related questions. On the one side the concepts of interoperability, Distributed GIS, Internet GIS and Web services gave us a broader vision in terms of the communication process to access and retrieve data from different sources. On the other hand, the theories behind data harmonization, data conflation and data fusion showed us the complexity of combining heterogeneous data to visualize and extract meaningful information from the integration of those sources.

3. Feasibility of User Generated spatial Content for Land Use/Land Cover

3.1. Introduction

Prior to the development of the integration data model, it was important to explore the feasibility of UGsC data to be used as a source of information to help in the process of LULC databases production. This would demonstrate the significance of such development.

In this chapter we describe the studies we have developed exploring different UGsC initiatives to investigate their potential to be used in the process of LULC databases

production. Table 2 provides a list of such studies grouped by initiative and type of publication.

Initiative	Type of data	Type of publication	Study area	Reference
OSM	Vector	Book chapter	Continental Portugal	Estima and Painho (2015)
		Conference	Continental Portugal	Estima and Painho (2013a)
Flickr	Photos	Conference	2 Portuguese Municipalities	Estima, Fonte and Painho (2014)
			Continental Portugal	Estima and Painho (2013b)
Flickr + Panoramio	Photos	Journal	Continental Portugal	Estima and Painho (2014)

Table 2 - List of studies undertaken to explore the suitability of UGsC initiatives for the purpose of helping LULC activities

For each study, we applied a methodology that evaluates the respective UGsC source data against a reference LULC database, the CORINE (COoRdination of INformation on the Environment) Land Cover (CLC) database, within a defined study area.

The CLC database used is composed by the version 16 (04/2012) for the CLC2006 inventory, downloaded from the European Environment Agency (EEA). This dataset was developed using the European Terrestrial Reference System 1989 (ETRS89) with the Lambert Azimuthal Equal Area, also known as ETRS89-LAEA. Using a Minimum Mapping Unit (MMU) of 25 Ha, the land cover is classified according to the CLC nomenclature, shown in Table 3, which is hierarchically divided into three levels of classes: Level 1, 2 and 3, with the granularity increasing from the former towards the latter.

Level 1	Area (Ha)	Level 2	Area (Ha)	Level 3	Area (Ha)
1 Artificial surfaces	309716.89	11 Urban fabric	227482.56	111 Continuous urban fabric	12234.34
				112 Discontinuous urban fabric	215248.23
		12 Industrial, commercial and transport units	47821.49	121 Industrial or commercial units	33895.51
				122 Road and rail networks and associated land	7678.06
				123 Port areas	1945.27
				124 Airports	4302.65
		13 Mine, dump and construction sites	21149.09	131 Mineral extraction sites	13659.71
				132 Dump sites	971.58
				133 Construction sites	6517.80
		14 Artificial, non-agricultural vegetated areas	13263.75	141 Green urban areas	1763.71
				142 Sport and leisure facilities	11500.04
2 Agricultural areas	4199177.27	21 Arable land	1245009.51	211 Non-irrigated arable land	981677.22
				212 Permanently irrigated land	210509.59
				213 Rice fields	52822.70
		22 Permanent crops	592974.48	221 Vineyards	228965.31
				222 Fruit trees and berry plantations	100983.22
				223 Olive groves	263025.95
		23 Pastures	41871.11	231 Pastures	41871.11
		24 Heterogeneous agricultural areas	2319322.18	241 Annual crops associated with permanent crops	404000.98
				242 Complex cultivation patterns	607041.55
				243 Land principally occupied by agriculture	686819.25
				244 Agro-forestry areas	621460.40
3 Forest and semi natural areas	4259642.22	31 Forests	2016515.84	311 Broad-leaved forest	1007003.84
				312 Coniferous forest	533981.79
				313 Mixed forest	475530.21
		32 Scrub and/or herbaceous vegetation associations	2074423.48	321 Natural grasslands	171861.61
				322 Moors and heathland	284552.04
				323 Sclerophyllous vegetation	206613.41
				324 Transitional woodland-shrub	1411396.42
		33 Open spaces with little or no vegetation	168702.90	331 Beaches, dunes, sands	11148.98
				332 Bare rocks	23862.88
				333 Sparsely vegetated areas	100830.47
				334 Burnt areas	32860.57
				335 Glaciers and perpetual snow	0.00
4 Wetlands	28777.11	41 Inland wetlands	1138.71	411 Inland marshes	1138.71
				412 Peat bogs	0.00
		42 Maritime wetlands	27638.40	421 Salt marshes	18457.26
				422 Salines	7228.50
				423 Intertidal flats	1952.64
5 Water bodies	110906.66	51 Inland waters	72859.65	511 Water courses	19874.09
				512 Water bodies	52985.56
		52 Marine waters	38047.01	521 Coastal lagoons	8521.46
				522 Estuaries	26680.68
				523 Sea and ocean	2844.87

Table 3 - CLC nomenclature and respective areas for continental Portugal[9]

[9] Source: http://www.igeo.pt/gdr/pdf/CLC2006_nomenclature_addendum.pdf

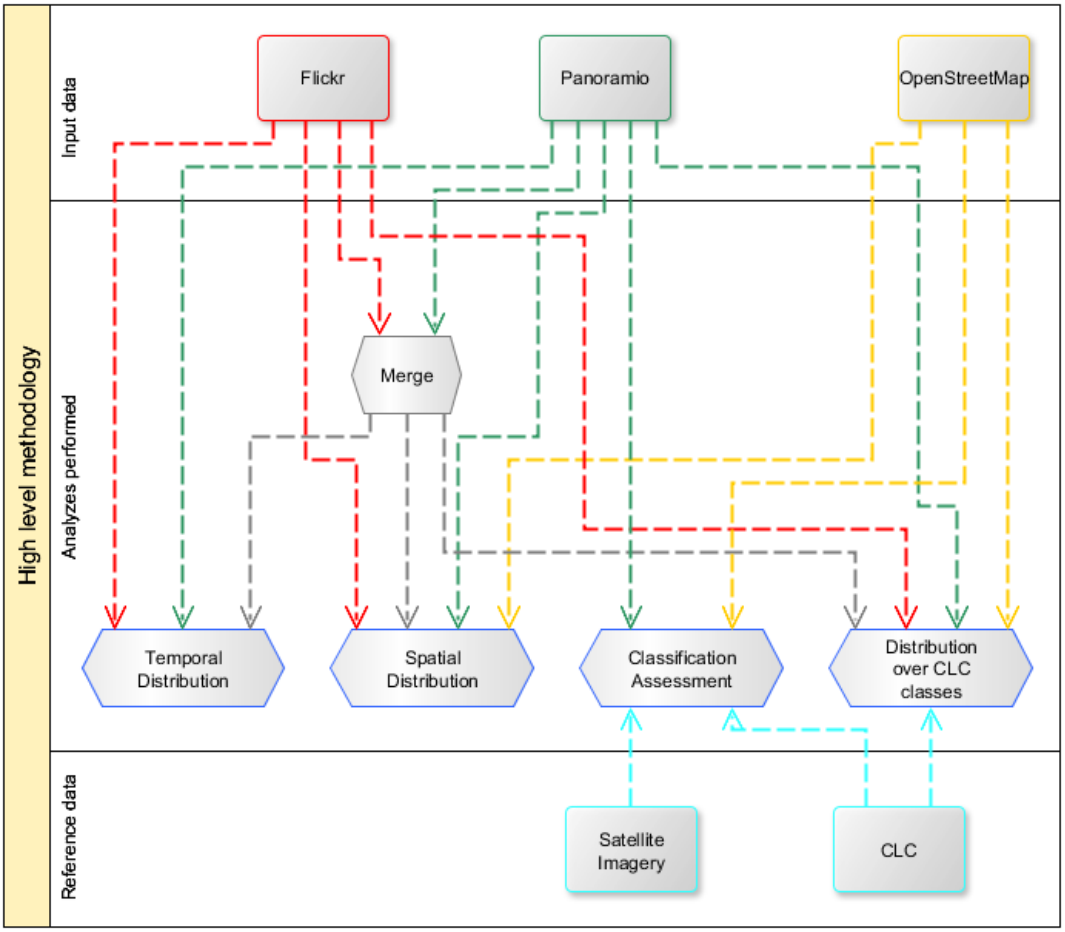

Figure 4 - High-level global methodology

Figure 4 provides a high level view of the global methodology applied in these studies. Four main analyzes were developed depending on the UGsC initiative being investigated: 1) analysis of the temporal distribution; 2) analysis of the spatial distribution; 3) Assessment of the classification; and 4) analysis of the distribution over CLC classes.

In the following sections we provide a detailed description of each study, including the methodology, results and discussions. Therefore, this chapter is divided into two main sections: 1) the OpenStreetMap initiative and 2) photo based initiatives. Regarding the OpenStreetMap initiative we present two studies where we explored respectively polygon features and PoI's (Points of Interest). Concerning the photo based initiatives we present three studies involving the Flickr initiative, the Panoramio initiative and a study where we merged data from both initiatives. We conclude by offering a discussion on the feasibility of UGsC data as a source of information for LULC activities, followed by some final remarks.

3.2. The OpenStreetMap initiative

As already mentioned, this initiative is one of the best known and most studied VGI initiatives (Elwood et al., 2012). To explore the suitability of OSM for the purpose of using it to help in the LULC databases production process, we downloaded the OSM database from the Geofabrik website[10]. This database is current as of July 23, 2013, and is divided in six datasets: places, points, railways, roads, waterways, buildings, landuse and natural areas. Places and points are represented by points; railways, roads and waterways by lines; and buildings, landuse and natural areas by polygons. We have analyzed different datasets in two separate studies:

1. Exploratory analysis of OpenStreetMap for land use classification;

[10] http://www.geofabrik.de/data/download.html

2. Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal.

3.2.1. Exploratory analysis of OpenStreetMap for land use classification

In this study we conducted an exploratory analysis of data from the OpenStreetMap initiative. Using the CLC database as reference and continental Portugal as the study area, we established a possible correspondence between both classification nomenclatures, evaluated the quality of OpenStreetMap polygon features classification against CLC level 1 classes, and analyzed the spatial distribution of OpenStreetMap classes over continental Portugal.

3.2.1.1. Study area and datasets used

The defined study site is Continental Portugal, located in the southwestern side of Europe, which is constituted with 18 districts and 278 municipalities covering a total area of 8908220.16 Ha. The land cover is mainly composed by agricultural and forest areas covering around 95% of the country.

The OSM database under analysis covers the area of continental Portugal and was downloaded from the Geofabrik website.

The nomenclature used to classify features in the OSM datasets is available in the OSM wiki Website (OpenStreetMap, 2014), along with pictures and descriptions for each class. Table 4 shows the OSM nomenclature classes identified over continental Portugal for natural areas and landuse classes. Regarding the buildings dataset, as

the majority of the features do not have a class defined, it was decided to assign a generic class “urban” to all of them. It is important to refer that the generalization we are doing for this specific case might have a negative impact, mainly in rural areas and this should be further investigated in the future.

“Landuse” classes				
Abutters	Farm	Harbour	Park	Scrubs
Allotments	Farmland	Industrial	Public	University
Basin	Farmyard	Landfill	Quarry	Village_green
Beach	Field	Leisure	Railway	Vineyard
Brownfield	Garages	Meadow	Reservoir	Waste_water_plan
Cemetery	Garden	Military	Residential	Water
Commercial	Grass	Museum	Retail	Wood
Conservation	Greenfield	Not_known	Salt_pond	Greenhouse_horti
Construction	Greenhouse	Orchard	Scrub	Recreation_groun
“Natural areas” classes				
Forest	Park	Riverbank	Water	

Table 4 - OSM datasets' classes over continental Portugal

3.2.1.2. Assumptions

For the correct understanding of this study, it is important to refer that we assume that the time difference between CLC and OSM databases (2006 for CLC and 2013 for OSM) would not represent a major issue. Considering a yearly average change value of land cover in Europe of 0.23% (Büttner, Kosztra, Maucha, & Pataki, 2012), for the purpose of this exploratory analysis, we believe that the impact of such change rate between both periods does not depreciate this study. In a more in depth analysis, data from similar periods shall be used.

3.2.1.3. Methods

The adopted methodology to conduct this exploratory analysis was, according to Figure 3, as follows:

1. Analysis of the defined OSM datasets. We have explored the three polygon based OSM datasets defined in the previous section in terms of nomenclature and area of coverage. We have also analyzed the areas of overlap to identify eventual existing inconsistencies;
2. Analysis and establishment of a relationship between the classification nomenclatures used by the different databases (CLC and OSM). In this step we established a correspondence between CLC and OSM classes defined by their respective nomenclatures, extremely important to develop the subsequent steps in this methodology;
3. Analysis of the coverage of each OSM class using the CLC level 1 classes as reference: 1) artificial surfaces, 2) agricultural areas, 3) forests and semi-natural areas, 4) wetlands, 5) water bodies. According to the relationship between OSM and CLC established in the previous step, we first merged all the OSM datasets and gave each OSM class the corresponding CLC level 1 class. We have then dissolved all the polygons by each CLC class value to have a resultant map with only 5 classes plus the areas without corresponding CLC class. In the last step we have removed overlapping areas in conflict. Then a comparison between the resultant areas and the correspondent ones from the CLC database was made;

4. Analysis of the matching degree between related classes. In this step, the area covered by each class that matched the correspondent CLC level 1 class was determined by intersecting both datasets, and the accuracy of OSM classification calculated;
5. Analysis of the OSM spatial distribution. In this final step, we intersected the dataset resultant from the previous step with a dataset representing the Portuguese districts, an administrative division that splits the country in 18 areas.

It is important to refer that in steps 3, 4 and 5 the developed analyses were restricted to the level 1 of the CLC. This was due to multiple correspondence issues detected in the step 2. Solutions to solve this multiple correspondences need further investigation that is outside the scope of this study.

3.2.1.4. Analysis of OSM datasets

The first step was to explore the three datasets in terms of nomenclature, area of coverage and overlapping areas to identify eventual existing inconsistencies. Table 5 describes the areas of coverage of each dataset in Ha and the respective percentage relative to continental Portugal.

Dataset	Area in Ha	Country coverage (%)
Natural areas	140006.95	1.57%
Landuse	144350.23	1.62%
Buildings	7057.61	0.08%

Table 5 - Areas of coverage of the used OSM datasets

Landuse is the dataset with more extensive coverage, covering 1.62%, followed by the natural areas dataset covering 1.57% and the buildings dataset covering 0.08% of the country. These three datasets together cover a total of 3.27% of the study area. In order to have a more realistic value, once some of the features represented in these datasets are totally or partially superimposed, the overlapping areas were deducted. The determined overlapping area was approximately 3017.18 Ha representing 0.03%, making the real coverage area to decrease by 3.24%.

Before deducting the overlapping areas, the three OSM datasets were also intersected to identify existing classification inconsistencies in those areas. Table 6 summarizes the different classifications recognized in those common areas. These different classifications do not represent a real conflict but rather the combination of different features/classes in the same location, seen probably by their contributors at different scales. A good example of that, extracted from Table 6, would be a place classified as park in natural areas, residential in landuse and café, church or museum, etc. in buildings. This example represents actually something that happens in reality with these datasets.

The total value of overlapping areas with different classification shown in Table 6, 9.47 Ha, is significantly lower that the total area of overlapping areas show above, 3017.18 Ha, which gives us a good indicator that the classification has some consistency.

Natural areas dataset	Landuse dataset	Buildings dataset	Area (Ha)
Forest	Military	None	5.24
	Residential	Reservoir_cover	0.02
	Recreation_ground	Hospital	0.25
Park	Commercial	None	0.01
	Residential	Museum	0.39
		Cafe	0.05
		Chapel	0.01
		Church	0.00
		House	0.03
		Library	0.03
		Museum	0.08
		Public	0.02
		Public_building	0.37
		Restaurant	0.03
		Roof	0.01
		Theatre	0.03
		Toilets	0.00
		Yes	0.01

Table 6 - Existing classification differences within the three OSM datasets

3.2.1.5. Correspondence between OSM and CLC nomenclatures

Each database (CLC and OSM) uses different nomenclatures for classification. It is therefore necessary to find correspondence between both systems before proceeding to the next steps. Although the OSM wiki page already has a possible correspondence, some of the tags present in the study area are not mentioned there. Thus, in Table 7 we propose a tentative to relate both CLC and OSM nomenclatures, developed based on the description of each CLC and OSM class available at the OSM wiki Website mentioned before and the CLC illustrator guide, respectively.

OSM classes	CLC classes		
	Level 3	Level 2	Level 1
Landuse dataset			
Abutters	111-112-121	11-12	1
Allotments	242	24	2
Basin	512	51	5
Beach	331	33	3
Brownfield	133	13	1
Cemetery	111-112	11	1
Commercial	121	12	1
Conservation	313-312-311	31	3
Construction	133	13	1
Farm	222-231-241-242	22-23-24	2
Farmland	222-231-241-242	22-23-24	2
Farmyard	222-231-241-242	22-23-24	2
Field	**?**	**?**	**?**
Garages	122	12	1
Garden	142	14	1
Grass	231-321	23-32	2-3
Greenfield	321-322-323-324	32	3
Greenhouse	211	21	2
Greenhouse_horti	211	21	2
Harbour	123	12	1
Industrial	121	12	1
Landfill	132	13	1
Leisure	142	14	1
Meadow	231	23	2
Military	**?**	**?**	**?**
Museum	121	12	1
Not_known	**?**	**?**	**?**
Orchard	222-241	22-24	2
Park	142	14	1
Public	121	12	1
Quarry	131	13	1
Railway	122	12	1
Recreation_groun	142	14	1
Reservoir	512	51	5
Residential	111-112	11	1
Retail	121	12	1
Salt_pond	422	42	4
Scrub	324-323-322-321	32	3
Scrubs	324-323-322-321	32	3
University	121	12	1
Village_green	141	14	1
Vineyard	221	22	2
Waste_water_plan	121	12	1
Water	511-512	51	5
Wood	313-312-311	31	3
Natural areas dataset			
forest	313/312/311	31	3
park	313/312/311	31	3
riverbank	512/511	51	5
water	523/522/511/512/511	52/51	5

Table 7 - Correspondence between CLC and OSM classes

(CLC classes according to the CLC nomenclature presented in Table 3)

Difficulties arouse trying to establish a direct relation between some classes from the two nomenclatures. In this sense, three types of issues occurred: 1) two OSM classes were not identified at all due to absence of any description (case of OSM

classes “field” and “not_known”) in the OSM wiki; 2) one OSM class didn’t match with the description of any CLC (the “military” class); and 3) some OSM classes did not fit in the description of only one CLC class resulting in multiple correspondences. In the first and second cases, a unique correspondence was not possible to provide.

It is noticeable that the difficulty in finding correspondence rises when the level of detail increases, e.g. more multiple correspondences can be verified in the level 3 than in the level 1. Actually, for the level 1 only one case of multiple correspondence was identified: the “grass” class. In the description of this class it is stated that it should be used to represent “areas covered with grass” and, as a complement, it is also specified that the user should “consider landuse=meadow for meadow and landuse=pasture for pasture”. According to the description of CLC level 1 classes, two CLC classes can match this OSM class: agricultural and forest and semi natural areas making it a multiple correspondence case.

The following steps in the analysis used the level 1 classes of CLC database assuming that the OSM “grass” class has only one correspondent CLC level 1 class that is the class 3, forest and semi-natural areas.

3.2.1.6. Coverage analysis of OSM datasets

In the next step we used the OSM merged dataset from the previous step and gave to each feature the corresponding CLC level 1 class. Then we dissolved the resultant dataset by CLC level 1 class and removed overlapping areas in conflict, e.g. all the overlapping areas with a different CLC level 1 class were removed. These areas

perform a total of 4004.05 Ha representing 1.39% of the OSM area. It is important to refer that these areas were not deducted but totally removed from the analysis. We then calculate the coverage area of each new class group and compare them with those from CLC database.

Table 8 shows the results of this analysis. For each class we have the corresponding area from the CLC database in the second column and the area from OSM database in the third column. The fourth and fifth columns shows, the percentage covered by each OSM class over each respective CLC class and over continental Portugal, respectively.

CLC classes	Area from CLC (Ha)	Area from OSM (Ha)	Class coverage (%)
Unclassified	---	7036.75	---
1 Artificial Surfaces	309716.89	62407.48	20.15
2 Agricultural Areas	4199177.27	34309.93	0.82
3 Forest and Semi Natural Areas	4259642.22	98536.62	2.31
4 Wetlands	28777.11	64.59	0.22
5 Water Bodies	110906.66	82621.61	74.50

Table 8 - Coverage areas from CLC level 1 and OSM

Some interesting indicators can be seen in Table 8. Comparing the coverage area, by class, between OSM and CLC, Water Bodies had a very interesting value of 74.5% followed by Artificial Surfaces covering 20.15%. Agricultural Areas, Forest and Semi Natural Areas and Wetlands have poor coverage with values under 10%. The “unclassified” areas, OSM classes without correspondent CLC level 1 class, represent a total of 7036.75 Ha that, comparing with the other values displayed in Table 5, covers 0.08 % of the country.

3.2.1.7. Analysis of OSM classification accuracy

In the following step, the verification of classifications in overlapping areas was made. We based this analysis using a confusion matrix shown in Table 9. Values in shaded cells represent areas with the same classification in both databases.

		OSM classes					
		Artificial Surfaces	**Agricultural Areas**	**Forest and Semi Natural Areas**	**Wetlands**	**Water Bodies**	**Total**
CLC classes	**1 Artificial Surfaces**	**44160.56**	1059.00	4086.69	0.00	663.20	52369.87
	2 Agricultural Areas	12934.72	**31884.28**	10716.09	4.94	12088.20	68459.87
	3 Forest and Semi Natural Areas	5182.27	1214.07	**83362.66**	0.07	6322.15	99843.05
	4 Wetlands	42.27	114.77	238.65	**59.57**	4402.91	4870.53
	5 Water Bodies	87.66	37.81	132.53	0.00	**59145.14**	59433.67
	Total	62407.48	34309.93	98536.62	64.59	82621.61	284976.99

Table 9 - Confusion matrix of CLC vs. OSM classifications

Some calculations can be derived from Table 9 to have an idea about the classification provided by OSM comparing with the one obtained using CLC.

The accuracy index for each CLC class is an important indicator that shows which are the classes where the areas wrongly classified are higher. It is calculated dividing the area correctly classified in each OSM class (diagonal cell in the table) by the total area of each CLC class (sum of each line).

$$Class\ Accuracy = \frac{e_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij}}$$

(where: e represents the value, i the line index and j the column index)

The Global Accuracy (GA) represents the proportion of area where the classification matches in both databases over the total overlapping area, given by the formula:

$$Global\ Accuracy = \frac{\sum_{i=1}^{n} e_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij}}$$

(where: e represents the value, i the line index and j the column index)

Table 10 shows the resultant values for the accuracy of each class and the global accuracy. Wetlands obtained the worse result, around 1.2% followed by Agricultural Areas with an interesting value of 46.6%. All the other classes had very encouraging results with Water Bodies getting an impressive accuracy value of 99.5%. The GA value is also very interesting and promising around 76.7%.

Class	Classification accuracy (%)
1 Artificial Surfaces	84.3%
2 Agricultural Areas	46.6%
3 Forest and Semi Natural Areas	83.5%
4 Wetlands	1.2%
5 Water Bodies	99.5%
Global	76.7%

Table 10 - Classification accuracy

3.2.1.8. Analysis of the OSM spatial distribution

In the final step the spatial distribution of OSM areas were analyzed, using the dataset resultant from the previous step. Figure 5 shows the spatial distribution of all OSM classified areas and the distribution of classes' coverage areas by continental Portuguese districts, left and right maps respectively. Both maps demonstrate a larger and more balanced coverage near the biggest cities and touristic places. On the opposite side, the interior area of Portugal shows less coverage and homogeneity among the different classes. Also, in Évora and Beja districts, most of

the coverage is related to Water Bodies due to the existence of important dams, such as the Alqueva dam.

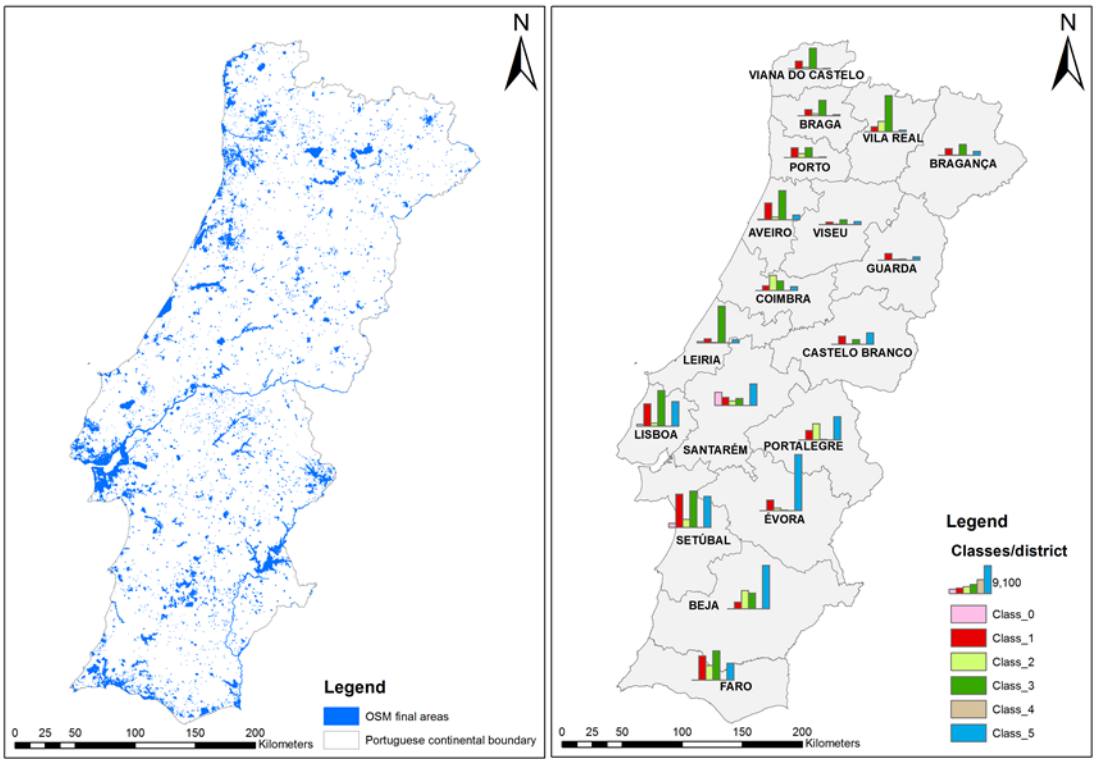

Figure 5 - Spatial distribution of OSM classified areas over continental Portugal (left) and Distribution of classes' coverage areas by continental Portuguese districts (right)

3.2.2. Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal

In this study we explored the PoI's dataset in terms of content and coverage, we established a relationship between each point type and the CLC classes, based on their description documented on the OSM Map Features website (OpenStreetMap, 2014), and, for each point location, we compared the classification given in the

previous step with the respective class extracted from the CLC database, using a confusion matrix approach. We also analyzed the classification accuracy for each OSM point type.

3.2.2.1. Study area and data

Continental Portugal was the defined study site, already described in section 3.2.1.1, and the datasets used are composed by the CLC database already described in section 3.1 and the PoI dataset of the OSM database.

Regarding the CLC database and for the purpose of this study, we used the five classes of level one: 1) artificial surfaces, 2) agricultural areas, 3) forests and semi-natural areas, 4) wetlands, 5) water bodies.

3.2.2.2. Methods

The methodology adopted to conduct this analysis was as follows:

1. We explored the point dataset defined in the previous section in terms of content and coverage;
2. We established a relationship between each point type and the CLC classes, based on their description documented on the OSM Map Features website (OpenStreetMap Map Features 2014);
3. For each point location, we compared the classification given in the previous step with the respective class extracted from the CLC database, using a con-

fusion matrix approach. We also analyzed the classification accuracy for each OSM point type.

3.2.2.3. Analysis of the OSM dataset

In the first step we explored the point dataset in terms of content and cover-age. This data are composed of a collection of 49,861 Points of Interest (PoI) within the study area, classified according to type of PoI. A list of predefined types is available for use when a new point is registered (OpenStreetMap, 2014), but each user can also define new types. Although this possibility gives a lot of flexibility in the mapping and classification process, it creates additional difficulties to perform further analysis, mainly related to the lack of proper descriptions but also to the possibility of introducing spelling errors.

Table 11 shows a list of PoI types found within the collection of points. A closer look shows some types that are not of interest for the purpose of our study, mainly because they do not represent any type of LULC or related feature, or the relation is not clear (e.g. “attraction”, “heritage”, “no”, “yes”). Different spelling for the same type were also found (e.g. “community_centre”, “comunity centre”, and “Comunity_centre”), a typical error related to the possibility of users creating their own types.

Taking into account the description available for each feature type, and only for those types available in the wiki list, the types marked with asterisk (*) in Table 11 were considered attributable to a CLC class and selected for further analysis. This

represents a total of 26,290, corresponding to around 52% of the total number of initial points.

arts_centre*	charging station*	flagpole	marketplace*	reservoir*	tertiary
adit	charging_station*	food_court*	mast	reservoir_covere*	tertiary_link
alpine_hut	chimney	footway	measurement_stat	residential*	theatre*
animal_shelter	cinema*	ford*	megalith	resort	theme_park*
antenna	city_gate*	forester's lodge	memorial	rest_area*	toilets
archaeological_s	clinic*	fort	milestone	restaurant*	tower
artwork	clinica fisiote	forte de sao jo	mineshaft	road*	townhall*
ashtray	clock	fountain*	mini_roundabout*	ruins	track
atm*	college*	fuel*	moinho do cuco	satellite_centre	traffic_signals
attraction	communications_t	gasometer*	monument*	school*	traffic-signs
baby_hatch	community_centre*	gate	motel*	scout_hut	trail_riding_sta
bank*	comunity centre*	give_way	motorcycle_parki	secondary	tram_stop*
bar*	comunity_centre*	grave_yard	motorway_junctio*	seguranca socia	trunk_junction
battlefield	conference_centr	guest_house	museum*	service	turning_circle*
bbq	construction	halt	newspaper*	services*	turntable*
beacon	convent	health	newstand	shelter	undefined
beauty	courthouse*	health_centre	nightclub*	shop*	university*
bed & breakfast	coutada	healthcare	no	shower*	user defined
bench	crane*	heritage	nursing_home*	silo*	vending_machine
biblias e casa	critpy	horses	oil_tank	snack_bar	veterinary*
bicycle_parking	cross	hospital*	old_cafe	social_centre*	viewpoint
bicycle_rental	crossing	hostel*	optical	social_facility*	waste_basket
biergarten	dentist*	hotel*	park*	solicitor	waste_deposal
boundary_stone	disused	hunting_stand	parking*	souvenirshop	waste_disposal
bridge*	diving_center	ice_cream*	parking_entrance*	spa	waste_dispostal
brothel	doctor*	icon	parking_space*	spa	wastewater_plant*
buffer_stop	doctors*	incline	passing_place	speed_camera*	water_tank
buoy	drinking_water	incline_steep	path	sport clube leir	water_tower*
buoy	driving school	incline_up	pharmacy*	station*	water_well
bus_station*	driving_school	info	picnic_site	steps	water_works*
bus_stop	elevator	information	pier*	stop	waterfall
café*	embassy*	junction	pillar buoy	storage_tank	watering_place
cairn	emergency_access	kindergarten*	place_of_worship*	street_lamp	watermill
caixa geral de d	emergency_phone	laboratory	police*	studio*	wayside_cross
camp_site	escola superior	landmark	post_box	subway entrance*	wayside_shrine
camping park	ev_charging*	lavoir	post_office*	subway_entrance*	wifi
capela	farmacia	lawyer*	posto abastecime	survey point	wind_turbine
car_rental*	fast_food*	leisure_centre	primary_link	survey_pillar	windmill
car_wash*	ferry_terminal	level_crossing*	prison*	survey_point	works*
chalet	fitness_center	lookout_tower	register_office*	telephone	
caravan_site	fire_hydrant*	library*	pub*	swimming_pool*	yes
castle*	fire_station*	lift	public_building*	taxi	zoo*
cemiterio	first_aid	lighthouse	recycling	teahouse	

Table 11 - List of types of OSM PoIs

Legend: types marked with asterisk (*) were considered attributable to a CLC class and selected for further analysis

Figure 1 shows the spatial distribution of the selected PoIs over the study area. It is possible to observe the concentration of points over the coast, where touristic places and larger cities are represented, as well as along some of the main roads.

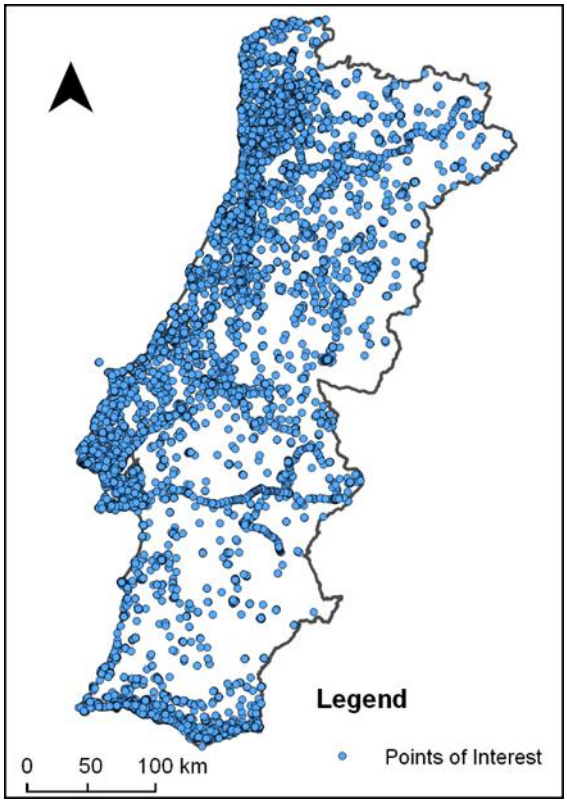

Figure 6 - Spatial distribution of the points of interest over the study area

3.2.2.4. Correspondence between OSM point types and CLC classes

After selecting the types of PoI to use in the previous task, a CLC equivalent class was attributed to each type according to their description in the wiki website. Only two CLC classes were used: classes 1 and 5, representing Artificial Surfaces and Water Bodies, respectively. This was already expected due to the higher probability of more volunteers visiting places fitting in these classes. There were some special cases where we also took into account our knowledge of the feature type class versus their surroundings. The case of the “bridge” feature type, which would apparently be classified as Artificial Surfaces, was classified as Water Bodies since

bridges are usually over water bodies and are not represented in LULC databases due to their size. Table 12 shows the list of PoI types for each given CLC level 1 class.

Artificial Surfaces class					Water Bodies Class
arts_centre	crane	lawyer	post_office	subway entrance	bridge
atm	dentist	level_crossing	prison	subway_entrance	ford
bank	doctor	library	pub	swimming_pool	pier
bar	doctors	marketplace	public_building	theatre	reservoir
beauty	embassy	mini_roundabout	register_office	theme_park	
bus_station	ev_charging	monument	residential	townhall	
cafe	fast_food	motel	rest_area	tram_stop	
car_rental	fire_hydrant	motorway_junctio	restaurant	turning_circle	
car_wash	fire_station	museum	road	turntable	
castle	food_court	newspaper	school	university	
charging station	fort	nightclub	services	veterinary	
charging_station	fountain	nursing_home	shop	wastewater_plant	
cinema	fuel	park	shower	water_tower	
city_gate	gasometer	parking	silo	water_works	
clinic	hospital	parking_entrance	social_centre	works	
college	hostel	parking_space	social_facility	zoo	
community_centre	hotel	pharmacy	speed_camera		
comunity centre	ice_cream	place_of_worship	station		
courthouse	kindergarten	police	studio		

Table 12 - CLC classes given to each PoI type

3.2.2.5. Classification accuracy analysis

After assigning a CLC level 1 class to each PoI type, the evaluation of the classification was the next step. In this task we first filled the PoI dataset with the CLC class, based on the correspondence defined in the previous step. We then intersected it with the CLC database to have, for each point location, the classification defined by the PoI description and the classification taken from the CLC database. A new attribute was created to identify agreements/disagreements between the two classifications. This agreement/disagreement is depicted, along with

their spatial distribution, in Figure 7. red points represent locations where both classifications are not matching and green points represent locations where both classifications are equal.

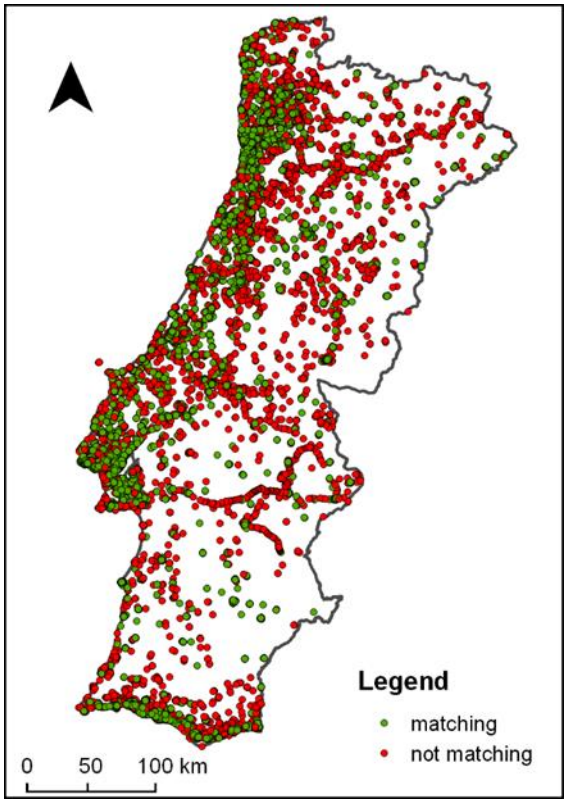

Figure 7 - PoI type class vs. CLC class

Table 13 summarizes the classification of the OSM point accuracy. Points classified as Artificial Surfaces and Water Bodies classes obtained 77.96% and 1.47% correct classification, respectively, when compared with the CLC classification for the same locations. One of the reasons for the poor result of the Water Bodies class might be related with the MMU of 25 Ha of the CLC database. It is natural that body areas of small dimension do not represent the predominant class when using such a MMU value.

		Classification based on OSM points	
		1 Artificial Surfaces	5 Water Bodies
CLC classes containing the point locations	1 Artificial Surfaces	20,421	1
	2 Agricultural Areas	4,110	46
	3 Forest and Semi Natural Areas	1,556	20
	4 Wetlands	19	0
	5 Water Bodies	85	1
	Total	26,191	68
	Correct	77.96%	1.47%
	Wrong	22.03%	98.53%

Table 13 - Classification of OSM points

Finally we analyzed the classification accuracy for each OSM point type. In Table 14, each PoI type is classified according to its range of accuracy. This is important to understand the suitability of each OSM PoI type to use in LULC data-bases. The lower accuracy of some OSM point type might be also related with the MMU. A “rest_area”, for instance, might be located within a forest crossed by a motor way. In the same way, a “water_tower” might be located within an area where another class is predominant.

3.3. Photo based initiatives

In this section we describe the studies we developed using the Flickr and Panoramio photo based initiatives, to explore their suitability for the purpose of using it to help in the LULC databases production process. We refer first to the Flickr initiative separately, presenting two studies, followed by an extended study where a comparison between the Flickr and Panoramio initiatives was performed.

Accuracy classes (%)						
0–50	**50–60**	**60–70**	**70–80**	**80–90**	**90–100**	**100**
water_tower	place_of_worship	works	clinic	fire_station	cinema	charging station
castle	social_facility	station	townhall	parking_space	bank	charging_station
rest_area	speed_camera	motel	hotel	car_wash	courthouse	comunity centre
motorway_junctio	water_works	city_gate	museum	hospital	university	doctor
zoo	silo	food_court	parking	bus_station	pharmacy	embassy
level_crossing	monument		mini_roundabout	nightclub	veterinary	ev_charging
pier	fire_hydrant		swimming_pool	arts_centre	theatre	fort
theme_park	residential		turning_circle	kindergarten	police	ice_cream
gasometer	studio		nursing_home	crane	car_rental	lawyer
services			fuel	public_building	post_office	newspaper
wastewater_plant			fountain	cafe	library	park
beauty			hostel	pub	dentist	parking_entrance
bridge				fast_food	doctors	prison
ford				school	marketplace	register_office
reservoir				restaurant	atm	road
shower				tram_stop	college	shop
turntable				bar		social_centre
				community_centre		subway entrance
						subway_entrance

Table 14 - Classification accuracy by PoI type

3.3.1. Flickr geotagged and publicly available photos: preliminary study of its adequacy for helping quality control of Corine Land Cover

In this paper we conducted a preliminary analysis of the adequacy of photos from the Flickr initiative in order to use them as a source of field data in the quality control of the Land Use/Cover databases production. We evaluated its temporal and spatial distributions over Continental Portugal and also its distribution over Land Use/Cover classes using as a reference the European CORINE Land Cover database. We conclude that this source is very valuable but needs to be combined with other sources due to some issues related with its uneven spatial distribution

3.3.1.1. Description of the study area and datasets

The defined study site is Continental Portugal, and the previously described CLC database was used to support our analysis.

The considered dataset is also composed by the geo-referenced Flickr photos' locations for the study period ranging between 2004 and 2012. It is originally in text format and each location is complemented by the following attributes: latitude, longitude, name, title and date of acquisition. The latitude and longitude values refer to the WGS84 Spatial Reference System (SRS) used by default in GPS receivers. These data were downloaded from the Flickr database using its own API. Initially, we downloaded all the publicly available locations (414,323) inside the Portuguese boundary. Based on the date of acquisition attribute, the old locations (photos taken before 2004) and the locations with missing information were removed. Therefore, the final dataset is constituted by 409,829 locations concentrated mainly over the main cities (Lisboa and Porto) and along the country coastal lines (Figure 8b).

The Portuguese official administrative boundaries database, with the original name of "Carta Administrativa Oficial de Portugal" (CAOP), downloaded from the Portuguese Geographic Institute website, was used to confront with the Flickr photos' locations and characterize them in terms of its spatial distribution over the country. Figure 8a shows the Portuguese country divided by its municipalities.

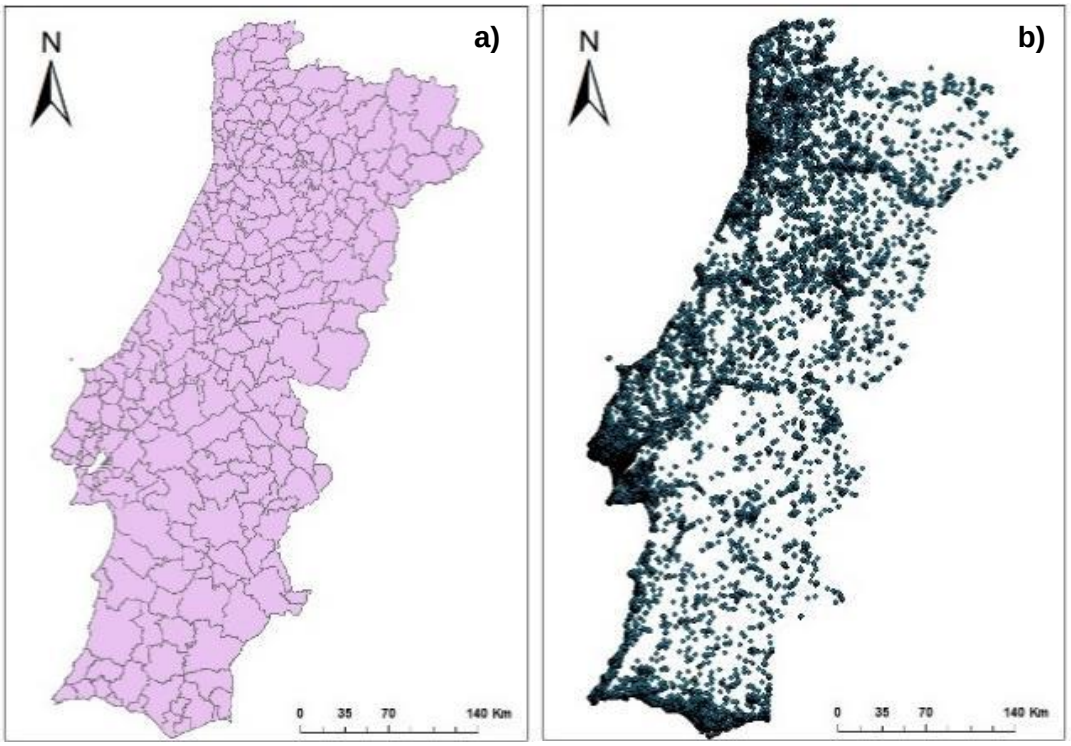

Figure 8 - a) Portuguese boundaries and b) distribution

3.3.1.2. Methods

To explore the suitability of Flickr data for the purpose of using it to help in the LULC databases production process, four main analyses were performed:

1. Analysis of the temporal distribution of photos considering the “date” tag. We examined distribution of the photos over the years to understand the evolution of the initiative, and over the months to understand the monthly distribution;

2. Analysis of the spatial distribution by confronting the photos with the Portuguese municipalities, verifying and comparing the number of photos between different municipalities;
3. Analysis of the distribution over the different CLC classes, prepared by overlaying the points with the CLC database. Each point was assigned the correspondent CLC value and the number of points for each CLC class was calculated. The CLC classes used are shown in Table 3;
4. Cross analysis to compare the distribution of photos over CLC classes along with the spatial and temporal distributions. In this case Portuguese districts were used, as administrative boundaries, for the spatial comparisons.

3.3.1.3. Temporal distribution of Flickr photos

Regarding the temporal distribution of Flickr photos we developed in this first step, we can verify, by observing the chart in Figure 9 (left) that the number of pictures has been growing since 2004. The number of uploaded photos has grown from 3469 in 2004 to 85310 in 2012 at an yearly average rate of around 61%. The highest growth happened from 2005 to 2006, the second and third years of the Flickr initiative, with 240% more photos in the later. This represents an enormous growth possibly explained by the early success of the project and the growth of GPS enabled devices.

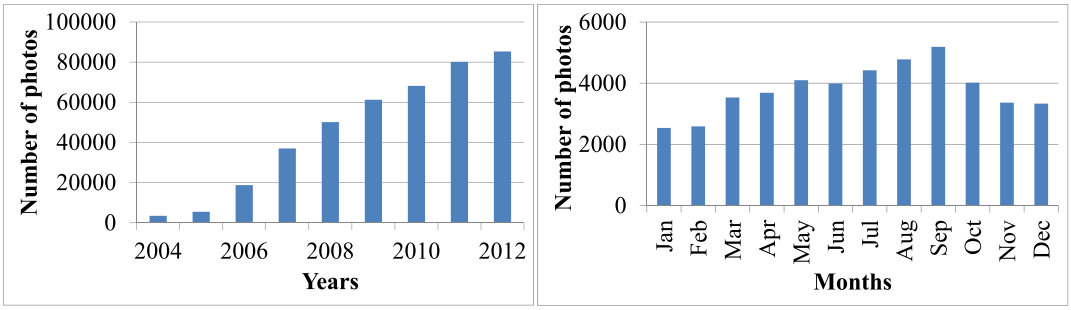

Figure 9 - Number of photos per year (left); monthly average of photos between 2004 and 2012 (right).

The chart in Figure 9 (right) depicts the monthly average of Flickr photos uploaded be-tween 2004 and 2012. Observing the chart we can see that January and February are the months with fewer photos with an average value of about 2500 photos per month, and September followed by August and July are the strongest months with an average value of around 5200, 4800 and 4400 photos per month respectively. This can be related with the fact that these are the most common vacation months

The fact that the number of photos has been growing every year since 2004 shows that the project has become more mature and this represents a positive aspect for its adequacy in LULC production activities. The number of photos is also reasonable distributed over the months and that is another positive characteristic. The fact that some types of LULC vary throughout the year means that we also need photos taken in different months in order to have a good monthly coverage. Also the satellite images used in the classification process are acquired in a specific period of the year and therefore should be assessed using information from a similar period.

3.3.1.4. Spatial distribution of Flickr photo locations

The next step was to analyze the spatial distribution of Flickr photo locations. Using the CAOP database a map was developed to demonstrate what was recognized by the visual inspection. Thereby the maps presented in Figure 10 shows the spatial distribution of the frequency of Flickr photos by each municipality: absolute number of photos (left) and normalized by area (right). This confirms what we realized visually: points are more concentrated around the biggest cities and also on the coastal side of Continental Portugal.

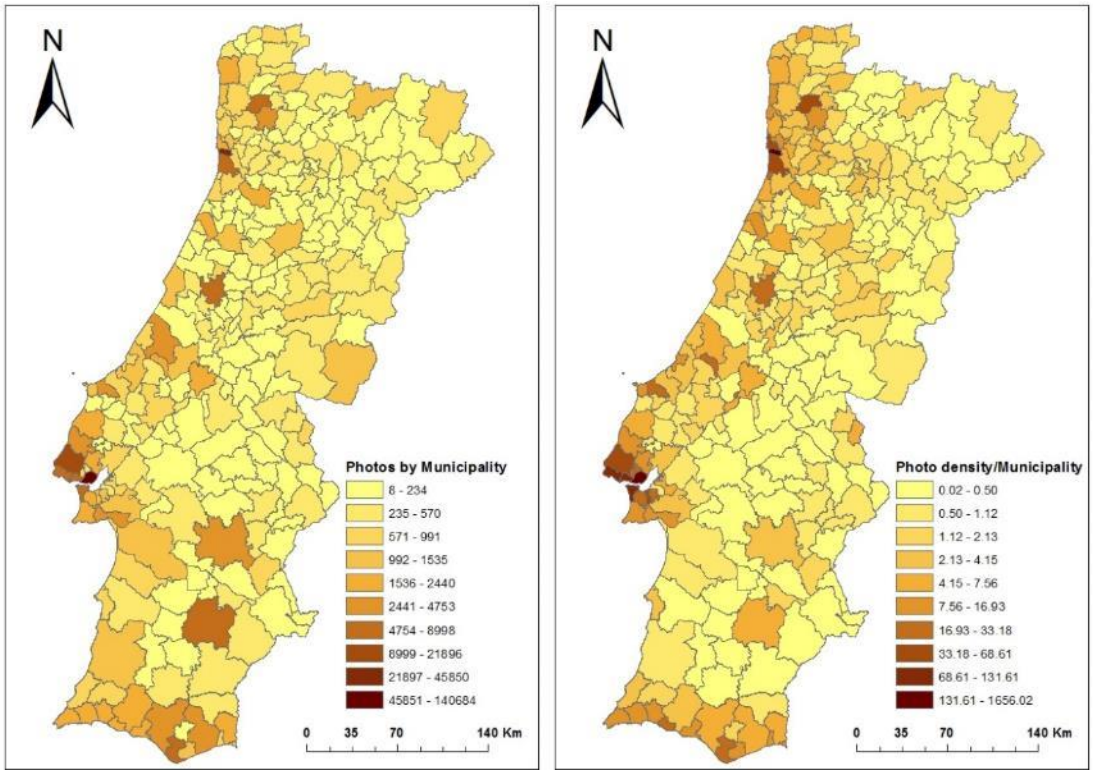

Figure 10 - Flickr photos frequency distribution by municipalities: absolute number of photos (left) and normalized by area (right)

From the 278 municipalities over Continental Portugal, 143 has less than 1 photo per Km2, 100 have between 1 and 10 photos, 28 have between 10 and 50 photos, and 7 have more than 50 photos per Km2.

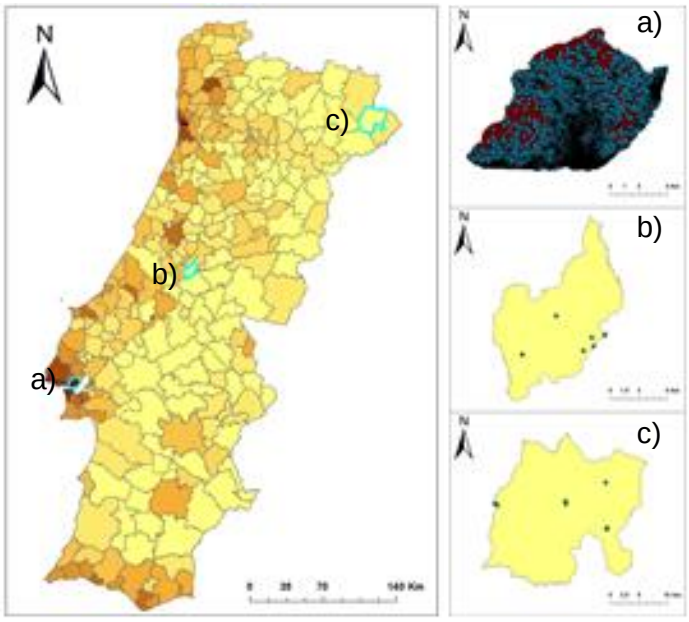

Figure 11 - Spatial distribution of Flickr photos over the municipalities of Lisboa (a), Pedrógão Grande (b) and Vimioso (c)

In depth observation was made to the municipalities with the highest and the lowest amount of points and also the highest and lowest density. Lisboa (Figure 11a), capital of Portugal, is the municipality with the highest values for number of photos and also photo density with respectively 140684 photos and around 1656 photos per Km^2. The municipalities of Pedrógão Grande (Figure 11b) and Vimioso (Figure 11c) have respectively the lowest number of photos, with 8 photos, and the lowest density with around 0.3 photos per Km2. In both cases the distribution of the points is

clustered with a bigger concentration in some places following also the tendency of the whole country.

The spatial distribution of Flickr photos over Continental Portugal is not homogeneous and the difference in number of photos and density between some municipalities is large. Although this is not a very positive characteristic, places with a larger number of photos and higher density are also more complex and, therefore, a higher number of photos will provide a better “picture” of those places.

3.3.1.5. Distribution of Flickr photo locations over CLC classes

The next step was performed to provide a picture about the distribution of the Flickr photos over the different CLC classes. Comparing the number of photos overlapping each CLC level 1 class with the corresponding area as presented in Table 15, the “artificial surfaces” class has the highest value in terms of density, very far from the other classes. This class has almost 105 photos per Km2 and all the other classes have less than half of that. The “forest and semi natural areas” and “agricultural areas” classes have even less than one photo per Km2. From the total number of photos, 322032 are located in artificial surfaces and 34270 in forest and semi natural areas representing respectively 78.58% and 9.06% from the total of Flickr photos. Agricultural areas have 8.36%, water bodies have 3.29% and Wetlands have 0.71%.

Table 16 demonstrates that, according to the level 2 of the CLC nomenclature, 248866 photos, representing 60.72% of the total photos, are located in urban fabric. The remaining 39.28% are distributed by the other categories with none of them

exceeding 10%. Therefore the photos are not well distributed over all the CLC classes with the artificial surfaces class and subclasses having more than 60% of the total.

CLC level 1 classes	Area (Km^2)	Number of photos	Density (photos/Km^2)	Percentage
1 Artificial surfaces	3088.01	322032	104.28	78.58%
2 Agricultural areas	41996.50	34270	0.82	8.36%
3 Forest and semi natural areas	42620.95	37129	0.87	9.06%
4 Wetlands	1012.31	2902	2.87	0.71%
5 Water bodies	361.75	13496	37.31	3.29%

Table 15 - Density of Flickr photos by level 1 classes of CLC

CLC Level 2 classes	Frequency	Percentage
11 Urban fabric	248866	60.72%
12 Industrial, commercial and transport units	34283	8.37%
13 Mine, dump and construction sites	408	0.10%
14 Artificial, non-agricultural vegetated areas	38475	9.39%
21 Arable land	5373	1.31%
22 Permanent crops	5094	1.24%
23 Pastures	426	0.10%
24 Heterogeneous agricultural areas	23377	5.70%
31 Forests	16709	4.08%
32 Scrub and/or herbaceous vegetation associations	12756	3.11%
33 Open spaces with little or no vegetation	7664	1.87%
41 Inland wetlands	37	0.01%
42 Maritime wetlands	2865	0.70%
51 Inland waters	6636	1.62%
52 Marine waters	6860	1.67%
Total	**409829**	**100.00%**

Table 16 - Frequency of Flickr photos by level 2 classes of CLC

3.3.1.6. Cross analysis

The cross analysis consisted of relating the different variables analyzed in the previous chapters. The chart presented in Figure 12, demonstrates that the different classes follow the same tendency and have more photos in summer and less in winter months. In the artificial surfaces class this tendency is even more evident.

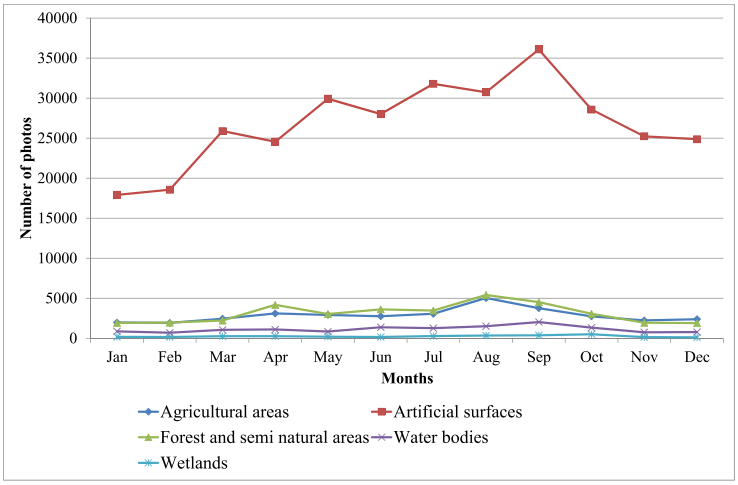

Figure 12 - Monthly distribution of photos in each CLC level 1 class

Figure 13 and Table 17 show respectively the monthly variation of photos by district (groups of municipalities) and the minimum and maximum values and respective months, and ratio of photos by district. We can verify that all the districts follow approximately the same pattern with more photos in summer and less in winter months.

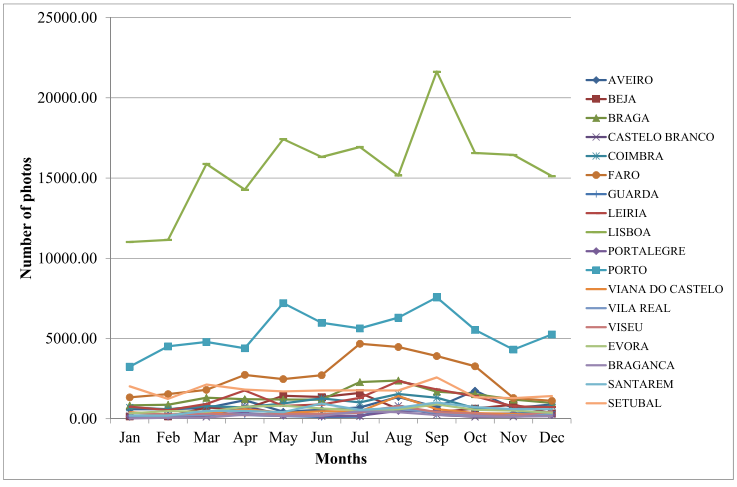

Figure 13 - Monthly variation of photos by district

Taking a closer look at Table 17 we can also realize that, in few cases, the difference between the months with minimum and maximum values is very high. The ratio value can give us an approximate idea about the seasonality of the places. In fact, the district of Bragança has around 26 times more photos in July than in January and the district of Beja has a positive variation of around 11 times the amount of photos between February and July. On the opposite side we have the district of Lisboa with a positive variation of less than 2 times the amount of photos between January and September.

District	Min		Max		Ratio
	Month	Value	Month	Value	
Aveiro	May	439	Oct	1748	3.98
Beja	Feb	140	Jul	1607	11.48
Braga	Jan	832	Aug	2382	2.86
Bragança	Jan	18	Jul	471	26.17
Castelo Branco	Mar	119	Aug	806	6.77
Coimbra	Nov	569	Aug	1542	2.71
Evora	Jan	328	Sep	897	2.73
Faro	Dec	1118	Jul	4669	4.18
Guarda	Jun	109	Aug	808	7.41
Leiria	Feb	536	Aug	2340	4.37
Lisboa	Jan	11018	Sep	21630	1.96
Portalegre	Feb	137	Aug	491	3.58
Porto	Jan	3247	Sep	7560	2.33
Santarem	Feb	197	Sep	1007	5.11
Setubal	Feb	1239	Sep	2573	2.08
Viana do Castelo	Feb	221	Aug	1413	6.39
Vila Real	Jan	154	Aug	596	3.87
Viseu	Nov	173	Aug	798	4.61

Table 17 - Minimum and Maximum values, respective months and ratio of photos by district

3.3.2. Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database

In this study we evaluated if geo-referenced and publicly available photos from the Flickr initiative can be used as a source of geographic information to help Land Use/Cover classification. Using the Corine Land Cover nomenclature, we compare the classification obtained for selected photo locations, against the classification obtained from high-resolution satellite imagery for the same locations.

3.3.2.1. Description of the study area and datasets

The defined study area is the Portuguese municipality of Coimbra, covering an area of approximately 300 km2.

Three datasets were used in this study: 1) the geo-referenced Flickr photos for the study area over the period ranging between 2004 and 2013, corresponding to a total of 4977 photos; 2) the CLC database, composed by the version 16 (04/2012) for the CLC2006 inventory, downloaded from the European Environment Agency (EEA) ; 3) the high resolution satellite imagery, with 30cm spatial resolution, available for the study area at the ArcGIS® software as basemap.

Figure 14 shows the CLC map for the study area (left) and the points corresponding to the spatial location of the photos situated in each of the three CLC classes used for this analysis, overlaid with the high resolution satellite images (right).

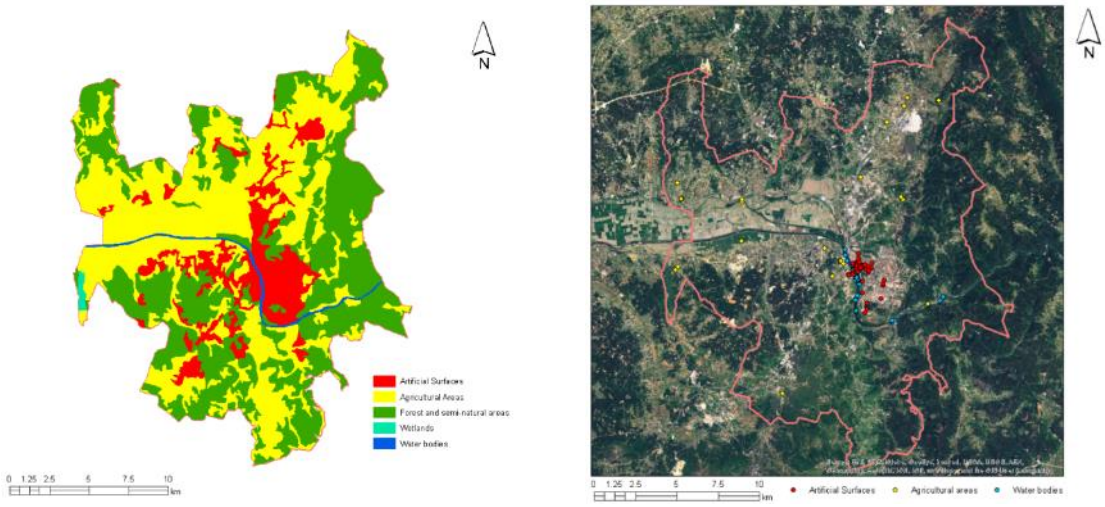

Figure 14 - CLC level 1 classes in Coimbra municipality (left) and Location of the sample Flickr photos used for the analysis (right)

3.3.2.2. Data processing

For the purpose of this preliminary study, the position associated to the 4977 photos was intersected with the CLC level 1 classes and the three classes that from a user perspective were more likely to have information were selected, namely classes 1, 2 and 5, respectively Artificial Surfaces, Agricultural Areas and Water Bodies, corresponding to a total of 4892 photos. Table 18 summarizes the distribution of selected photos over the three CLC classes.

CLC Classes	Flickr Photos
1 Artificial Surfaces	4703
2 Agricultural Areas	64
5 Water Bodies	125
Total	4892

Table 18 - Summary of Flickr photos

3.3.2.3. Methods

The methodology adopted to conduct this analysis was as follows:

1. A stratified sample of 60 photo locations was selected for each of the three classes chosen for the analysis, considering the CLC classes as strata;
2. An expert classification of Flickr photos was done, based on the image content interpretation, according to the CLC nomenclature. Using the photo assigned to each location, we first evaluate whether it was possible to attribute a class or not and, when possible, a class was then assigned to the corresponding location;
3. Flickr photo locations were overlaid with the high resolution satellite imagery, and a land cover class was assigned to each location based on the imagery interpretation.

3.3.2.4. Results and discussion

Following the methodology described in section 2.3, Table 19 and Table 20 show the resultant classification of the locations based on the interpretation of Flickr photos and satellite imagery respectively. Besides CLC level 1 classes, two more classes were considered: “Not Clear” and “Not Good”. The “Not clear” class refers to those photos where more than one class is present and it is not clear which one, if any, is predominant, and the “Not Good” class refers to those photos that do not show predominantly any type of landscape and therefore cannot be used in LULC classification.

		CLC Classes containing the photo's location		
		1 Artificial Surfaces	2 Agricultural Areas	5 Water Bodies
Classification based on Flickr photos	1 Artificial Surfaces	24	10	6
	2 Agricultural Areas	--	11	--
	3 Forest and semi natural areas	--	--	3
	4 Wetlands	--	6	--
	5 Water Bodies	--	1	34
	Not Clear	11	16	4
	Not Good	25	16	13
	Total of photos	60	60	60
	Correct	40.0%	18.3%	56.7%
	Wrong	0.0%	28.3%	15.0%
	Not clear	18.3%	26.7%	6.7%
	Not good	41.7%	26.7%	21.7%

Table 19 - Classification of Flickr photos

		CLC Classes containing the photo's location		
		1 Artificial Surfaces	2 Agricultural Areas	5 Water Bodies
Classification based on satellite imagery	1 Artificial Surfaces	60	1	--
	2 Agricultural Areas	--	39	--
	3 Forest and semi natural areas	--	--	7
	4 Wetlands	--	--	--
	5 Water Bodies	--	--	52
	Not Clear	--	20	1
	Total of points	60	60	60
	Correct	100.0%	65.0%	86.7%
	Wrong	0.0%	1.7%	11.7%
	Not clear	0.0%	33.3%	1.7%

Table 20 - Classification of Flickr photos' locations based on the satellite imagery

Having a closer look to the spatial distribution of the photos relatively to the CLC classes (see Figure 14) it is clear that for Artificial Surfaces class they are centered in the more touristic places of the city of Coimbra and for Water Bodies most photos are located in the region of the river where touristic boats operate. A more even distribution can be seen for the Agricultural Areas class.

Results from the interpretation of Flickr photos are shown in Table 19. The percentage of photos considered “not good” for LULC classification is relatively high, with 41.7% for Artificial Surfaces, 26.7% for Agricultural Areas and 21.7% for Water Bodies. Another negative aspect is related with the percentage of photos classified as “not clear”, with 18.3%, 26.7% and 6.7% for classes Artificial Surfaces, Agricultural Areas and Water Bodies respectively. These two classes together, representing photos that do not fit in any CLC class, embody a high percentage of photos with classes Artificial Surfaces, Agricultural Areas and Water Bodies getting respectively 60%, 53.4% and 28.4%. In the opposite direction, the value for locations correctly classified is very low for all the classes with the Agricultural Areas class getting the worst value, below 20%. Looking at the value for photos wrongly classified, we can see a good result for Artificial Surfaces, with 0%, while Agricultural Areas and Water Bodies had 28.3% and 15% respectively.

During the classification process, however, some problems related to the use of the Flickr photos became apparent, contributing to increase the negative aspects of this source. Among the collection of photos analyzed, we have seen photos showing predominantly people, photos taken inside houses, photos showing small details and photos taken far from what is shown in the image reflecting a high level of zoom. This last case was particularly present for photos considered inside Water Bodies, where although the picture shows mainly water it is easy to realize that the pictures were taken from land.

The assignment of classes to the photo locations using the satellite imagery produced the results shown in Table 20. It can be seen that 100% of the points

located at the Artificial Surfaces areas in the CLC map where actually assigned to the Artificial Surfaces class, with values of respectively 65% and 87% for the classes Agricultural Areas and Water Bodies.

At some locations it is not clear to which class the point should be assigned, due to the mixture of classes observed at the vicinity of the point and to the fact that the minimum mapping unit of the CLC map is 25ha, which means that the class choice cannot be done analyzing only what exists at each point, but also looking at a larger vicinity. In Table 20 it can be seen that this difficulty occurred for 20 points. However, a closer analysis showed that only 4 of these points correspond to different locations. The other 16 points, even though corresponding to different Flickr photos, actually were assigned exactly to the same spatial location, meaning that the volunteer assigned the same coordinates to a large number of photos. Moreover, an analysis of the photos as well as the photos tags also showed that there are also other photos wrongly geotagged, since the coordinates assigned are far from the real location were the photo was taken.

3.3.3. Photo based UGsC initiatives: a comparative study of their suitability for helping quality control of Corine Land Cover

In this study we conducted a preliminary analysis of the adequacy of photos from the Flickr and Panoramio initiatives in order to use them as a source of field data in the quality control of the Land Use/Cover databases production. We evaluated their temporal and spatial distributions over Continental Portugal and also their distribution

over Land Use/Cover classes using as a reference the European Corine Land Cover database.

3.3.3.1. Material and Methods

The defined study site was Continental Portugal.

The CAOP database (Portuguese official administrative boundaries database), already described in the section 3.3.1.1, and the CLC database were used.

The considered dataset is also composed by geo-referenced photos' locations resulting from two UGsC initiatives: a) Flickr and b) Panoramio. The data were downloaded using their own APIs for the study period ranging from the beginning date of each initiative (2004 for Flickr and 2005 for Panoramio) to the end of 2012, and clipped by the Portuguese boundary, in a total of 404,691 for Flickr and 261,943 for Panoramio. Originally in text format, each location is complemented by attributes such as: latitude, longitude, name, title, date of acquisition, among others. Latitude and longitude values refer to the WGS84 Spatial Reference System (SRS) used by default in GPS receivers. Some errors were found in the data downloaded from Flickr initiative and therefore, based on the date of acquisition attribute, the old locations (photos taken before 2004) and the locations with missing information were removed. The final dataset is thus constituted by 404,691 locations from Flickr (Figure 15-b), more concentrated mainly over the main cities (Lisboa and Porto) and along the country coastal lines, and 261,943 from Panoramio (Figure 15-c) better distributed over the country.

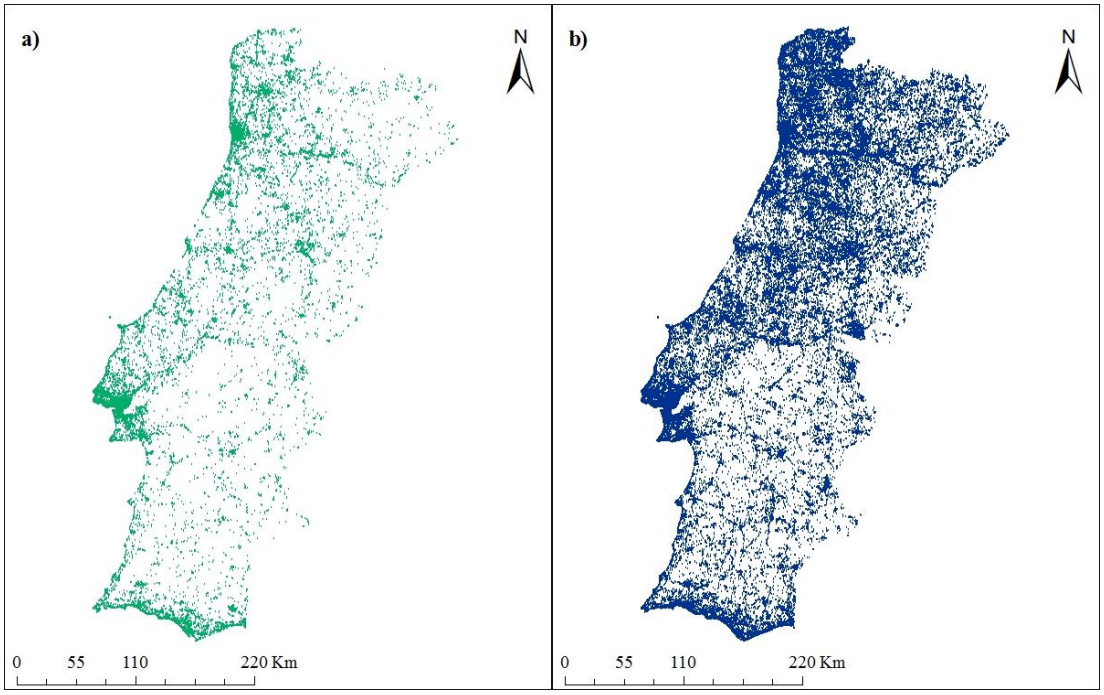

Figure 15 – Photo datasets used: a) Flickr photos' locations, and b) Panoramio photos' locations

Note: These maps give the wrong idea that Flickr have more photos than the Panoramio initiative. This is due to the fact that Flickr photos are more concentrated near to the biggest cities and touristic places, whereas Panoramio photos are spatially better distributed over the study area.

3.3.3.2. Methods

In this work we studied the distribution of photos from Flickr and Panoramio initiatives in terms of their temporal and spatial distributions, and distribution over CLC classes. In this sense, four main analyses were carried out:

1. Analysis of the temporal distribution of photos considering the date tag. We examined how the photos are distributed over the years to understand the evolution of the initiative, and over the months to understand the distribution

throughout the year. A comparison between photos from both initiatives was also executed;

2. Analysis of the spatial distribution by confronting the photos with the Portuguese municipalities, verifying and comparing the number of photos between different municipalities as well as verifying differences between both initiatives. For each municipality, the number of photos of each initiative and respective densities were calculated. To determine which initiative has more influence in each municipality, the difference of densities for each initiative was calculated and a map showing which one has positive values was produced;
3. Analysis of the distribution over the different CLC classes, prepared by overlaying the photos' locations from both initiatives with the CLC database. Each point was assigned the correspondent CLC value and the number of points for each CLC class was calculated. The CLC classes used are shown in the Table 2 and represent the level's 1 and 2 of the CLC nomenclature. A comparison between the results from both initiatives was also completed;
4. Cross analysis to compare the distribution of photos over CLC classes along with the spatial and temporal distributions. In this case Portuguese districts (groups of municipalities) were used, as administrative boundaries, for the spatial comparisons.

3.3.3.3. Temporal distribution of photo locations

Regarding the temporal distribution of photos we can verify, by observing the chart in Figure 16(a) that the number of photos has been growing since 2004 for both

initiatives. This number has grown from 3,445 in 2004 to 83,836 in 2012 for Flickr and from 6 in 2005 to 54,890 in 2012 for Panoramio. Both initiatives had a big growth at the beginning of their lives but after 2008 Panoramio has growth in a more contained way, with actually a small decrease in the number of contributions from 2008 to 2009.

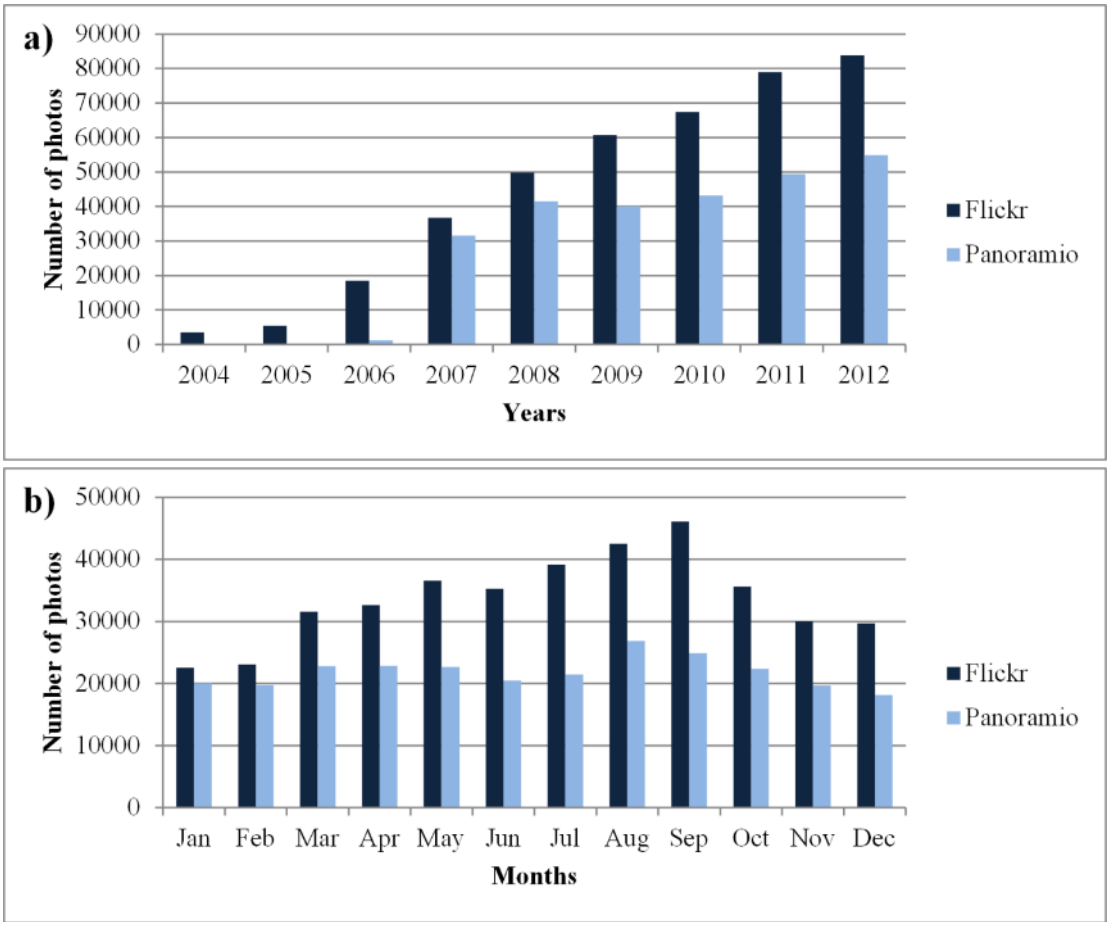

Figure 16 - a) Number of photos per year; b) Monthly distribution of photos between 2004 and 2012

The chart in Figure 16(b) depicts the monthly distribution of photos uploaded between 2004 and 2012. Observing the chart we can see that the months with the lowest number of photos are January for Flickr, with 22,532 pictures, and December for Panoramio with 18,138 pictures. On the other side, the months with more number of photos were September for Flickr with 46,102 pictures and August for Panoramio with 26870 photos. This can be related with the fact that August and September are the most common vacation months in opposition to December and January.

The fact that the number of photos has been growing every year since the beginning shows that these projects have become more mature and this represents a positive aspect for its adequacy in LULC production activities. The number of photos is also reasonable distributed over the months and that is another positive characteristic. The fact that some types of LULC vary throughout the year means that we also need photos taken in different months in order to have a good monthly coverage. Also the satellite images used in the classification process are acquired in a specific period of the year and therefore should be assessed using information from a similar period.

3.3.3.4. Spatial distribution of photo locations

Looking at the maps presented in Figure 15 (a and b) it is possible to observe that the spatial distribution of photos from Panoramio over the study area is more homogeneous than Flickr. Those from Flickr are more concentrated around the biggest cities and along the Atlantic and Mediterranean coasts, forming a clustered distribution, whereas from Panoramio, only a portion of the country, at the center south region, is less covered. The clustered distribution of Flickr photos can be

explained by a bigger number of people living in these places and also by the presence of a higher number of touristic attractions and beaches. For the case of Panoramio, the higher level of homogeneity in their photo distribution might have a direct connection with the approach of the initiative, more focused on exploiting places rather than personal content.

Intersecting the CAOP database with the photos' locations of both initiatives, maps presented at Figure 17 were developed to demonstrate what was recognized by the visual inspection. Thereby the presented maps show the spatial distribution of photos' densities over the study area from Flickr (a) and Panoramio (b). This confirms what we realized visually: Panoramio initiative has a more homogeneous distribution while photos from Flickr initiative are more concentrated around the biggest cities and also on the coastal side of continental Portugal. Figure 17(c) shows which initiative has the higher value by municipality. In fact, Panoramio has higher density values than Flickr in most of the municipalities (green vs. yellow), probably as a consequence of its higher homogeneity in the distribution of photos over the study area.

Table 21 shows a comparison of the number of municipalities with different densities of photos, between both initiatives and their sum. This gives us a good idea about the spatial distribution for each source and, as expected, from the 278 municipalities, Flickr has 143 with less than 1 photo per Km2, confirming that photos are clustered around biggest cities and touristic places. This number decreases significantly to 45 if Panoramio is used and, consequently, if both sources are used, only 25 municipalities have less than 1 photo per km2. Only the class of municipalities with

more than 50 photos per km2 does not follow this trend, with 7 for Flickr against 2 for Panoramio, although this number increases to 11 if both sources are used.

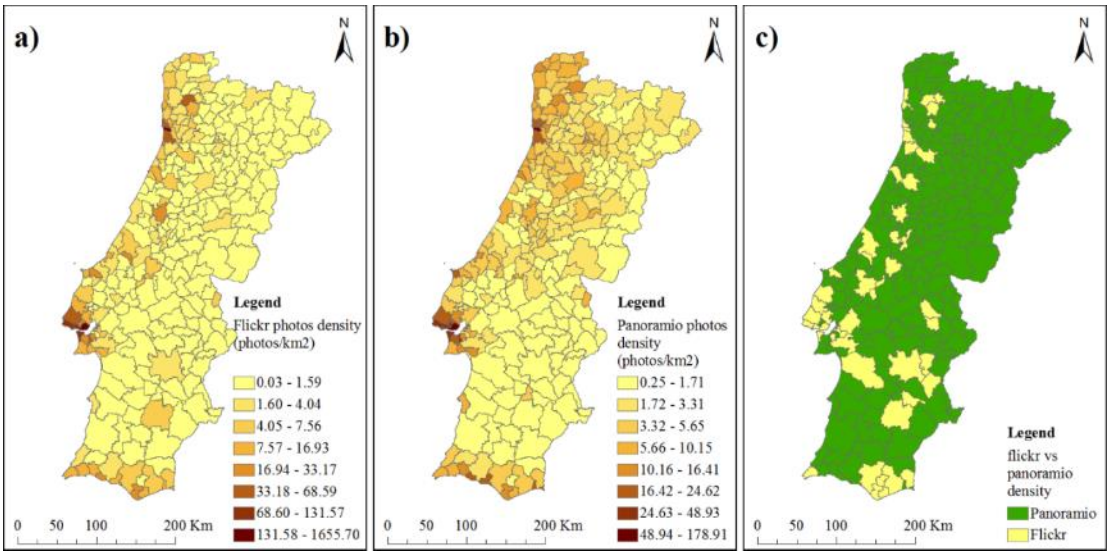

Figure 17 - Spatial distribution of photos density: a) Flickr photos density b) Panoramio photos density and c) Flickr (yellow) vs. Panoramio (green).

Note: The maps a and b demonstrate the concentration of photos near the biggest cities and touristic places. They have different scales due to their difference in terms of number of photos.

Density (photos/km^2)	Flickr	Panoramio	Flickr + Panoramio
< 1	143	45	25
≥ 1 and < 5	82	165	145
≥ 5 and < 10	18	33	51
≥ 10 and < 20	18	19	23
≥ 20 and < 50	10	14	23
≥ 50	7	2	11

Table 21 - Number of municipalities with different densities

These two sources of photo based initiatives have significant differences in terms of spatial distribution with Panoramio being more homogeneous than Flickr on one side, and Flickr having around more 50% of photos than Panoramio on the other side. It is therefore clear that using both sources together the spatial distribution of

photos becomes more balanced. In any case places with larger number of photos and higher density are also more complex and, therefore, a higher number of photos will provide also a better "picture" of those places.

3.3.3.5. Distribution of photo locations over CLC classes

This study was performed to verify the distribution of photos over the different CLC classes. Table 22 and Table 23 summarize the results of this analysis against CLC level 1 and CLC level 2 classes, respectively. Regarding level 1 classes, class 1, artificial surfaces, obtains the highest values for all sources of photos in terms of number and density, very far from the other classes. On the opposite side, class 4, wetlands, gets the lowest values.

CLC level 1 classes	Area (Km²)	Number of photos			Photo density (photos/km²)			Distribution (%)		
		Flickr	Panoramio	Both	Flickr	Panoramio	Both	Flickr	Panoramio	Both
1. Artificial surfaces	3097.17	323916	108428	432344	104.59	35.01	139.59	80.04	41.39	64.85
2. Agricultural areas	41991.77	31178	73921	105099	0.74	1.76	2.50	7.70	28.22	15.77
3. Forest and semi natural areas	42596.42	35091	62221	97312	0.82	1.46	2.28	8.67	23.75	14.60
4. Wetlands	287.77	2357	2341	4698	8.19	8.13	16.33	0.58	0.89	0.70
5. Water bodies	1109.07	12149	15032	27181	10.95	13.55	24.51	3.00	5.74	4.08
Total	**89082.20**	**404691**	**261943**	**666634**				**100.00**	**100.00**	**100.00**

Table 22 - Density of Flickr and Panoramio photos by level 1 classes of CLC

Following the trend identified in the spatial distribution of photo locations, photos from Panoramio shows, also here, a better distribution over CLC classes, which have

a positive influence when both sources are used simultaneously. In this case, the minimum density value is of 2.28 photos per km2 for class 3, Forest and semi-natural areas, whereas class 1, Artificial surfaces got an impressive value of approximately 140 photos per km2.

Table 23 demonstrates that, according to the level 2 of CLC nomenclature, class 11, Urban fabric, gets the highest values, with 246985 Flickr photos and 88902 Panoramio photos, representing respectively 61.03% and 33.94%, while the lowest values belong to class 41, Inland wetlands, with 39 Flickr photos and 67 Panoramio photos, representing 0.01% and 0.03% correspondingly.

CLC Level 2 classes	Number of photos			Distribution (%)		
	Flickr	Panoramio	Both	Flickr	Panoramio	Both
11 Urban fabric	246985	88902	335887	61.03	33.94	50.39
12 Industrial, commercial and transport units	35552	9538	45090	8.78	3.64	6.76
13 Mine, dump and construction sites	365	1202	1567	0.09	0.46	0.24
14 Artificial, non-agricultural vegetated areas	41014	8786	49800	10.13	3.35	7.47
21 Arable land	4615	11731	16346	1.14	4.48	2.45
22 Permanent crops	4661	11055	15716	1.15	4.22	2.36
23 Pastures	424	1484	1908	0.10	0.57	0.29
24 Heterogeneous agricultural areas	21478	49651	71129	5.31	18.95	10.67
31 Forests	16215	20326	36541	4.01	7.76	5.48
32 Scrub and/or herbaceous vegetation associations	11338	28037	39375	2.80	10.70	5.91
33 Open spaces with little or no vegetation	7538	13858	21396	1.86	5.29	3.21
41 Inland wetlands	39	67	106	0.01	0.03	0.02
42 Maritime wetlands	2318	2274	4592	0.57	0.87	0.69
51 Inland waters	6644	7227	13871	1.64	2.76	2.08
52 Marine waters	5505	7805	13310	1.36	2.98	2.00
Total	**404691**	**261943**	**666634**	**100.00**	**100.00**	**100.00**

Table 23 - Number and distribution of photos by CLC level 2 classes

3.3.3.6. Cross analysis

The cross analysis tried to relate the different variables analyzed in the previous chapters. Since has become clear that using photos from both initiatives together leads to better results, this cross analysis was done by using them as the source of photos instead of analyze each one separately.

The chart presented in Figure 18, demonstrates that the different classes follow the same tendency, determined in the temporal analysis, having more photos in summer and less in winter months. This trend becomes even more evident for the Artificial surfaces class.

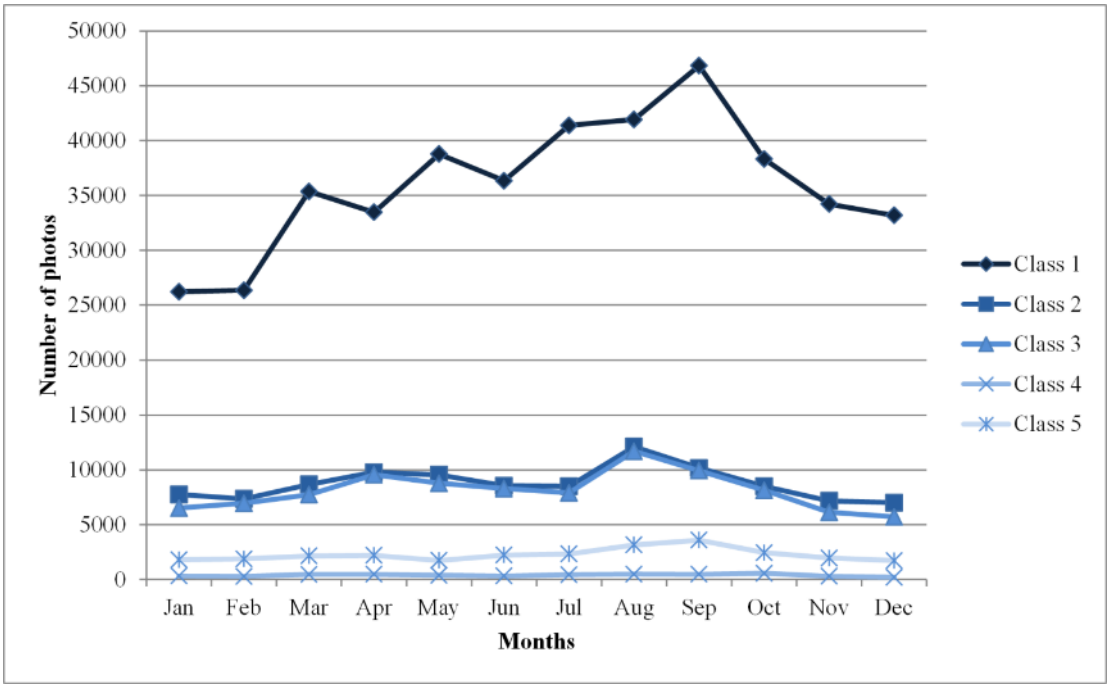

Figure 18 - Monthly distribution of photos in each CLC level 1 class

More important than looking for absolute numbers, is to verify how well covered are classes in each of the study site regions. In this sense, Figure 19 shows the density of photos by CLC level 1 class for each district, where two districts, actually the two most populated and most important regions in Portugal, stand out clearly from the other regions: Lisboa and Porto. Class 1 is also dominant in most of the regions but some of them, such as Leiria, Viana do Castelo and Porto shows an equilibrium between class 1 and 5, with the last one having actually a higher density for class 5.

District	Min		Max		Ratio
	Month	Value	Month	Value	
Aveiro	Jan	1319	Oct	2799	2.12
Beja	Feb	666	May	2181	3.27
Braga	Feb	1968	Aug	3922	1.99
Bragança	Nov	626	Aug	1987	3.17
Castelo Branco	Nov	801	Apr	1985	2.48
Coimbra	Nov	1368	Aug	3307	2.42
Évora	Dec	647	Apr	1282	1.98
Faro	Dec	2625	Aug	7403	2.82
Guarda	Feb	930	Aug	2136	2.30
Leiria	Jan	1335	Aug	3777	2.83
Lisboa	Feb	13602	Sep	24240	1.78
Portalegre	Feb	519	Apr	1128	2.17
Porto	Jan	4793	Sep	9282	1.94
Santarém	Jan	828	Sep	1834	2.21
Setúbal	Dec	2455	Sep	5818	2.37
Viana do Castelo	Jul	1426	Aug	2861	2.01
Vila Real	Jan	990	Aug	1814	1.83
Viseu	Dec	1383	Aug	2516	1.82

Table 24 - Min and Max values, respective months and ratio of photos by district

Table 24 shows the min and max values and respective months, and ratio between those values for each district. Ratio values are relatively low and equilibrated which proves that joining Panoramio photos to those coming from Flickr initiative improves and balances the distribution in opposition of using only photos from Flickr as presented in Estima and Painho (2013b). To give an example, that study revealed, for the district of Bragança, a ratio of around 26 times more photos in the month with higher level, in opposition to the month with a lower value, whereas here, the ratio for

that same district went down to a value of around 3, showing a better distribution between different months.

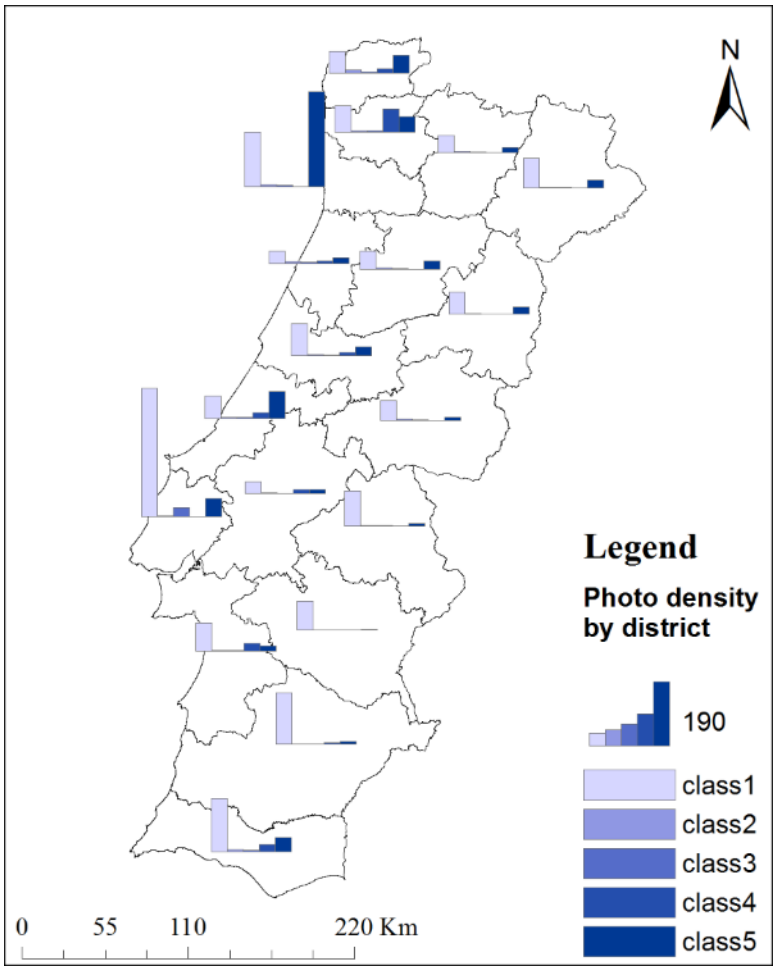

Figure 19 - Density of photos from both initiatives by district and CLC level 1 class

3.4. Conclusion

In this chapter we described the studies we have developed exploring different UGsC initiatives to investigate their potential to be used in the process of LULC

databases production. We have explored vector based and photo based initiatives and analyzed their suitability in terms of their temporal and spatial distribution, and distribution over the different LULC classes using the CLC nomenclature as a reference. We developed also a quality evaluation of a photo based initiative by comparing the classification of their photos with the classification of satellite imagery and the CLC database at the same locations with promising results.

These studies revealed strengths and weaknesses of each of the analyzed sources and two important conclusions were drawn in all the studies: 1) These sources have the potential to help in the process of LULC databases production; and 2) although some sources shown interesting results, they cannot be used alone for this purpose, and the integration of diverse sources has been advised. Such conclusions proved the importance and relevance of having a model that allows the integration of data from different UGsC initiatives into a common platform and therefore support the development of this study.

4. User Generated spatial Content-Integrator model

4.1. Introduction

As different UGsC initiatives have different goals, interests and audiences, different types of data are produced, stored with different structures and made available by different types of access. This represents additional challenges to retrieve, analyze, extract and visualize useful information from various sources and requires the development of integration models that overcome their dissimilarities.

Before starting the development of a model that integrates diverse sources of data, important decisions have to be taken. First we have to list and analyze the available data sources. Second we have to define a set of minimum requirements needed for a data source to be therefore integrated and list all the relevant sources that fulfill these requirements. Finally we need to decide on which type of integration model best fits the purpose and best integrates the selected sources.

This chapter is organized as follows. First we look at the different initiatives of UGsC, establish a list of minimum requirements for an initiative to be included in the UGsC integration model and select the initiatives that follow these requirements. Then we discuss the most important dissimilarities among the selected sources and finish by proposing a conceptual UGsC-Integratior model and drawing some final remarks.

4.2. Sources of User Generated spatial Content

Following the inventory made by Elwood et al. (2012), 99 initiatives were identified in 2009 and the most recent version of the list count 100 initiatives but no update date is mentioned (Vgi-net, 2013). The comprehensive list of initiatives is presented in appendix 1, and each initiative was checked for availability resulting in 61% of initiatives still active without changes, 3% having changed their name and 36% being not active anymore. Nevertheless, the most important and well known initiatives, referred in chapter 2, such as 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 user's location 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 need to fulfill 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:

- Type of spatial context. In this matter we found 2 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 (Hollenstein & Purves, 2010). For instance when one refers the name of a city, there is no accurate position of that city. Coordinates refer to a location with much more precision and therefore are of more interest for this study.
- Type of spatial phenomena: landscape, user position, high dynamic phenomena (natural like fires, tornados, etc., or artificial such as cars,

animals, people, etc.), static entities (buildings, roads, farms). User position and high dynamic phenomena are not of interest for this study because they do not represent physical aspects of the earth.

- 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 stated in the previous point, can have a very imprecise location. Geometries are usually georeferenced by their coordinates representing precisely geographic data.
- Type of access: no public access, access using public API's, access using private API, access using direct URL's 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 internally privacy issues so by using them, only publically available content will be accessed. Consequently, the second type is of more interest to this project.
- Type of data license: Open Data Commons Open Database License (ODbL), license to public use, license that belongs to the contributor, among others, are some of the used types of data license. It is important to note that this model will only use publically available data and will not store nor commercially exploit the data used.

- 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 to the entire globe. Depending of 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 we know exactly where it is concerning 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 removed from the study. In the same sense, for legal reasons, all the data without a free type of license were not included. Consequently, a list of essential requirements that any initiative should follow to be included in the model, presented in Table 25 , was developed:

Type of requirement	Requirement
Spatial context	Data has to be georeferenced by coordinates
Spatial phenomena	Data has 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 LULC being produced and the area of the earth being classified

Table 25 - List of essencial requirements that any initiative must have

Table 26 shows identified UGsC initiatives that follow the defined requirements, based on appendices 1 and 2, and will be used subsequently in the development of this study.

#	Name	Since	**Spatial context** (data georeferenced by coordinates)	**Spatial phenomena**	**Coverage**		**Data type**			**Access type**	**Availability**
					Global	Regional	Vector data	Photos	Descriptions		
1	Degrees Confluence	1996	X	X	X			X	X	URL	Public
2	Flikr	2004	X	X	X			X	X	API	Public
3	OpenStreetMap	2004	X	X	X		X			API	Public
4	GeographUK	2005	X	X		X				API	Public
5	Panoramio	2005	X	X	X			X	X	API	Public
6	Wikimapia	2006	X	X	X		X		X	API	Public
7	Twitter	2006	X	X	X				X	API	Public
8	Instagram	2010	X	X	X			X	X	API	Public

Table 26 – Selected UGsC initiatives

All the initiatives have the data referenced by coordinates, representing physical aspects of the earth, and are publically available. Except 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 API's to access their data, except the Degrees Confluence project where the access has to be made using

photo specific URL's. Finally, concerning the type of data, two initiatives have vector data, five are based on photos and seven of them have textual descriptions incorporated.

4.2.1. Description of the selected UGsC initiatives

For each selected UGsC initiative a brief description is provided here. These descriptions allow us to have a broader understanding of the initiatives and therefore enable the identification of similarities/dissimilarities among them.

4.2.1.1. Degree confluence[11]

This project was started in February 1996 by Alex Jarrett with the goal of "*visit each of the latitude and longitude integer degree intersections in the world, and to take pictures at each location*"[12].

The idea is to provide volunteers with a repository to upload pictures and descriptions/stories for each location creating thus "*an organized sampling of the world*". As these pictures, as well as the descriptions, are focused on describing the landscape of those locations, they are of huge interest for this project. Figure 20 shows the website of the project.

[11] http://confluence.org/

[12] http://confluence.org/infodcp.php#history

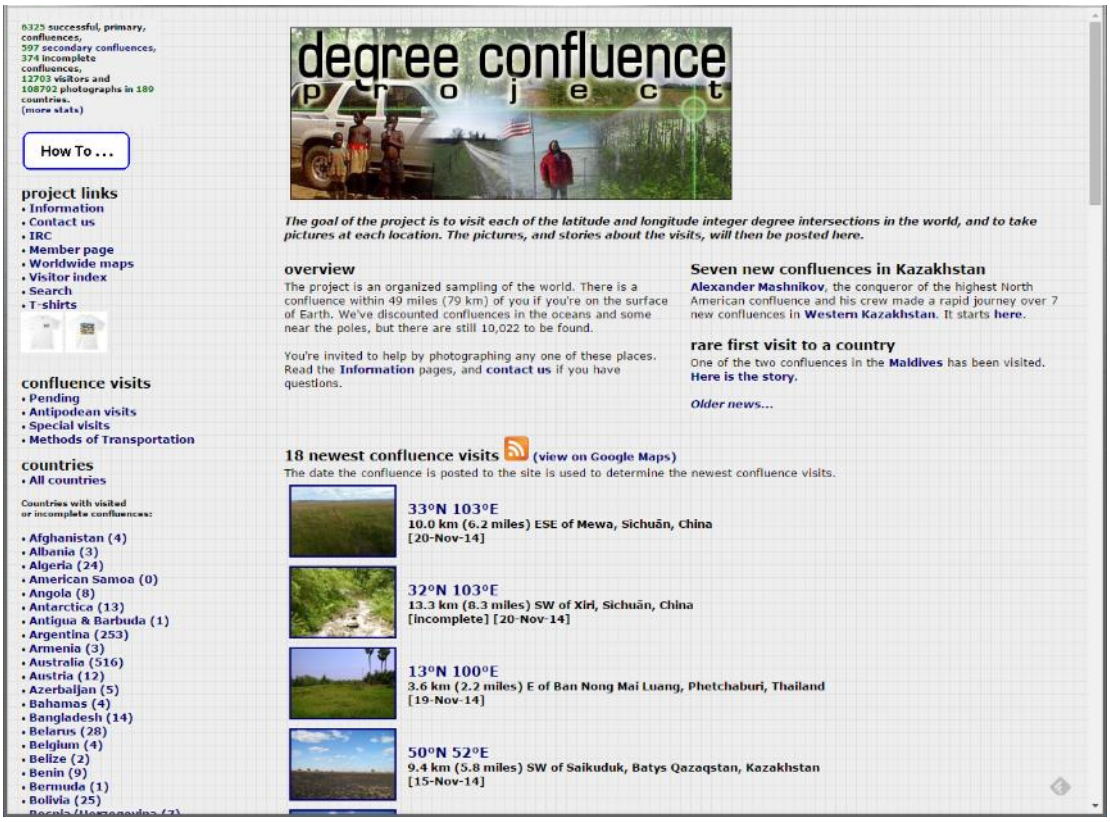

Figure 20 - Degree confluence project website

For each location volunteers are requested to follow some requirements[13]. Here is an example of the requirements related to confluences' photos:

- Preferred:
    - One (1) picture <u>of</u> the general area of the confluence, taken within 100 meters of the confluence;

[13] Extracted from http://confluence.org/infovisit.php#checklist

  - Four (4) pictures from the confluence, taken in the four cardinal compass directions (north, south, east, west), or one or more panoramic views FROM the confluence;
  - One (1) GPS photo (if a GPS was used) taken at the same location as the other photos. The photo must show the WGS84 position, and if the GPS allows for it, the altitude, reported error, and date/time.
- Minimum:
  - Two (2) pictures of or from the confluence, taken within 100 meters of the confluence;
  - Confluence visitor(s) and items belonging to them are not allowed in these photos;
  - These photos must be single-view shots (no montages), however panoramic photos are allowed.

A structured process ensures some quality control of the submitted information. After a visit to a confluence, pictures and descriptions can be uploaded through the Website. The submission will remain pending until it is validated by a regional coordinator that will ensure that all the requirements were met.

The main issue with this initiative is related with the access to pictures and descriptions. They can be accessed by navigating the Website where an interactive map is provided to navigate among visited confluences and the latest visited confluences can also be accessed through feed technology. No API is provided but a direct URL to each confluence can be used by providing their coordinates, e.g. to

access the location with a latitude of 40 degrees north and a longitude of 8 degrees west, the following URL should be used:

```
http://confluence.org/confluence.php?lat=40&lon=-8
```

This way it is possible to access the corresponding confluence Webpage where descriptions and photos can be visualized.

4.2.1.2. Flickr[14]

This initiative, already described in chapter 2, is an online application that allows photo storage and sharing where a huge amount of pictures are publically available with geotag information, in the form of latitude and longitude, among other types of tags. Photos can then be associated with a point location in a map using this spatial tag. It is therefore an interesting initiative for this study. Figure 21 shows an online map where geotagged Flickr photos can be explored. Unlike the Degrees of Confluence initiative, described previously, this initiative doesn't have any structure for quality control of photos and only the license rules and terms of use are checked

The public API[15] provided represents the easiest way to search and integrate data from this initiative with other applications, websites, etc. Particularly, it provides a search method that allows, among other filtering arguments, searching inside a

[14] http://www.flickr.com

[15] http://www.flickr.com/services/api/

certain bounding box (bbox) using a comma-delimited list of 4 values defining the area to be searched.

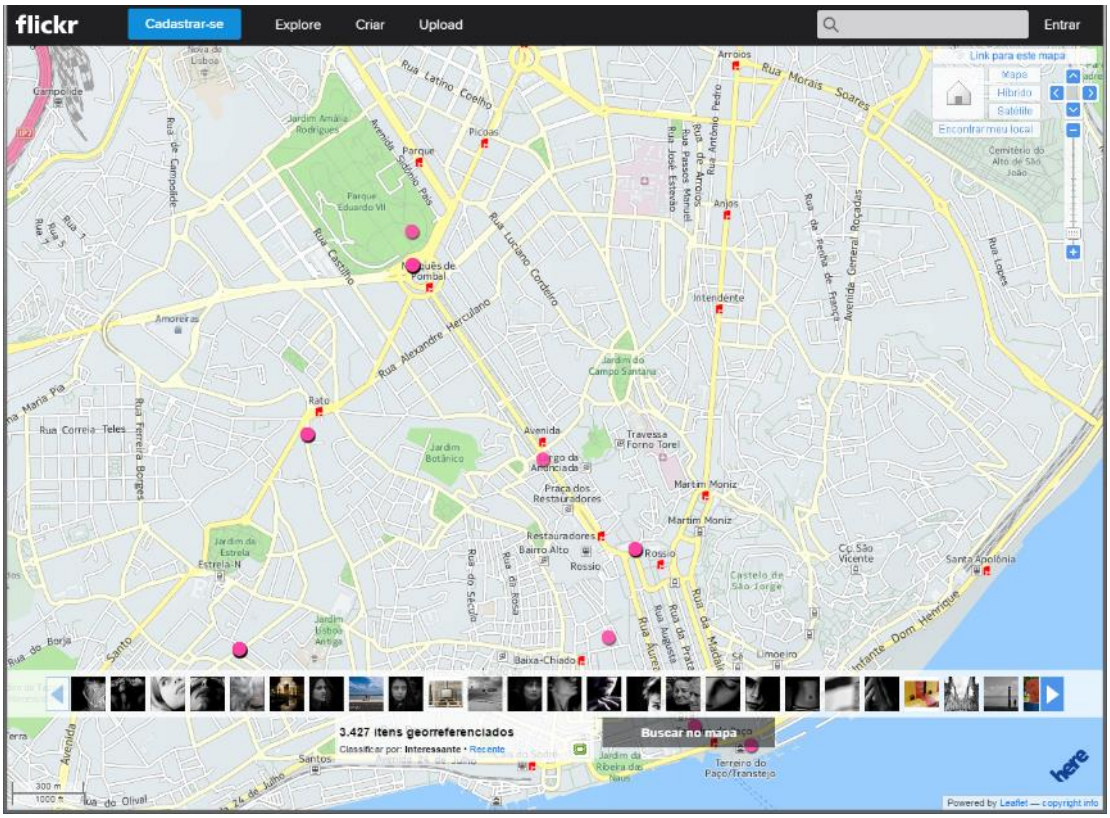

Figure 21 - Online map of the Flickr initiative

This particular method is called “flickr.photos.search” and request a list of photos according to a set of parameters. The most important elements within the server response, for the purpose of the current study, are presented in Table 27. Those elements allow to represent each photo by a point, using their coordinates, and access other important information about the photos, such as tags, titles and description, as well as the URL to open them.

Element	Description
“date_taken”	Date when the photo was taken
“latitude”	Latitude of the photo location
“longitude”	Longitude of the photo location
“tags”	Tags associated to the photo
“title”	Title given to the photo
“description”	Description of the photo
“url_?”	The direct URL to the photo, where the question mark (?) has to be replaced to the wanted size (e.g. n to small, b to large or o to original, among other possibilities)

Table 27 - Important elements from a request response from Flickr

4.2.1.3. OpenStreetMap[16]

OSM is one is the best and most studied VGI initiatives in scientific research (Elwood et al., 2012). It is a free and editable map of the whole world allowing free access to the full map dataset. Data can be stored in the form of nodes (which define a point in space), ways (which define linear features and areas) and relations (used to define the relation between other elements), and each element can also incorporate tags describing what features represent in reality. Figure 22 shows the OpenStreetMap online map that users can use to explore and also to edit and contribute.

OSM initiative provides various ways to access and manipulate data using either the official website or their public API’s. There are also third party websites and API’s with specific functionalities such as the OpenStreetMap Cycle Map[17], the OpenStreetMap Routing Service[18], among others.

[16] http://www.openstreetmap.org

[17] http://www.opencyclemap.org/

[18] http://www.yournavigation.org/

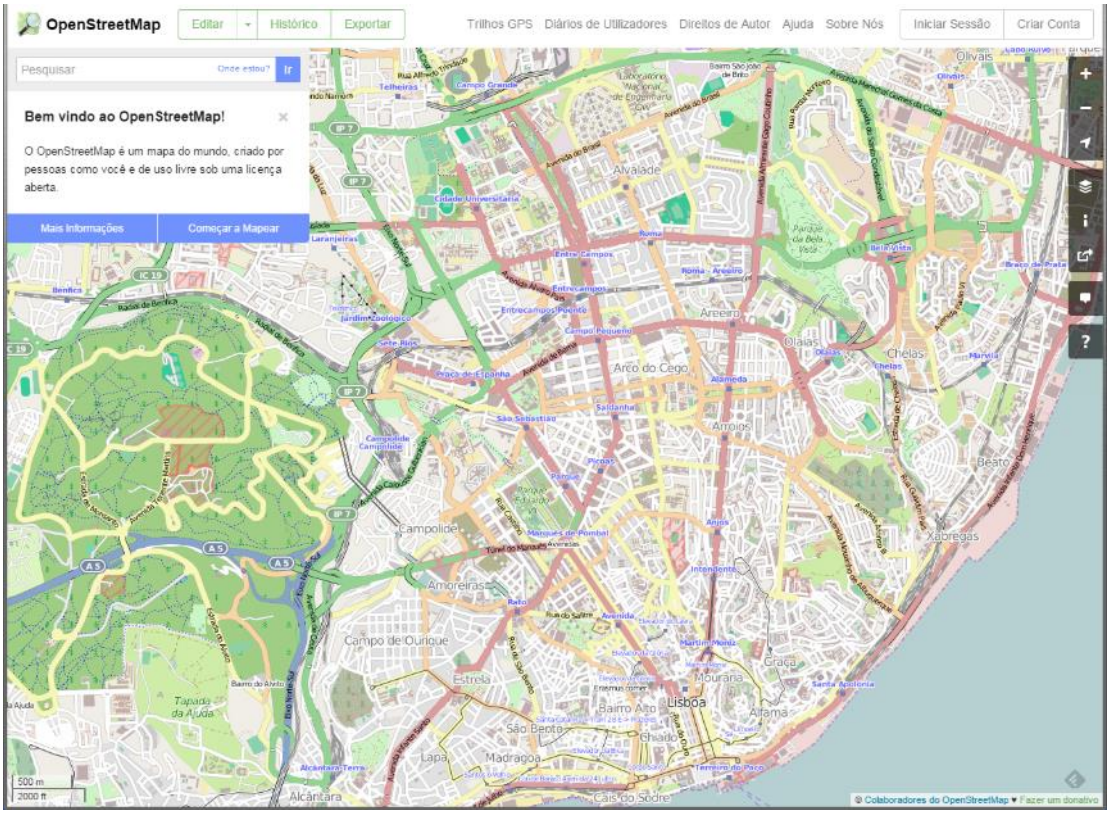

Figure 22 - OpenStreetMap online map

As the main OSM API is optimized for edition, and we are more interested only on downloading the features for displaying purposes, the best solution is to use the read-only Overpass API[19]. With this API it is possible to request all the data existing in a given bbox and get an XML response with all the elements found, along with their respective tags according to the OSM nomenclature.

[19] http://wiki.openstreetmap.org/wiki/Overpass_API

4.2.1.4. GeographUK[20]

The Geograph Britain and Ireland project (GeographUK) aims to collect geographically representative photographs and information for every square kilometer of Great Britain and Ireland.

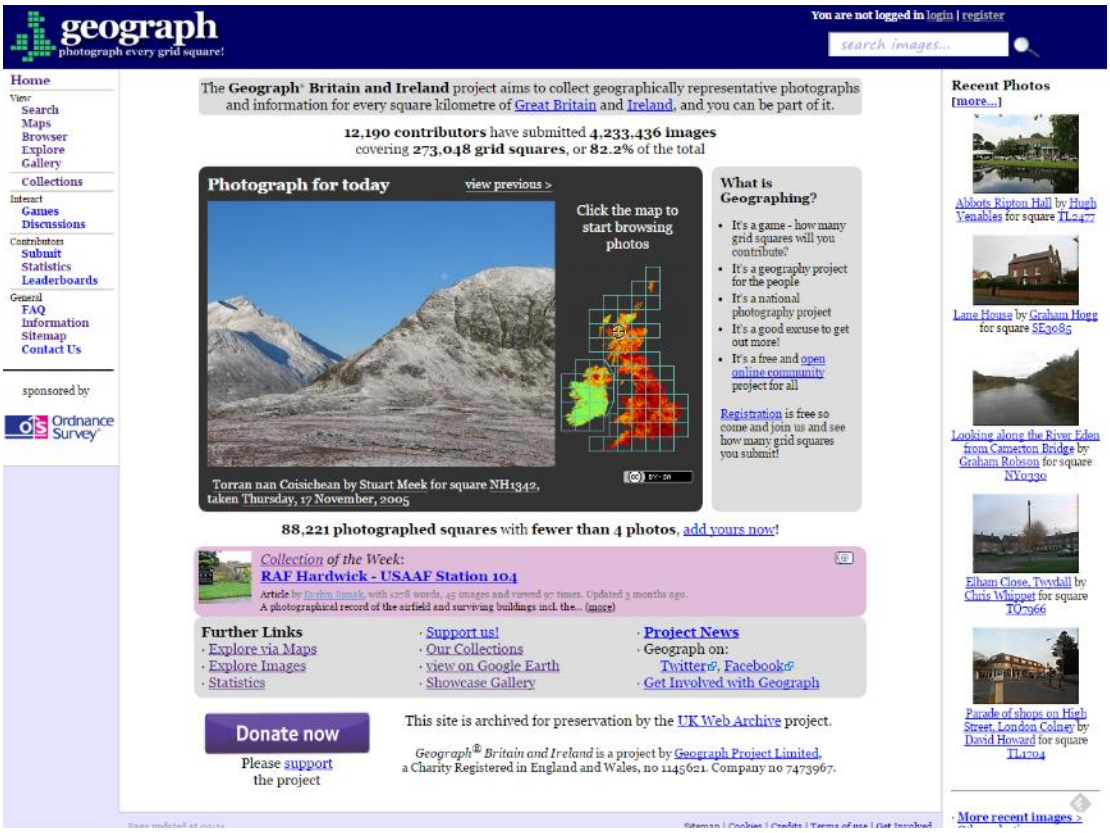

Figure 23 - GeographUK initiative website

According to the statistics available in the website, as of November 21, 2014, there were a total of 4,233,224 images with an average of 15.5 images per square within

[20] http://www.geograph.org.uk/

the 273,048 photographed squares (81.84% of the 331,983 total squares), contributed by 12,190 contributors.

The images are accessible using their official website or using their public API's. The website, shown in Figure 23, allows to explore the data using, for instance, an embedded search engine, a map engine, and a gallery.

4.2.1.5. Panoramio[21]

Panoramio initiative is a community-powered site for exploring places through photography. It is a photo sharing initiative, like Flickr, but with a remarkable difference: the photos illustrate places and do not have usually friends or family posing, which makes it very interesting and appropriate for our study. Photos can be accessed and explored through a website where a world map for browsing by location is available (Figure 24) and a public API.

The method "get_panoramas" allow to search for photos within a given bbox, in the same way as Flickr. The most important elements contained in a resulting response are presented in Table 28. Two main issues can be pointed to this API: 1) the date when the photo was taken is not available, and 2) the tags are also not provided. These issues represent a major limitation on using this source of information, and the only way to overcome them is to use the "photo_url" element to extract those elements directly from the photo webpage.

[21] http://www.panoramio.com/

Elements	Description
“photo_title”	Title of the photo
“photo_url”	URL of the page that contains the photo
“photo_file_url”	Direct URL to the photo
“longitude”	Longitude of the photo location
“latitude”	Latitude of the photo location
“upload_date”	Date when the photo was uploaded to Panoramio

Table 28 - Important elements from a request response from Panoramio

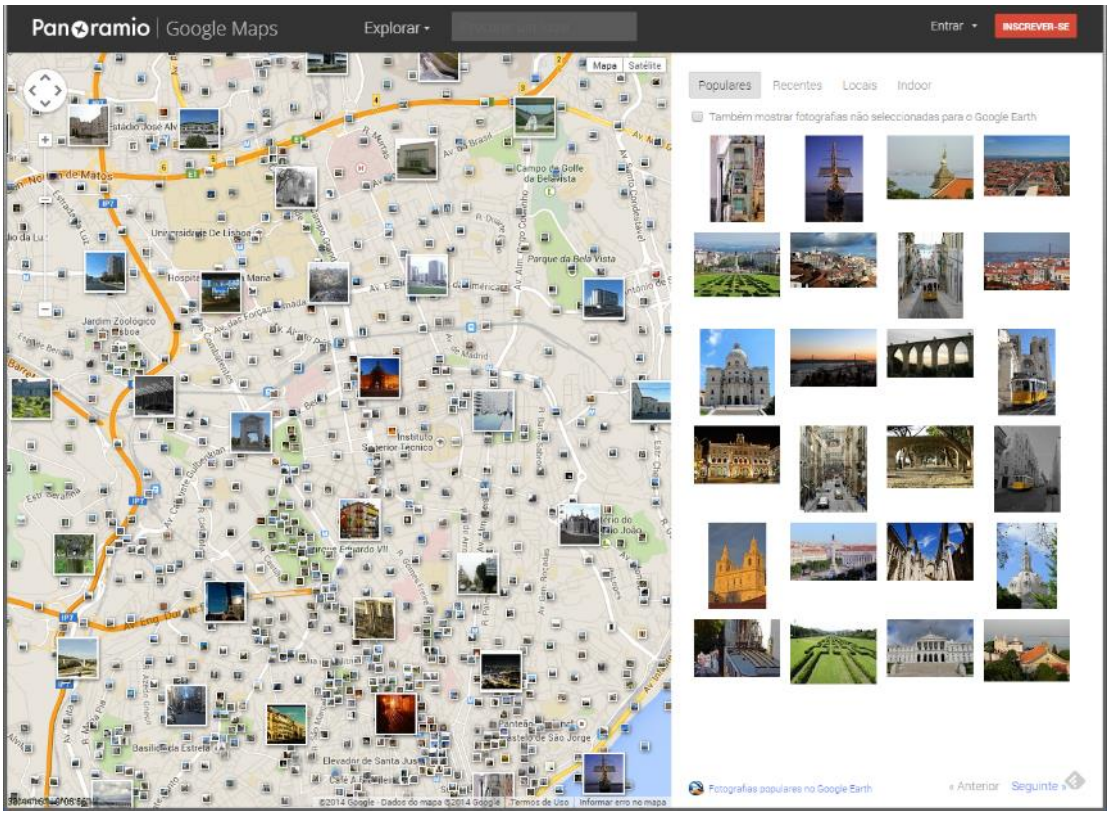

Figure 24 - Panoramio online map

4.2.1.6. Wikimapia[22]

Wikimapia is a multilingual open-content collaborative map, where anyone can create place tags to share their local knowledge. Launched in May 24, 2006, this initiative aims to describe the whole world by compiling as much useful information about all geographical objects as possible, organize it and provide free access to the public.

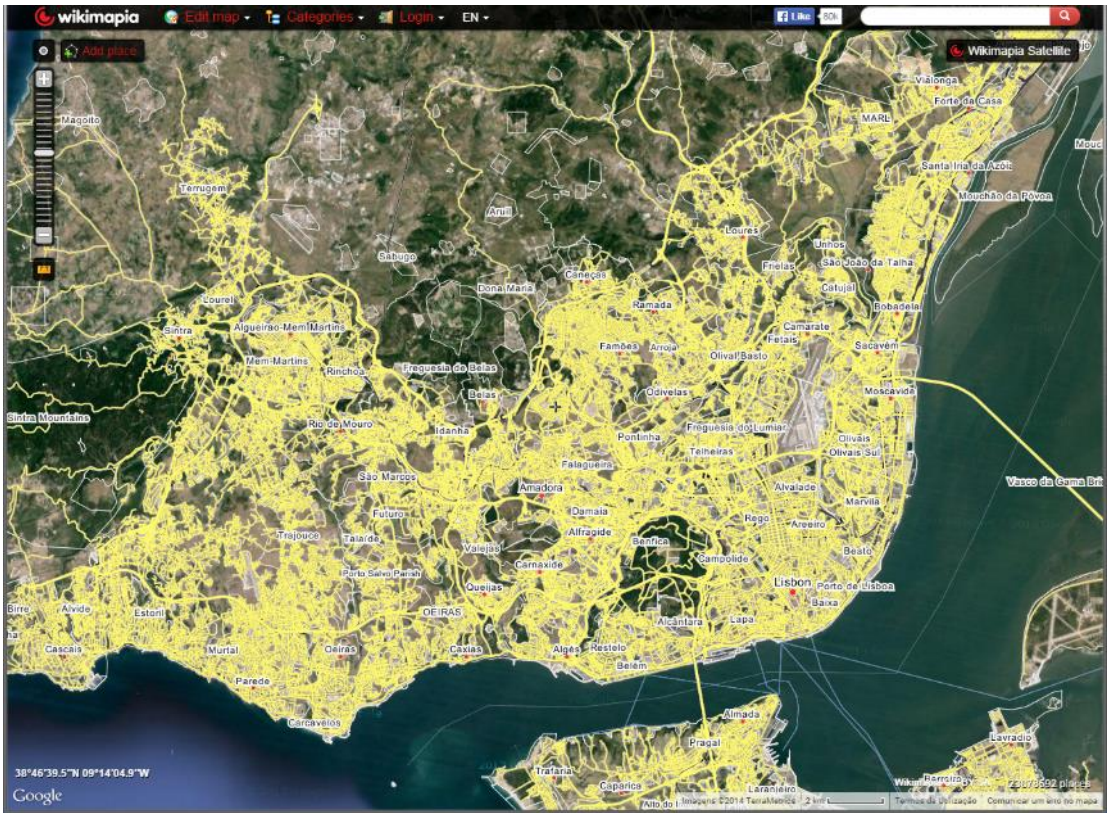

Figure 25 - Wikimapia online map

[22] http://wikimapia.org/

Data can be accessed and explored using an online GIS portal (Figure 25) and a public API. The public API allow to search for all the data available within a given bbox and the most important elements available in the response are presented in Table 29.

Elements	Description
“main”	Main information about place: url, title, description, categories, if place is a building, if it's a region. Also if it is deleted.
“geometry”	place geometry on map: polygon or rectangle
“edit”	“user_id” and name of last editor and timestamp. If the place is in deletion state this info will be in the edit block also
“location”	Place location: lat/lon coordinates, north/south/east/west coordinates, zoom level, country, state, city id and name, Wikimapia Cityguide domain name, street id and name
“attached”	Places attached to selected one or parent place of selected one, only basic info: url, title, categories. Also if child place is deleted
“photos”	Photos of current place: urls to thumb, big and fullsize photo, id, size, author id and name, date of photo uploading, last editor of this photo, photo status (deleted/active)
“comments”	Place comments: number, language of comment, author id, his ip and name, comment text, positive and negative votes, moderator id, name, and date of deletion if the comment was removed
“translate”	Languages available for selected place

Table 29 - Main elements on a search response from the Wikimapia API[23]

4.2.1.7. Twitter[24]

Twitter is an initiative that helps people create and share ideas and information using short messages, called tweets. If the user is using a mobile device with the location functionality activated, then the messages will get the coordinates of their location at the time of sharing and messages can be automatically georreferenced. Several studies using geotagged tweets have already been reported in the literature, as mentioned in chapter 2, for different purposes.

[23] http://wikimapia.org/api#placegetbyarea

[24] https://twitter.com

Users share their messages with their followers, and access messages shared by whom they are following, using either the twitter website (Figure 26) or a mobile application. For developers, a public API is also available.

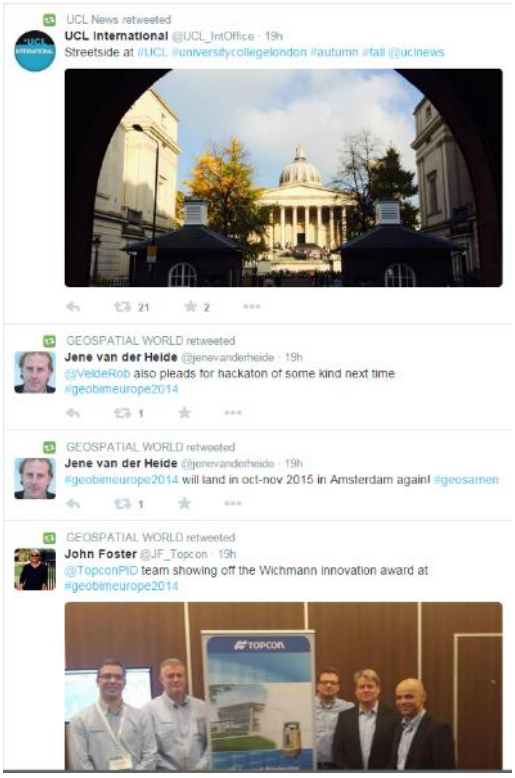

Figure 26 - Twitter timeline website

For the purpose of our study, although this can be considered geographic information, it would only be useful if text mining techniques to extract useful information from messages, outside of the scope of this research, are used.

4.2.1.8. Instagram[25]

Instagram is an initiative that aims to allow users to share their lives with friends through a series of pictures. The authors believe in a world more connected through photos.

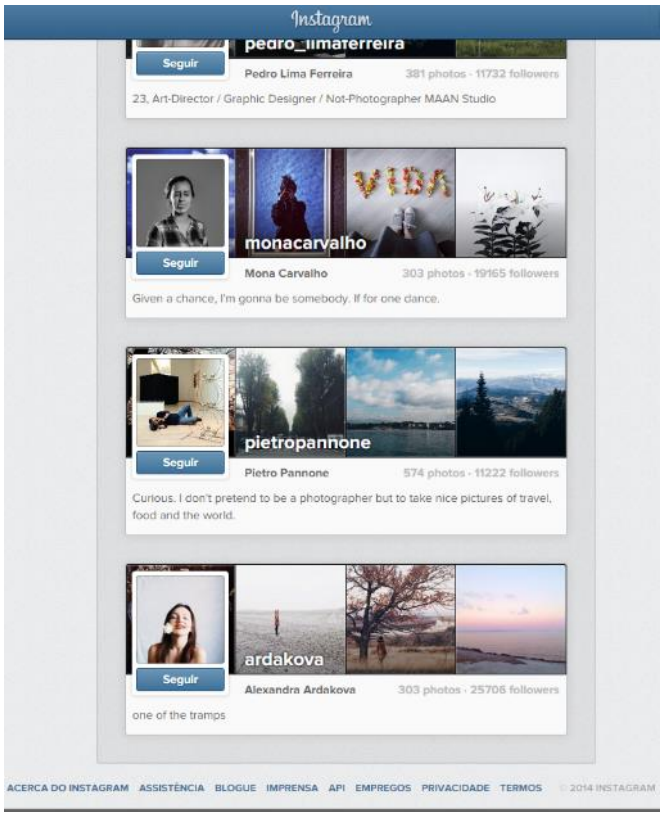

Figure 27 - Instagram website

All photos are publically available by default and accessible via Instagram mobile applications or the Instagram website (Figure 27) which do not offer the possibility of exploring content using a map. For developers, a public API to access the photos,

[25] http://instagram.com/

including the possibility to search by location using geographic coordinates, is also available.

4.2.2. Structural similarities and dissimilarities among the selected initiatives

As stated earlier, 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 explored the selected UGsC initiatives to find structural similarities and dissimilarities among them, 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 in the first type 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 could already be recognized. The first difference is related with 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.

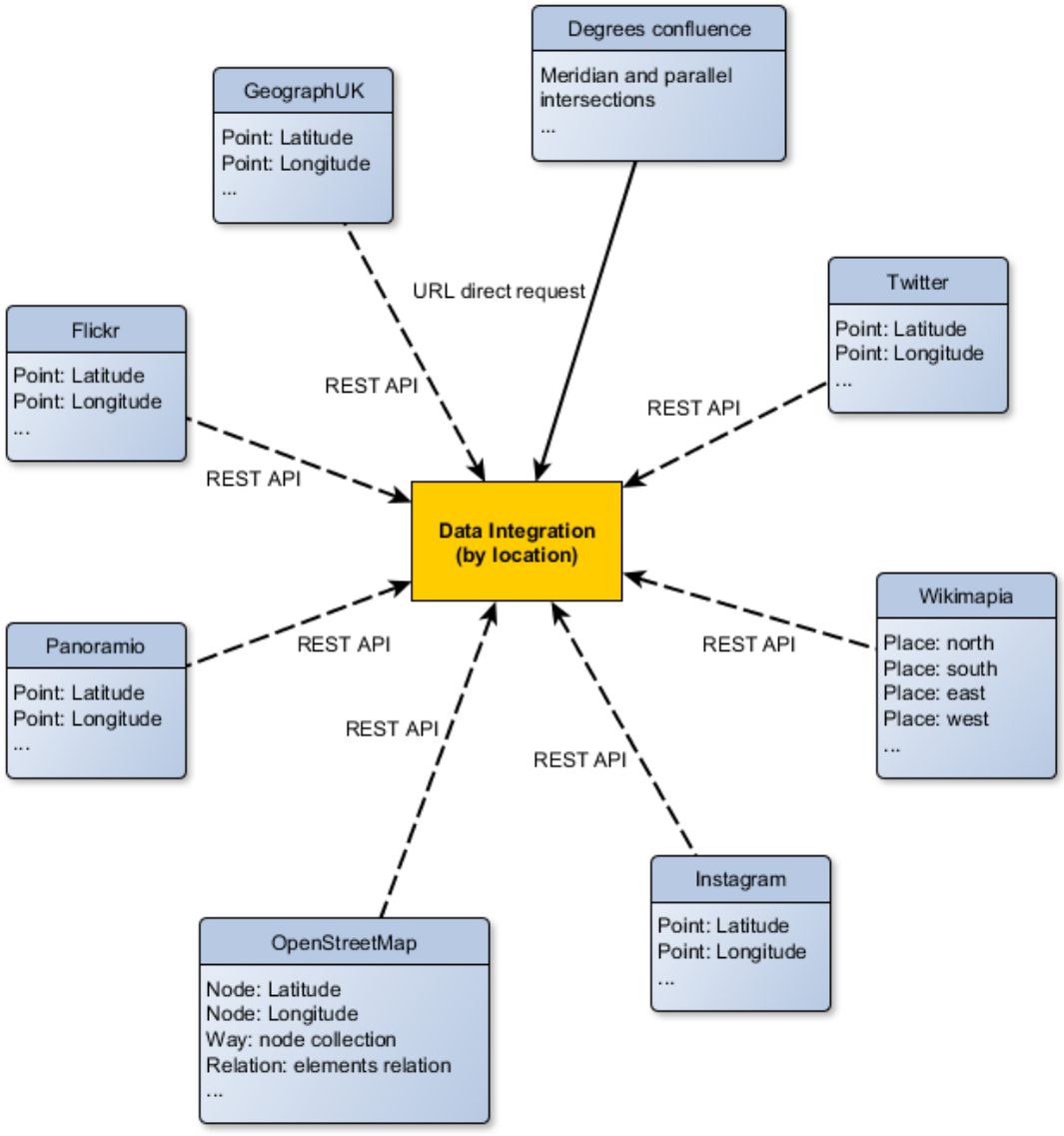

Figure 28 - Data integration by location

The former do not provide a search mechanism and needs a tailored development to retrieve information for very particular locations: the intersections of meridians with parallels degrees. The latter provides a specific interface, publically available, with known operations to retrieve the desire information from the source. Although the majority of the initiatives provide a public API to access their data, should be noted that the operations implemented by their interfaces are different from each other. Figure 28 provides a general overview of this common characteristic pointing also the type of access for each of the selected initiatives.

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

4.3. User Generated spatial Content-Integrator

4.3.1. Virtual versus materialized integration approach

There are two approaches to integrate several and diverse sources of data: 1) the virtual approach, where the information is queried and retrieved from the source on-the-fly; 2) the materialized approach, where a centralized database is developed to store data previously queried to the data sources; and 3) the hybrid approach composed by a mixture of the previous two approaches (Hull & Zhou, 1996). According to these authors, the virtual approach fits better when the information

sources are changing frequently, whereas the materialized approach would be better case the changes happen with a lower frequency.

As we already discussed, UGsC data is of the type that changes frequently. Therefore, the data integration model based on a virtual approach fits better the type of data we are dealing with, with the advantage of accessing always to the most recent data available.

4.3.2. Model architecture

The data integration model will be following 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 (Wiederhold, 1992) that stays between the application tier and the UGsC sources, as shown in Figure 29.

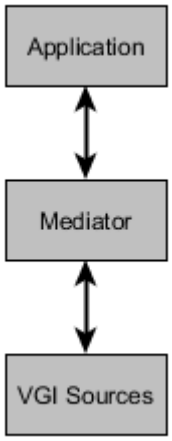

Figure 29 - High level architecture

Speaking broadly, the aim of such architecture is to ensure that the query made by the user on the application tier gets translated to the different UGsC sources, automatically, without the user having to know the structure or access type of such sources.

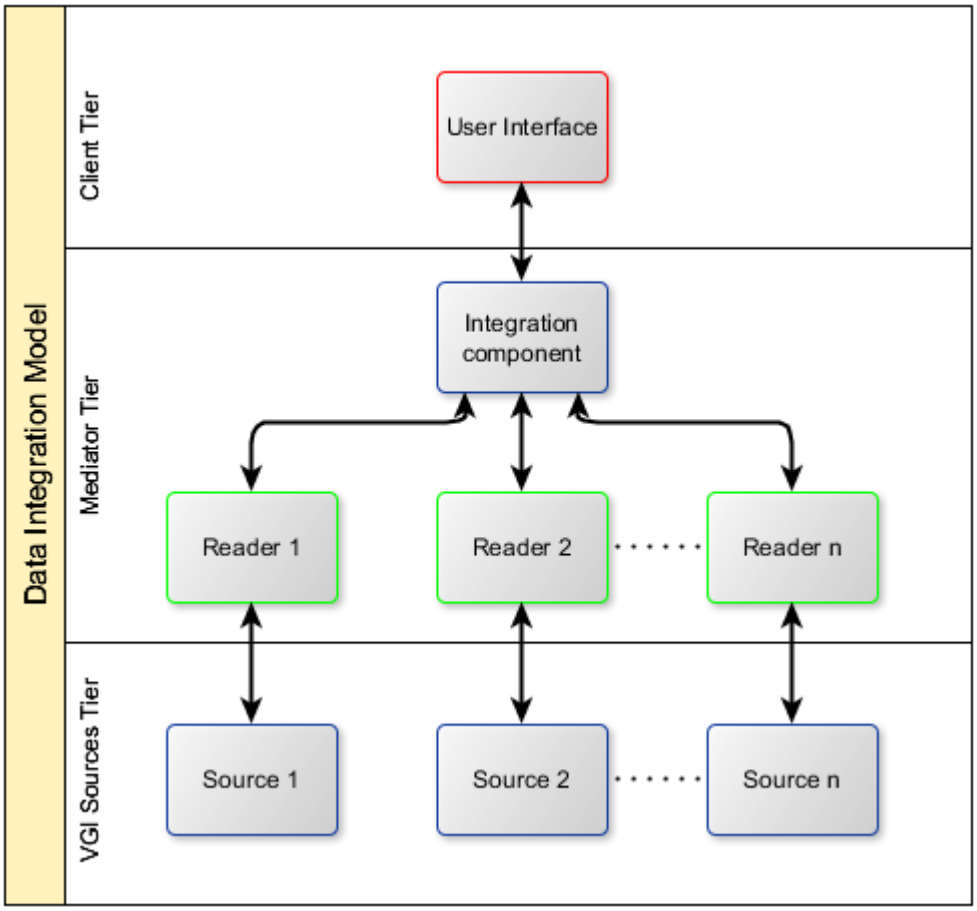

Figure 30 – Data integration model architecture

This architecture is based on three tiers or levels: the application, the mediator and the UGsC sources. The application tier is at the user level and it is responsible for

displaying in interacting with the information. The UGsC sources tier represents the sources of UGsC data containing the required information to be queried by the user. The mediator tier embodies a set of readers establishing the communication between the application and UGsC sources tiers, by translating the queries from the first towards the latter, and an integration component that incorporates the data coming from the different UGsC sources.

As already shown in Figure 28, the integration is made by overlaying the different data using their location parameters. Figure 30 presents a detailed version of the architecture of the data integration model at the three levels.

The next sections will describe each one of the model tiers in more detail.

4.3.2.1. Client tier

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

4.3.2.2. Mediator tier

The mediator tier is the core of the data integration model. As shown in Figure 31, it is composed by 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 fulfill the minimum requirements defined, by developing a specific reader for each source and adding them 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 take the decision in a fully automated way.

4.3.2.3. UGsC sources tier

This tier is composed by the data sources themselves. As already said 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 e format the response.

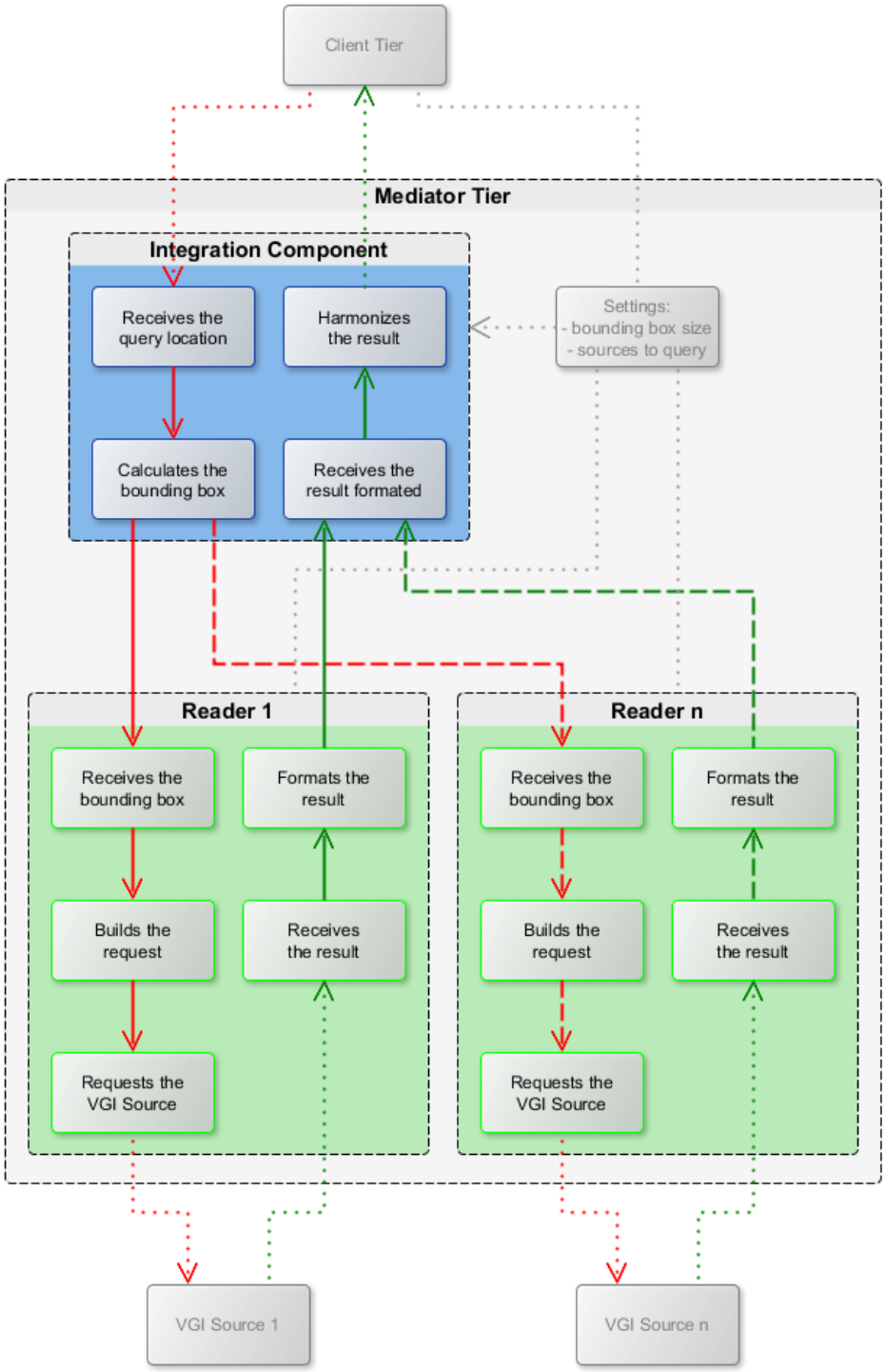

Figure 31 – Detail of the mediator tier

Note: input, output and settings' workflows respectively in red, green and grey colors

4.4. Conclusion

In this chapter we provided the architecture of a data integration model that combines diverse sources of UGsC in a common platform, 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. (2012), we identified and discussed important characteristics for the purpose of this study, and defined a set of minimum requirements that any UGsC source must fulfill to be included. A list of the current UGsC initiatives satisfying such requirements was also developed, and the identified similarities and dissimilarities were taken into account in the design of the model.

It is important to mention that the defined architecture is structured to allow the future evolution of the model by allowing the incorporation of new sources of UGsC as well as techniques that might give already some preliminary quality indicators and, ultimately, automate the decision making process by providing final quality indicators about the LULC database in evaluation.

At least one very important UGsC initiative was not included in this model. Geo-Wiki does not fulfill the minimum requirements in terms of type of access. Although the data collected has been made publically available, no API is available and the only way to get the data, besides accessing the project's online platform, is to download the entire dataset at once. This option does not fit into the virtualized integration approach used in this model.

5. Prototype development and implementation

5.1. Introduction

To validate the UGsC-Integrator model proposed in the previous chapter, a prototype was developed and implemented. In this chapter we describe the development and implementation of such prototype.

The chapter is organized as follows. We first start by defining a set of important use cases to understand which features must be included. Use cases are an important and widely used tool to capture system requirements (Neill & Laplante, 2003) that

are then used to design the system. Then we used the developed prototype to solve these use cases and prove the validity of the model.

5.2. Defining the use cases

Use cases have been one of the most used techniques for defining system requirements. To determine the requirements for the development of the prototype we started by defining important actors that could benefit and use such an application. We identified four main actors: 1) a photo-interpreter who would use the application to clarify the classification of certain dubious places; 2) a cartography validator who would use the application to help in the validation process of produced cartography; 3) a landscape architect who would use the application mainly to look at pictures around a specific location to get a sense on the surroundings; and 4) a programmer who would use the application to download the data available at a certain location and use it for other related purposes. Then we defined important cases for each of these actors taking into account their specific needs. These four use cases were identified to demonstrate also the broader application of the model proposed in this study.

Figure 32 shows an integrated view of these four actors and their respective cases in an integrated view. The basic cases, such as defining location, selecting initiatives to query, visualizing the retrieved features in an integrated map, or selecting features by tag, are shared by all the users. Advanced tasks are more related and useful to specific uses. The photo-interpreter and the landscape architect are more interested in observe data directly in the platform and therefore the view feature's info case is

more useful to them. The photo-interpreter is additionally interested to look at the tag statistics. On the other hand, the cartography validator and the programmer are also interested in downloading the data for further analysis or for use it in external applications. Consequently, the export data case is very useful for them.

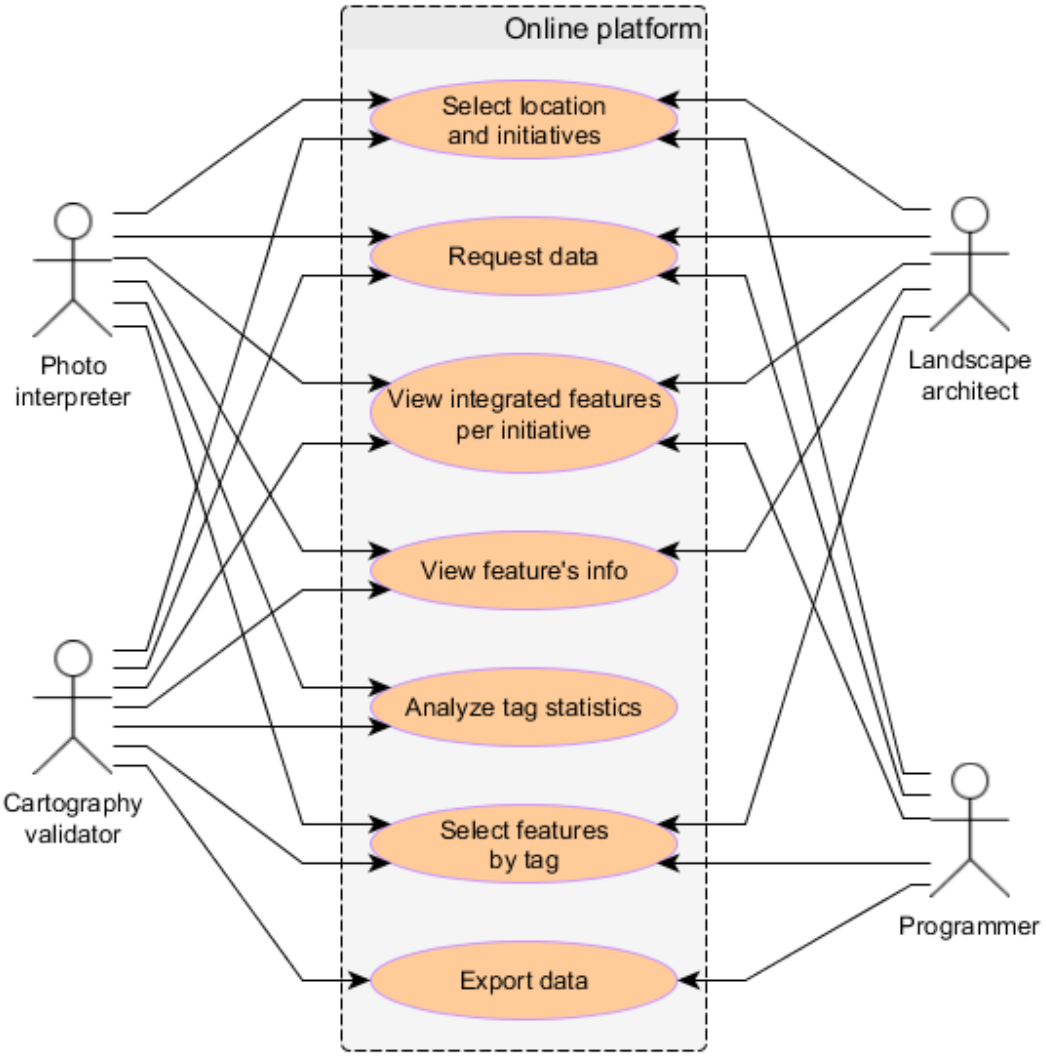

Figure 32 - Integrated view of the four identified use cases

Each one of these use cases is depicted in the next sections where more detailed diagrams are provided. Besides, each case will be solved using the developed prototype, thus demonstrating its usefulness.

5.2.1. Photo interpretation use case

As already described, LULC databases production is mainly done by means of satellite imagery interpretation. During this process, it often happens that the interpretation is ambiguous or not clear. This first use case, shown in Figure 33, illustrates how a photo-interpreter can use the platform to get and analyze ancillary data from UGsC sources to help in the classification process.

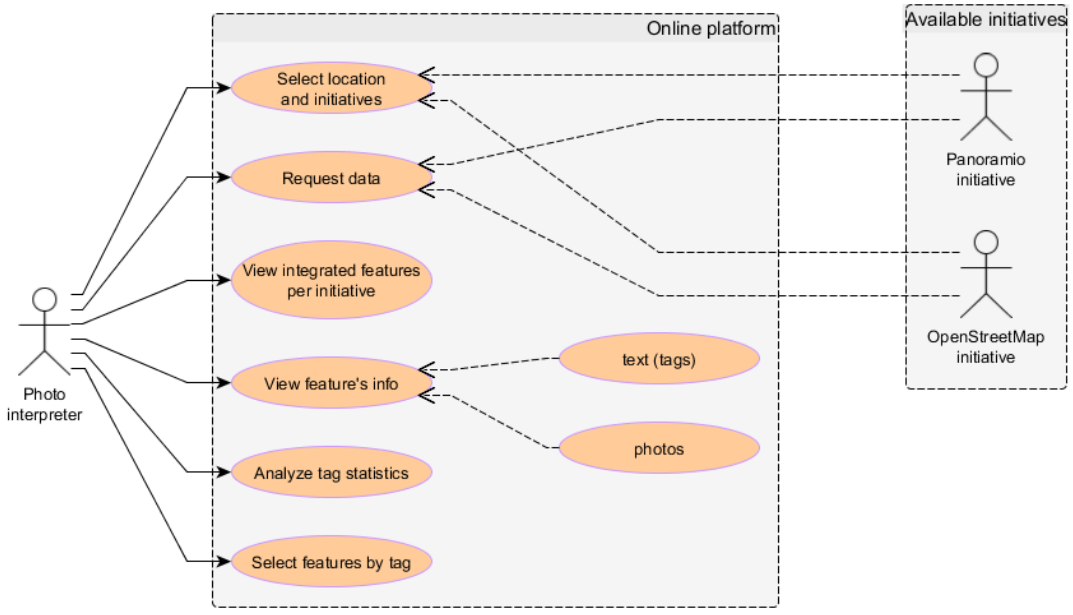

Figure 33 - Photo-interpreter use case diagram

The photo-interpreter accesses the application, selects the location as well as the initiatives he wants to query and requests the data based on these input parameters. The available data is then downloaded and presented in an integrated manner based on the geo-location of their features. The photo-interpreter can then view the integrated features spatially represented in a map and filter them by initiative, additional information available for each feature that might include text and/or photos, and also some basic statistics related to the features' tags. The selection of features by tag is also available to analyze features with specific tags.

5.2.2. Cartography validation use case

In this second use case, depicted in Figure 34, the possible uses of a cartography validator are demonstrated. The validation process is a very important and one of the last steps of any cartography or spatial databases production chain, such as LULC production. As already said, this step is performed to calculate quality indicators about any produced cartography or spatial database, by comparing them, for randomly selected sites, to reference data. The idea behind this use case is to use data from UGsC sources as reference data to validate produced cartography or spatial databases.

In this case the validator accesses the application, defines the inputs, including the location that he wants to validate and the UGsC initiatives he wants to query, and requests the data. The resulting data is then integrated in a map using the features geo-location giving the user the possibility to explore them together or filtered by initiative. For each feature it is possible to access the respective attributes including

text and/or photos. Basic statistics on tags are provided as well as the possibility to filter features by tag. Because this use case might require additional analysis, the validator might export all or pre-selected features to further analyze them in a desktop software.

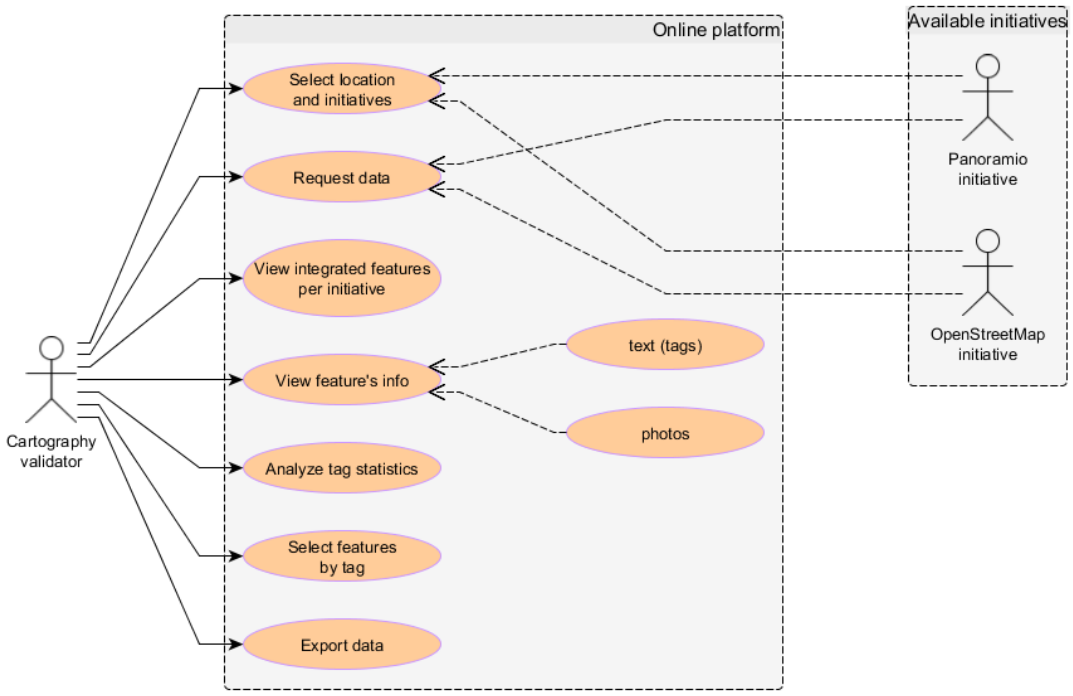

Figure 34 - Cartography validator use case diagram

5.2.3. Landscape architecture use case

In this use case, shown in Figure 35, a landscape architect uses the application to get a sense on the surroundings of a selected location. He accesses the application and defines the input parameters including the location to observe and the initiatives to query, followed by the request of available data. These data is then integrated in

an integrated map allowing to filter by initiative and select features by tag. The access to the attributes of each feature, including text but especially photos, is of extreme importance for this type of user as photos provide better insights and additional visual context on the surroundings.

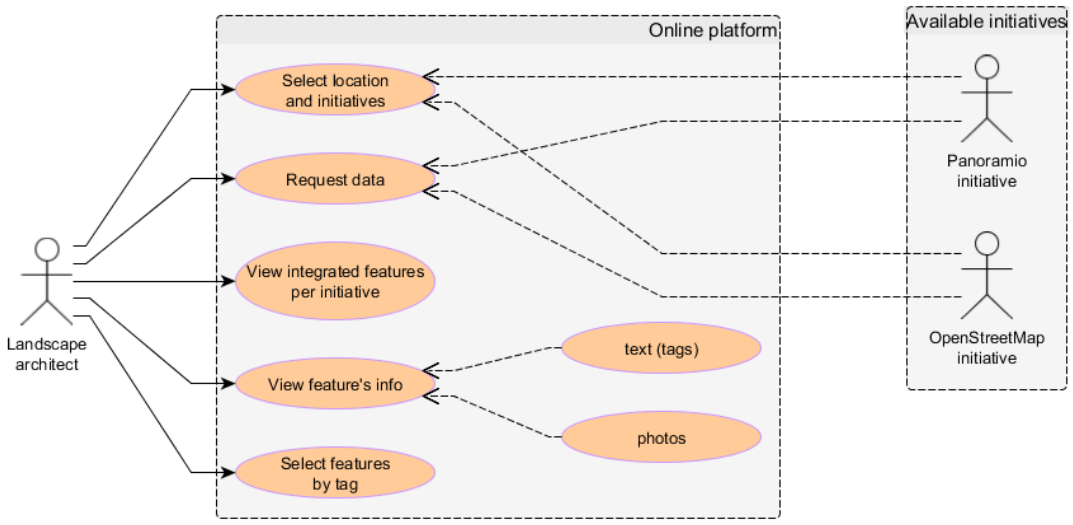

Figure 35 - Landscape architect use case diagram

5.2.4. Programmer use case

Depicted in Figure 36, this use case refers to a programmer that needs to download UGsC data to use for other applications. The programmer accesses the application, defines the input parameters including the location to observe and the initiatives to query, followed by the request of available data. He can then select features by tag and finally export either all the features or only the ones that have been selected to use the data externally.

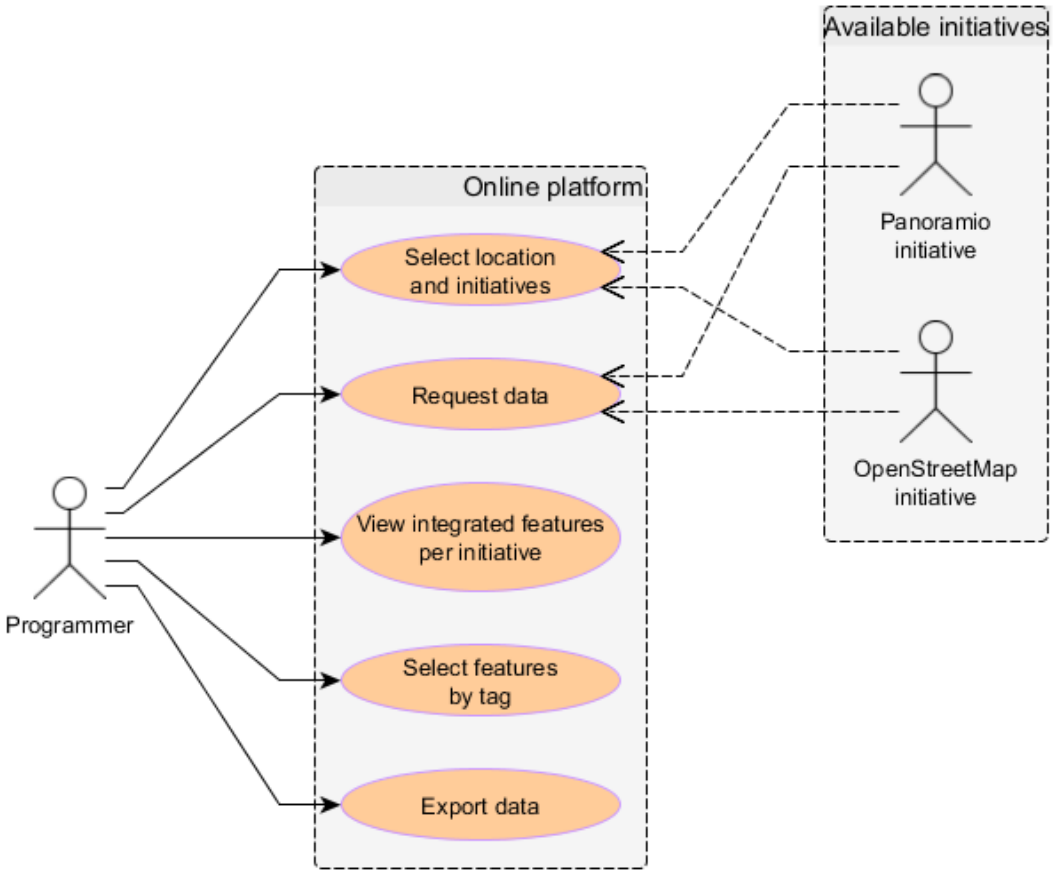

Figure 36 - Programmer use case diagram

5.3. 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 identified initiatives; 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 once the objective is not related with any evaluation of software or benchmark measurement, it was decided to select open source options with the necessary flexibility to implement interactive and user friendly solutions. Thus, the solution required two main structures: 1) a web-based framework and 2) a mapping framework. For the first case the framework Sencha Ext JS, version 4.2.2 (Sencha, 2013) 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 (OpenLayers, 2014).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 (Apache Software Foundation, 2014). This stack responds to all the defined requirements and has been used in several WebGIS implementations (Brovelli, Minghini, & Zamboni, 2014; Burdziej, 2012; Horanont, Basa, & Shibasaki, 2012; Le Cozannet, Bagni, Thierry, Aragno, & Kouokam, 2014; Okladnikov, Gordov, Titov, Bogomolov, & Martynova, 2013; Simeoni, Zatelli, & Floretta, 2014). Figure 37 shows the architecture of the prototype including the selected technologies.

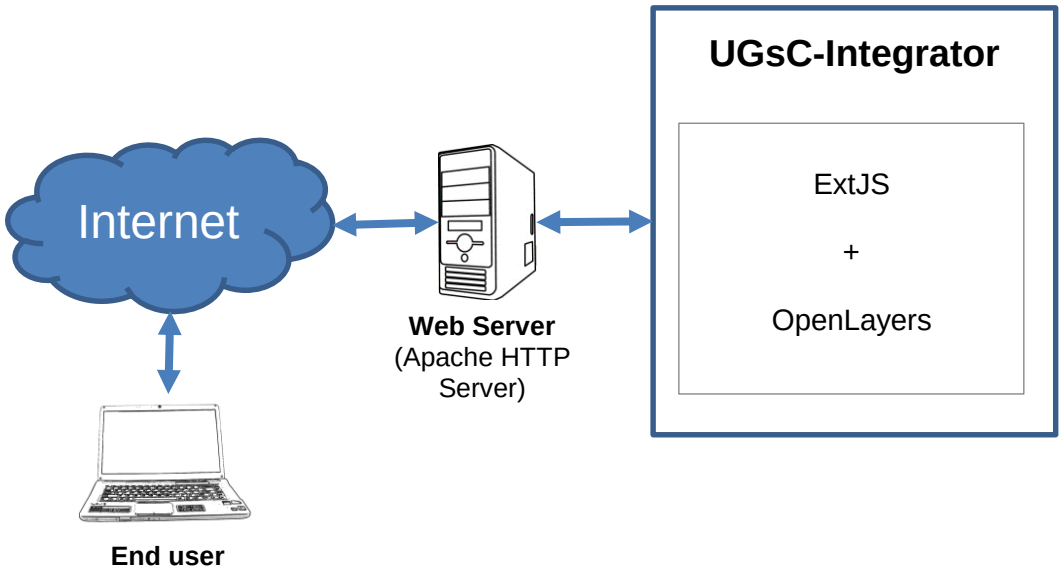

Figure 37 - Prototype architecture

Since the integration model was developed to integrate different data sources with different data types and structures, it was decided to include in the prototype two completely different sources of UGsC, already described in the section 4.2.1: 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 be able to select the location to analyze as well as the input parameters followed by the request and second the resulting data would be displayed in an integrated way allowing a certain level of interaction between the user and the displayed features, such as

feature selection, among others. Consequently, the final layout was divided in two main parts: 1) the initial map and input parameters definition shown in Figure 38; and 2) the features dashboard, depicted in Figure 39.

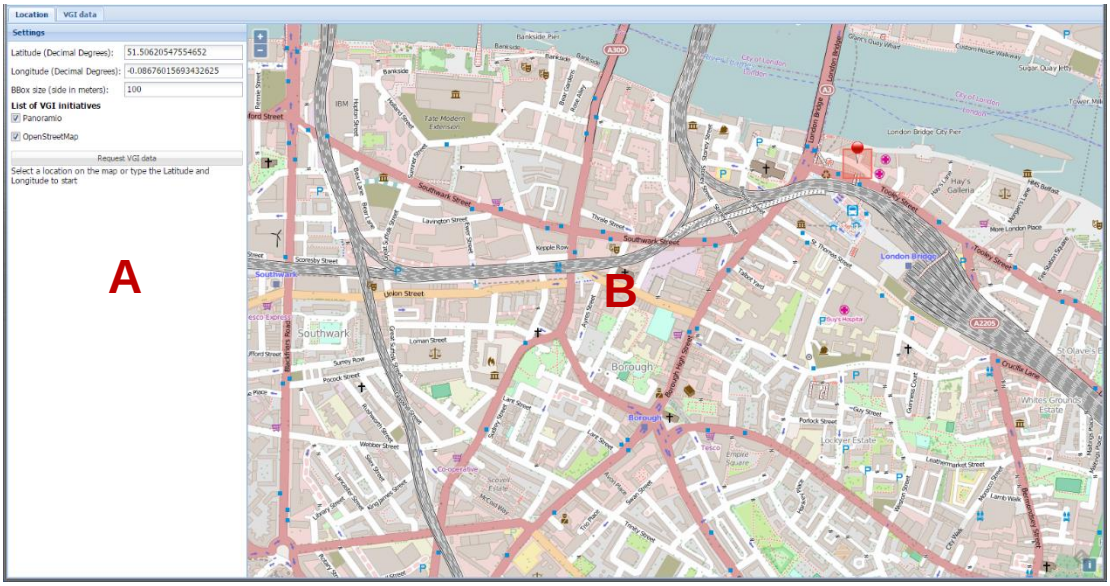

Figure 38 - Final layout (initial map)

Legend:A – Parameters to request data from UGsC initiatives; B – Map to select the location to query

In the initial map (Figure 38), the user is able to select the location to query and define additional input parameters. The selection of the location can be made using two ways: 1) by manually introducing the coordinates in the respective fields of latitude and longitude (B), or 2) by navigating throughout the map (C) and selecting a location that automatically captures the coordinates and fills the respective fields. In both cases, a pin is automatically inserted in the exact selected location on the map. The user has to define also the size of the bounding box used to query the initiatives, by inserting the size of its side in the respective field, and select the initiatives to

query, from the list of available initiatives. The final step is to select the button to start the request. In this phase, the bounding box used to query the initiatives is automatically shown around the selected location on the map.

After getting all the data, the tab to the features dashboard becomes available. In this second part of the final layout (Figure 39), the downloaded features are shown in the central map viewer (E) where they are integrated based on the geospatial attributes of each feature. This map supports the drag and drop of geographic layers, the user can drag, for instance, a polygon to validate and visually understand its limits and the features from the initiatives are spatially related. On the list of layers viewer (D), a list of layers in showed where each layer corresponds to each of the queried initiatives, and the user has the possibility to activate and deactivate each one of them.

The tag statistics viewer (F) displays a chart with the frequency of each tag combining all the selected initiatives, thus giving an initial idea on the most frequent tags. The tags list viewer (G) allows the user to select features by their respective tag. In this viewer it is possible to select one or multiple tags and all the features containing those tags will be selected on the map. The feature info viewer (H) shows all the available metadata for individual features selected on the map. The source code of the prototype is presented in Appendix 3.

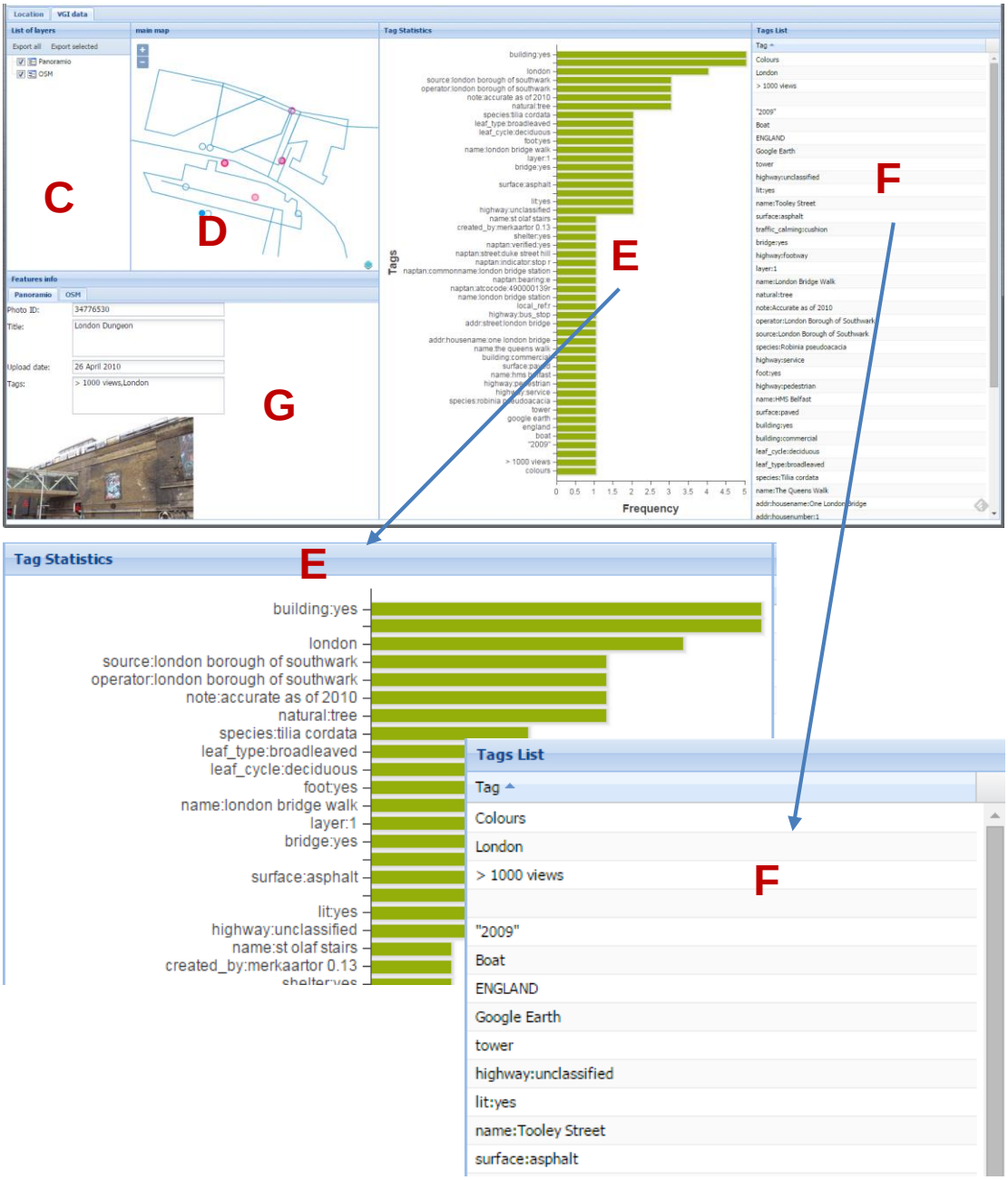

Figure 39 - Final layout (features dashboard)

Legend: C – List of layers; D – map representing data requested from the UGsC initiatives; E – Chart with statistics of tags; F – List of tags to select features on the map; G – Attributes of the selected feature

5.4. Solving the use cases

In this section, the developed prototype is used to demonstrate the ability to solve the defined use cases. For each use case, a step-by-step approach of all the identified activities is followed along with the respective description.

5.4.1. Photo interpretation use case

In this use case, a photo-interpreter accesses the application and uses the map to search and capture the location to clarify. Then he defines the bounding box size by inputting the side length of 200 meters as well as the initiatives to query: the Panoramio initiative in this case, followed by the request of the available data. Figure 40 shows the initial map with the input parameters for this use case.

When the requesting data button is pressed, the following code is fired.

```
Ext.data.JsonP.request({
  async: false,
  url: 'http://www.panoramio.com/map/get_panoramas.php',
  params: {
    set: 'public',
    from: panoramioPhotosFrom,
    to: panoramioPhotosTo,
    minx: minxy[0],
    miny: minxy[1],
    maxx: maxxy[0],
    maxy: maxxy[1]
    },
  success: function(result) {
    //Code to work the result
  },
  failure: function(result) {
       alert('Error requesting metadata from Panoramio initiative');
  }
});
```

Figure 40 - Initial map for the photo interpretation use case

Legend: The pin and square represents respectively the selected location and the boundingbox used in requesting data from the initiatives

This code receives the input parameters from the UI and contacts the Panoramio initiative via the Panoramio public API and requests all publically available data within the defined bounding box. If a success response is obtained, the following code runs inside the success function.

```
if (result.count != 0) {
  for (var i = 0; i < result.photos.length; i++){
    Ext.Ajax.request({
      async: false,
      url:
'http://localhost/phd_thesis/services/panoramiotags.php?photo_url='.concat(result.photos[i].ph
oto_url),
      method: 'POST',
      success: function(response){
        panoramioFeatures.addFeature(new ol.Feature({
          geometry: new ol.geom.Point(ol.proj.transform([result.photos[i].longitude,
result.photos[i].latitude], 'EPSG:4326', 'EPSG:3857')),
          upload_date: result.photos[i].upload_date,
          owner_name: result.photos[i].owner_name,
          photo_id: result.photos[i].photo_id,
          longitude: result.photos[i].longitude,
          latitude: result.photos[i].latitude,
          pheight: result.photos[i].pheight,
          pwidth: result.photos[i].pwidth,
          pheight: result.photos[i].pheight,
          photo_title: result.photos[i].photo_title,
          owner_url: result.photos[i].owner_url,
          owner_id: result.photos[i].owner_id,
          photo_file_url: result.photos[i].photo_file_url,
          photo_url: result.photos[i].photo_url,
          photo_tags: response.responseText,
          vgi_initiative: 'Panoramio'
        }));
        tags = response.responseText.split(",");
        for (var ii = 0; ii < tags.length; ii++) {
          allTags.push(tags[ii]);
          initMapController.tagsListPush(tags[ii]);
        };
      }
    });
  };
};
```

In this piece of code, the response data, composed by a JSON object containing the list of photos retrieved, is parsed and a spatial layer is created based on the latitude and longitude attributes of each photo. In the particular case of the Panoramio initiative, since photos tags are not available through the API, a new request is made to contact the Web page of the photo to extract them from the HTML code. At the

same time, the tags of each photo are added also to an array containing all the available tags to compute statistics and allow the selection of features by tag.

The layer is then added to the map by the following code.

```
vectorPanoramio.setSource(panoramioFeatures);
mainMap.addLayer(vectorPanoramio);
```

After the layer is created and added to the map, the tab for the features dashboard UI becomes available. By activating it, the photo-interpreter accesses to the downloaded data presented in a map as well as additional information such as statistics on tags and information on individual features. This UI is shown in Figure 41.

Each photo is represented by a point in the map and by selecting individual features, the photo-interpreter is able to access the respective photo metadata. This photo metadata includes the photo URL used to display it. The tag statistics chart shows the most frequent tags that might also help to clarify the classification. Additionally the photo interpreter is also able to select features by tag. In this case he would select one or more tags and the respective features would be highlighted in the map. This is useful if the photo interpreter is looking for something in particular, for instance a public building, a park, a forest, etc. The photo-interpreter might use all these available information as ancillary data and take the decision on which class fits better for that location.

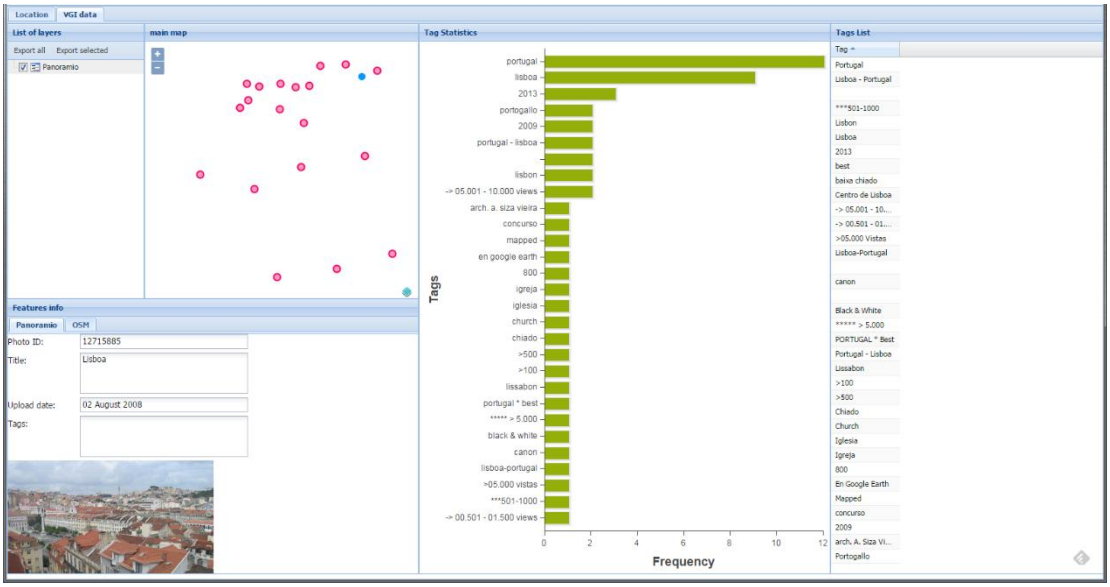

Figure 41 - Features dashboard for the photo interpretation use case

Legend: Red dots represent Panoramio photo locations and the blue highlighted dot represents the selected location

5.4.2. Cartography validation use case

In this use case, a cartography validator access the application and selects the location to validate using one of three ways: 1) by inputting the latitude and longitude of the location in the respective fields; 2) by searching on the map using the available zoom and pan tools and clicking the location; or 3) by dragging a KML file containing the location to use it as a reference and clicking on that location on the map. Then he defines the bounding box size by inputting the side length of 100 meters as well as the initiatives to query: Panoramio and OSM in this case, and click on the request data button. Figure 42 shows the initial map with the input parameters for this use case, as well as the dropped pin for location reference.

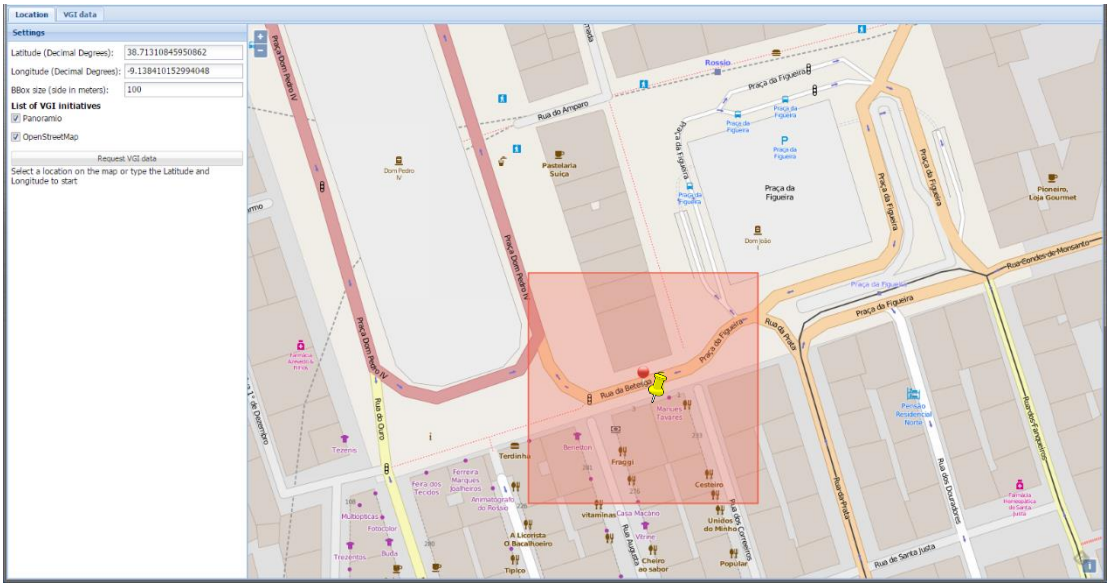

Figure 42 - Initial map for the cartography validation use case

Legend: The yellow and red pin represents respectively the dragged pin and the selected location whereas the red square depicts the boundingbox used in requesting data from the initiatives

When the requesting data button is pressed, the Panoramio and OSM initiatives are contacted. Since requesting data to the panoramio initiative was already described in the previous section, the code to request data from the OSM initiative is shown below.

```
var vectorSource = new ol.source.ServerVector({
  format: new ol.format.OSMXML(),
  loader: function(extent, resolution, projection) {
    var epsg4326Extent = ol.proj.transformExtent(extent, projection, 'EPSG:4326');
    var url = 'http://overpass-api.de/api/xapi?map?bbox=' + epsg4326Extent.join(',');
    Ext.Ajax.request({
      url: url,
      method: 'POST',
      success: function(response){
        vectorSource.addFeatures(vectorSource.readFeatures(response.responseXML));
        vectorSource.forEachFeature( function(feature) {
          for (key in feature.getProperties()) {
            if (key != 'geometry') {
              tag = key + ":" + feature.get(key);
              allTags.push(tag);
              initMapController.tagsListPush(tag);
            };
          };
      });
    });
  },
  strategy: function(){
    return [bbox3857];
  },
  projection: 'EPSG:3857'
});

var vector = new ol.layer.Vector({
  name: 'OSM',
  source: vectorSource,
});

mainMap.addLayer(vector);
```

In the same way as requesting data from Panoramio, the OSM initiative is contacted through its public API and requests all publically available data within the defined boundingbox. If a success response is achieved, the following, the function inside the success parameter is fired. In this case, OpenLayers already contains a function to read and parse OSM features and add them to a layer. After parsing the features, their tags are requested and added to the list of tags to calculate statistics and add them to the list of tags for further selection.

Figure 43 shows the features dashboard UI with all the features and respective metadata added to the different views. The main map view is now showing features from both initiatives spatially integrated. The statistics chart is displaying the frequency of each tag and the list of tags allows the selection, including multiple selection, of features by tag, both views integrating both initiatives.

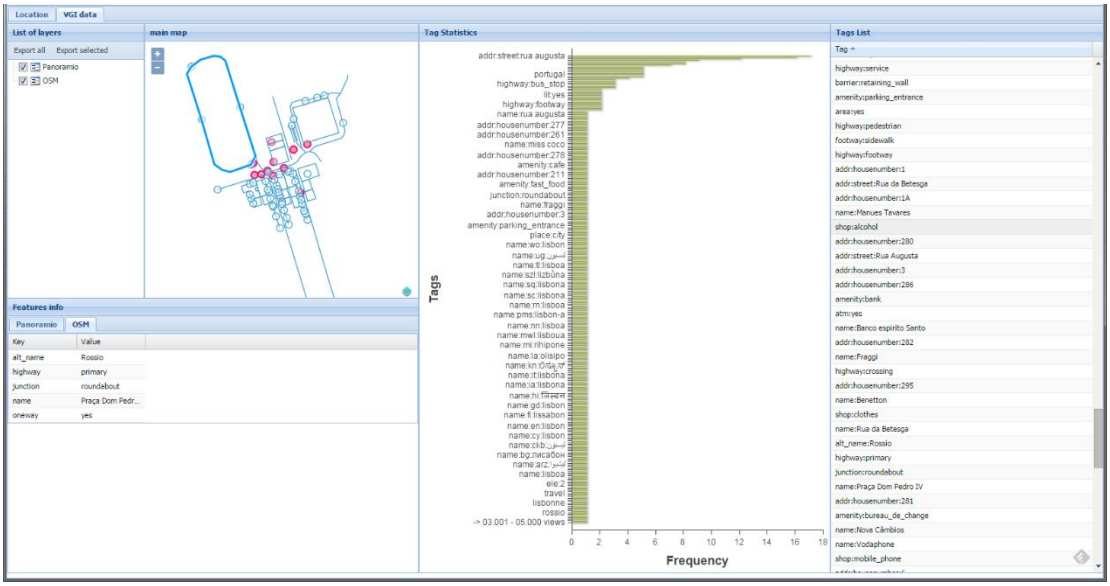

Figure 43 - Features dashboard for the cartography validation use case

Legend: The light blue features depicts OSM features, the red features represents Panoramio photos and the highlighted blue features represents the OSM selected feature

Figure 44 presents features info view showing the attributes of the selected OSM feature. Figure 46 shows the tag statistics view where it is easy to realize 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.

Figure 44 - Detail of the Features info view for an OSM selected feature

Figure 45 details the selection of features by tag. In this case, a multiple selection was made on the tag list view and all the features containing at least one of those tags were automatically selected on the main map view.

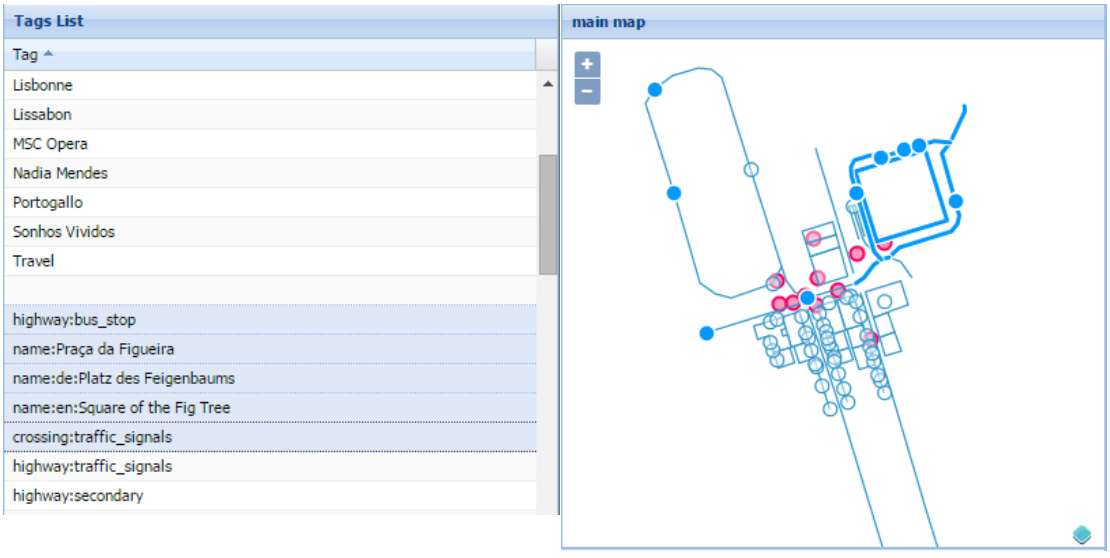

Figure 45 - Selecting features by tag with multiple tags selected

Legend: The light blue features depicts OSM features, the red features represents Panoramio photos and the highlighted blue features symbolizes features that have been selected

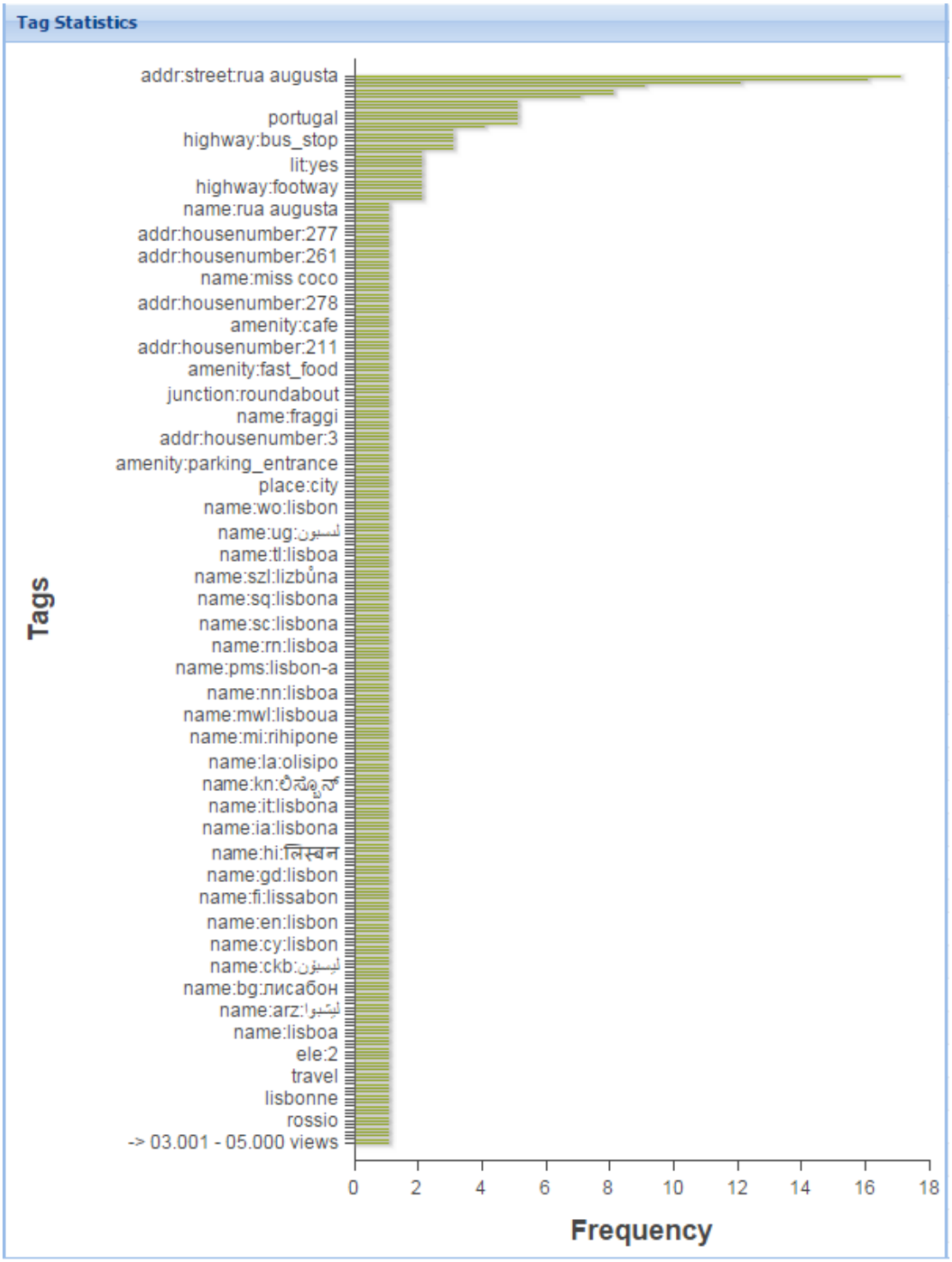

Figure 46 - Detail of the Tag statistics view

It is also possible to drop a polygon into the main map view. This is particularly useful in this use case since LULC cartography is usually constituted by classified areas, or polygons, and gives the validator the ability to overlay the polygon containing the location to validate. Figure 47 depicts such feature by showing the polygon overlaying the other features

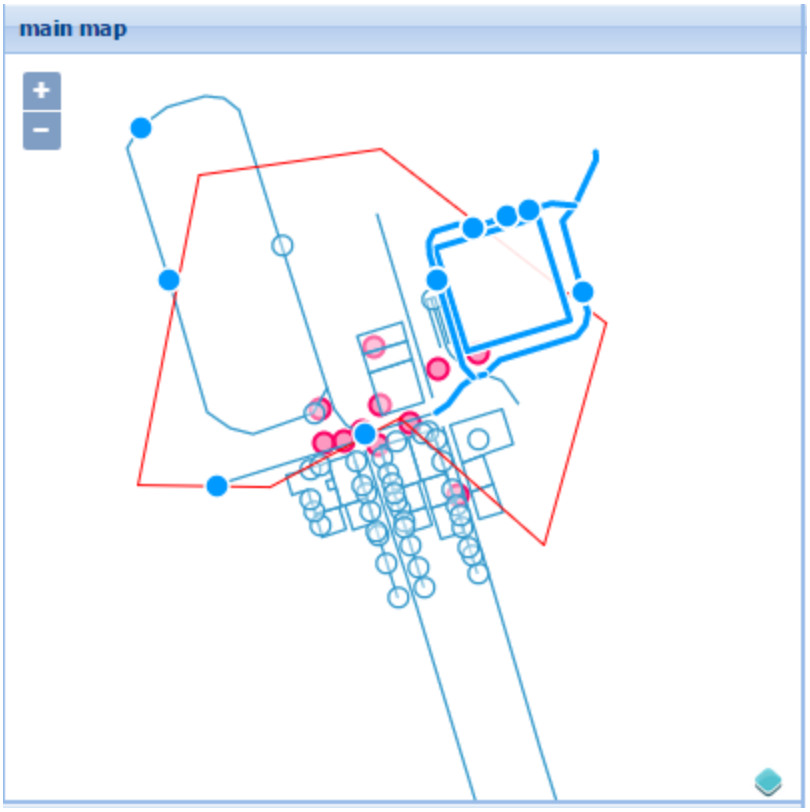

Figure 47 - Main map view with a dropped overlaying polygon

Legend: The light blue features depicts OSM features, the red circles represents Panoramio photos, the highlighted blue features represents the features that have been selected and the red feature depicts the dragged polygon

Finally, the validator can export either all the features present in the map or only the features that have been selected, for further analyze them in a desktop software by

using the appropriate buttons on the list of layers view. In any of these cases, the features being exported are converted to a GeoJSON format and automatically downloaded. The downloaded file can then be opened in any desktop GIS software that supports this format, e.g. QGIS.

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

5.4.3. Landscape architecture use case

In this use case, a landscape architect is interested to gather all the available photos for a certain location, to get a sense on the surroundings. He accesses the application, selects the location to analyze, defines the bounding box size by inputting the side length of 100 meters, selects the initiatives to query: Panoramio in this case, and clicks on the request data button. Figure 48 shows the initial map with the input parameters for this use case.

Figure 49 shows the features dashboard for this use case, where all the available photos from the Panoramio initiative are represented in a map. The architect can select the features one by one and analyze the photos surrounding the selected location on the features info view. By right clicking on a photo it is possible to save it for further inspection and analysis.

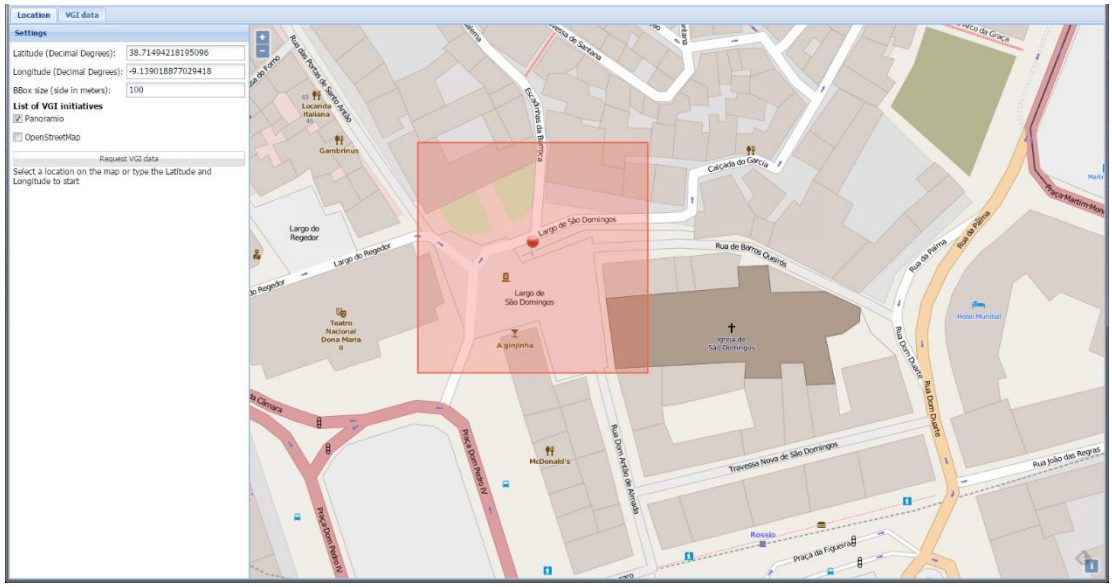

Figure 48 - Initial map for the landscape architect use case

Legend: The red pin represents the selected location and the red square depicts the boundingbox used in requesting data from the initiatives

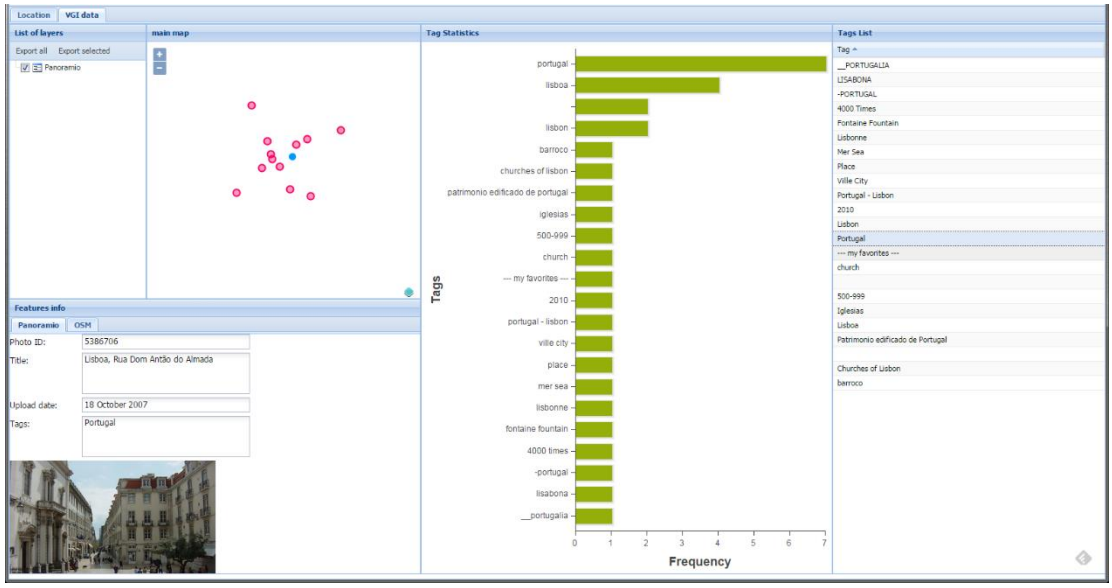

Figure 49 - Features dashboard for the landscape architect use case

Legend: The red circles represent the photo locations from the Panoramio initiative

By looking at photos surrounding the vicinity of the study area, the landscape architect can get an idea on how the area looks like and what kind of architecture exists in the field or the type of trees throughout the streets. If he is interested in a particular feature he can search for related tags available in the list of tags, select those features by tag and quickly look into them.

5.4.4. Programmer use case

In this case, the programmer needs to gather and export all the available data near a certain location to use it in an external application, for example to apply a machine learning approach and identify possible patterns. He accesses the application, selects the location to gather the data, defines the bounding box size by inputting the side length of 100 meters, selects all the available initiatives to query, and clicks on the request data button. Figure 50 shows the initial map with the input parameters for this use case.

This is a particular case where the programmer do not analyze any data on screen and only wants to gather the data and export everything to use the data externally. Figure 51 shows the features dashboard where the programmer only needs to click the export all button to download all the available data.

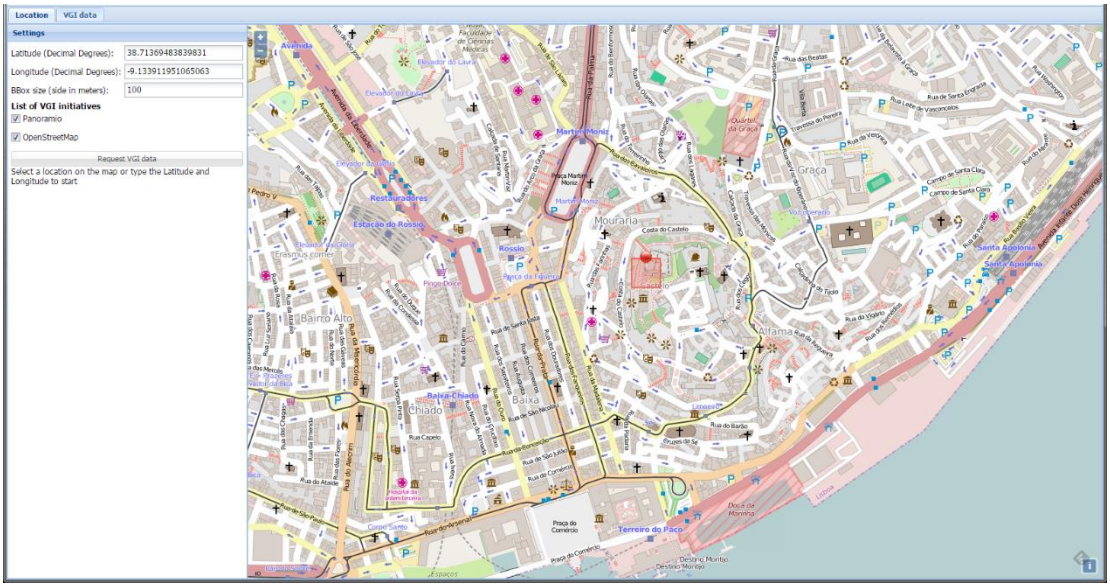

Figure 50 - Initial map for the programmer use case

Legend: The pin and square represents respectively the selected location and the boundingbox used in requesting data from the initiatives

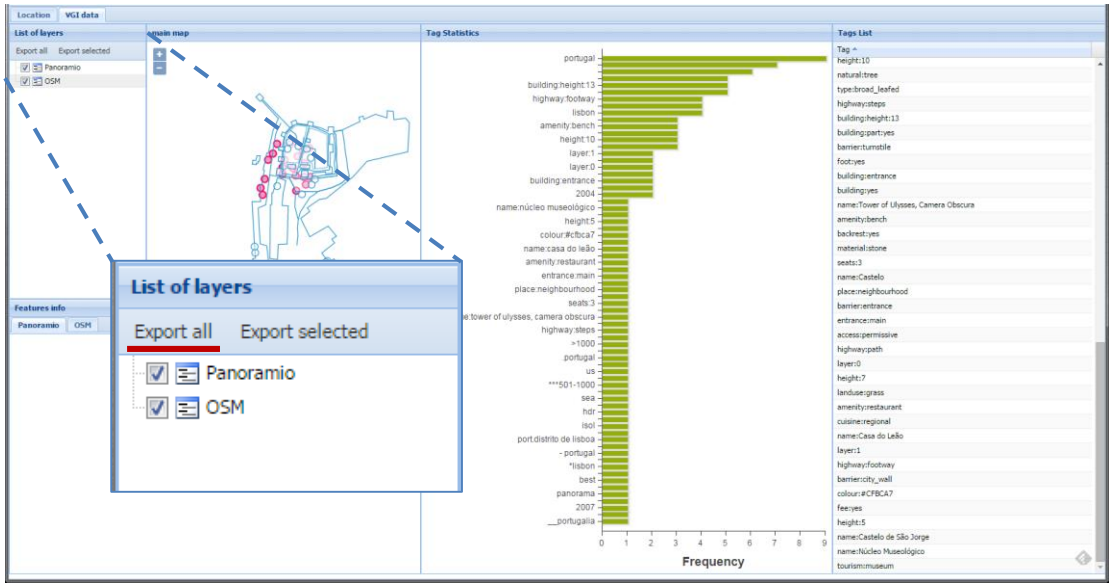

Figure 51 - Features dashboard for the programmer use case

Legend: Highlighted is the "Export all" tool to export all available features

5.5. Conclusion

In this chapter the prototype was described and tested against the set of defined use cases. It was possible to verify the integration of data coming from different sources with different structures 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.

Therefore the implementation of the model has been demonstrated and its validity verified in the perspective of four different users: a photo-interpreter, a cartography validator, a landscape architect and a programmer. For each user, all the defined cases have been successfully solved using the developed prototype. Similar users such as an urban planner, an archeologist, among others, with similar functionality needs can be identified, making the model useful for a large number of applications.

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.

6. Conclusion, contributions and future directions

6.1. Summary

The amount of Geographic Information (GI) created and shared by citizens through the Web during the last decade has increased exponentially. The amount of data associated with the local knowledge of their contributions represents a unique opportunity to explore and extract meaningful information to be used for other purposes and applications, such as LULC databases production. The main challenge

is related with their different nature and heterogeneity, coming from different projects with completely different aims and data structures.

With this in mind, in this study we aimed at analyzing the viability of using UGsC for validation of LULC databases and proposing a data model that integrates data produced by citizens from different sources with different formats for the purpose of using it in the production of LULC databases.

We started by presenting a review of what has been reported in the literature in terms of using UGSC for different applications emphasizing LULC databases production and validation. We reviewed also the concept of data integration and related topics such as interoperability, distributed GIS, data harmonization, conflation and fusion.

We explored the potential of Panoramio, Flickr and OpenStreetMap initiatives in different studies concluding that they have high potential for LULC databases production and validation if they are used together in an integrated way.

We analyzed the characteristics of different UGsC initiatives to collect their dissimilarities and investigate ways to integrate them. Based on these results we proposed a data integration model designed in a scalable way to allow the easy integration of new sources in the future.

We developed a prototype to assess and validate the proposed model. The system requirements were determined based on the development of four different use cases: 1) a photo-interpreter who would use the application to clarify the

classification of certain dubious places; 2) a cartography validator who would use the application to help in the validation process of produced cartography; 3) a landscape architect who would use the application mainly to look at pictures around a specific location to get a sense on the surroundings; and 4) a programmer who would use the application to download the data available at a certain location and use it for other related purposes. Finally the model has been validated by using the prototype to solve these use cases.

6.2. Discussion of Hypotheses

This study was conducted with the following hypotheses in mind:

1. Are the data from User Generated spatial Content (UGsC) initiatives feasible to help in the LULC databases production?
2. Which types of geographic data produced by citizens are more suitable to use in the production of LULC databases?
3. Is it possible to integrate them in a common data model/platform?

Different studies were conducted to explore the potential of using data from different UGSC initiatives, such as Panoramio, Flickr and OpenStreetMap, for LULC databases production and validation. The main advantages found are related with the amount of available data and their temporal distribution. On the other side, the spatial distribution has proved very uneven, more concentrated in urban areas and touristic places. We concluded that they have great potential and viability if they are used together, in an integrated way.

As different UGSC initiatives might have totally different aims, their type of data can also be of completely different type. We explored different types of data coming from a comprehensive list of UGSC initiatives and defined a set of minimum requirements based on their spatial context, spatial phenomena, data type, access type, data license, and coverage, which any initiative must have to be used in the production of LULC databases. Any initiative compliant with these minimum requirements can be used for this purpose and integrated in the model.

Based on the geographic characteristic of data from different UGSC initiatives it is possible to integrate them in a common model/platform. In this regard, and using the USGS initiatives compliant with the minimum requirements defined, we proposed a model to integrate them. We validated the model by developing a prototype and solving four pre-defined use cases, thus proving this hypothesis.

6.3. Main contributions to the scientific community

From this thesis, we would like to highlight the following contributions:

1. We explored the Panoramio, Flickr and OpenStreetMap initiatives and proved the feasibility of UGSC initiatives for LULC databases production, especially if they are used together in an integrated way.
2. We explored the characteristics of a comprehensive list of UGSC initiatives and proposed a set of minimum requirements any initiative needed to be used for LULC databases production. In this sense, any future initiative compliant with these requirements can be used and integrated.

3. We analyzed the list of compliant UGsC initiatives and proposed a data model that integrates them in a scalable way so any future initiative compliant with the minimum requirements defined can be added and integrated.
4. We defined four use cases to determine the system requirements for a platform to implements the proposed model. Based on that, a prototype was developed and used to validate the model by using it to solve the defined use cases.

6.4. Limitations

Most of the public API's of the UGsC initiatives have limitations in terms of number of requests a user can make, or the quantity of data that can be downloaded within a certain time interval. This represents a limitation to use the prototype for larger areas or with very high frequency.

Given the nature of the data from different UGsC initiatives, the UGsC-Integrator and prototype proposed in this study have some limitations that should be drawn. From exploring the viability of Panoramio, Flickr and OpenStreetMap for LULC databases production, the spatial distribution of data as well as the distribution over LULC classes was found uneven. Although the integration of different initiatives aimed to contribute to tackle such inequalities, there will always be a certain level of disparity.

Another important limitation is related with 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 for their integration. Tags are related with the language, the region or even the user environment. To overcome such 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 volunteers is much related with the introduction of the Web 2.0, the increasing availability of positioning equipment's 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 less available data, or even no data, for these locations. This phenomenon is identified as the Digital Divide (Sui et al., 2013) and represents a major limitation of the UGsC-Integrator and prototype for locations where such technologies are not used and therefore data is scarce or nonexistent.

6.5. Future work

The model proposed here is not finished and future developments and improvements can be expected.

The prototype was developed based on four use cases but more use cases can be defined to increase its comprehensiveness. We foresee also the integration of more and different UGsC initiatives to increase the reliability 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 in the platform, but further investigation is needed to determine their advantages.

Future research will also focus on improving the level of analytic tools available on the platform. 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.

Finally the development of a web service is planned. The main advantage would be related with the possibility of using the data resulting from the UGsC-Integrator directly in different and independent applications.

References

Adam, A., & Muraki, Y. (2011). Twitter for crisis communication : lessons learned from Japan ' s tsunami disaster. *International Journal of Web Based Communities*, *7*(3), 392–402.

Apache Software Foundation. (2014). Apache HTTP Server Project. Retrieved from http://httpd.apache.org/

Arsanjani, J. J., Helbich, M., & Bakillah, M. (2013). Exploiting volunteered geographic information to ease land use mapping of an urban landscape. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (Vol. XL). London, United Kingdom.

Arsanjani, J. J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, (September 2013), 1–15. doi:10.1080/13658816.2013.800871

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, n/a–n/a. doi:10.1111/tgis.12073

Bartholomeus, R. P., Witte, J.-P. M., van Bodegom, P. M., & Aerts, R. (2008). The need of data harmonization to derive robust empirical relationships between soil conditions and vegetation. *Journal of Vegetation Science*, *19*(6), 799–808. doi:10.3170/2008-8-18450

Batty, M., Hudson-Smith, A., Milton, R., & Crooks, A. (2010). Map mashups, Web 2.0 and the GIS revolution. *Annals of GIS*, *16*(1), 1–13. doi:10.1080/19475681003700831

Bishr, Y. (1997). *Semantic aspects of interoperable GIS*. Enschede: ITC.

Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, *12*(4), 299–314. doi:10.1080/136588198241806

Brodeur, J., Bedard, Y., Edwards, G., & Moulin, B. (2003). Revisiting the Concept of Geospatial Data Interoperability within the Scope of Human Communication Processes. *Transactions in GIS*, *7*(2), 243–265. doi:10.1111/1467-9671.00143

Brovelli, M. A., Minghini, M., & Zamboni, G. (2014). Public Participation GIS: a FOSS architecture enabling field-data collection. *International Journal of Digital Earth*, (July), 1–19. doi:10.1080/17538947.2014.887150

Burdziej, J. (2012). A Web-based spatial decision support system for accessibility analysis-concepts and methods. *Applied Geomatics*, *4*(2), 75–84. doi:10.1007/s12518-011-0057-x

Butenuth, M., Gösseln, G. V., Tiedge, M., Heipke, C., Lipeck, U., & Sester, M. (2007). Integration of heterogeneous geospatial data in a federated database. *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(5), 328–346. doi:10.1016/j.isprsjprs.2007.04.003

Büttner, G., Kosztra, B., Maucha, G., & Pataki, R. (2012). *Implementation and achievements of CLC2006*. Retrieved from http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2/#tab-documents

Caetano, M., Mata, F., & Freire, S. (2006). Accuracy assessment of the Portuguese CORINE Land Cover map. *Global Developments in Environmental Earth Observation from Space*, 459–467.

Chang, Y., & Park, H. (2006). XML Web Service-based development model for Internet GIS applications. *International Journal of Geographical Information Science*, *20*(4), 371–399. doi:10.1080/13658810600607857

Cihlar, J. (2000). Land cover mapping of large areas from satellites: Status and research priorities. *International Journal of Remote Sensing*, *21*(6-7), 1093–1114. doi:10.1080/014311600210092

Cihlar, J., & Jansen, L. (2001). From Land Cover to Land Use: A Methodology for Efficient Land Use Mapping over Large Areas. *The Professional Geographer*, *53*(2), 275–289. doi:10.1111/0033-0124.00285

Clark, M. L., & Aide, T. M. (2011). Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) for Collecting Land-Use/Land-Cover Reference Data. *Remote Sensing*, *3*(3), 601–620. doi:10.3390/rs3030601

Cobb, M. A., Chung, M. J., III, H. F., Petry, F. E., Shaw, K. B., & Miller, H. V. (1998). A Rule-based Approach for the Conflation of Attributed Vector Data. *GeoInformatica*, *2*(1), 7–35. doi:10.1023/A:1009788905049

Cowen, D. J. (2007). Why not a Geo-Wiki Corps. In *Workshop on Volunteered Geographic Information*. Santa Barbara, CA. Retrieved from http://ncgia.ucsb.edu/projects/vgi/participants.html

Elwood, S. (2008a). Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, *72*(3-4), 173–183. doi:10.1007/s10708-008-9186-0

Elwood, S. (2008b). Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. *GeoJournal*, *72*(3-4), 133–135. doi:10.1007/s10708-008-9187-z

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. doi:10.1080/00045608.2011.595657

Estima, J., Fonte, C. C., & Painho, M. (2014). Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database. In Huerta, Schade, & Granell (Eds.), *Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science.* Castellón, Spain.

Estima, J., & Painho, M. (2013a). Exploratory analysis of OpenStreetMap for land use classification. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '13* (pp. 39–46). New York, New York, USA: ACM Press. doi:10.1145/2534732.2534734

Estima, J., & Painho, M. (2013b). Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. In B. Murgante, S. Misra, M. Carlini, C. M. Torre, H.-Q. Nguyen, D. Taniar, … O. Gervasi (Eds.), *Computational Science and Its Applications – ICCSA 2013* (Vol. 7974, pp. 205–220). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-39649-6_15

Estima, J., & Painho, M. (2014). Photo Based Volunteered Geographic Information Initiatives: *International Journal of Agricultural and Environmental Information Systems*, *5*(3), 73–89. doi:10.4018/ijaeis.2014070105

Estima, J., & Painho, M. (2015). Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal. In J. Jokar Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience: experiences, research, applications* (pp. 273–293). doi:10.1007/978-3-319-14280-7

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. doi:10.1080/13658816.2013.867495

Fielding, T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. University of California. Retrieved from http://jpkc.fudan.edu.cn/picture/article/216/35/4b/22598d594e3d93239700ce79bce1/7ed3ec2a-03c2-49cb-8bf8-5a90ea42f523.pdf

Fischer, F. (2012). VGI as Big Data: A New but Delicate Geographic Data-Source. *Geoinformatics April/May*, (May), 46–47.

Fonte, C. C., Bastin, L., See, L., Foody, G., & Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*, (April), 1–23. doi:10.1080/13658816.2015.1018266

Foody, G. M. (2010). Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment*, *114*(10), 2271–2285. doi:10.1016/j.rse.2010.05.003

Foody, G. M., & Boyd, D. S. (2013a). Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(3), 1305–1312. doi:10.1109/JSTARS.2013.2250257

Foody, G. M., & Boyd, D. S. (2013b). Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(3), 1305–1312. doi:10.1109/JSTARS.2013.2250257

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., … Obersteiner, M. (2009). Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing*, *1*(3), 345–354. doi:10.3390/rs1030345

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., … Obersteiner, M. (2012). Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, *31*, 110–123. doi:10.1016/j.envsoft.2011.11.015

Funayama, T., Yamamoto, Y., Tomita, M., Uchida, O., & Kajita, Y. (2014). Disaster mitigation support system using Twitter and GIS. In *2014 Twelfth International Conference on ICT and Knowledge Engineering* (pp. 18–23). doi:10.1109/ICTKE.2014.7001528

Geograph. (2012). Geograph Website. Retrieved January 28, 2012, from http://www.geograph.org.uk/

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*(1), 115–125. doi:10.1016/j.dss.2014.02.003

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. doi:10.1007/s10708-007-9111-y

Goodchild, M. (2008). Commentary: whither VGI? *GeoJournal*, *72*(3-4), 239–244. doi:10.1007/s10708-008-9190-4

Goodchild, M., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, *3*(3), 231–241. doi:10.1080/17538941003759255

Gösseln, G., & Sester, M. (2004). Integration of geoscientific data sets and the german digital map using a matching approach. In *XXth ISPRS Congress* (Vol. 35, pp. 1249–1254). Istambul, Turkey.

Hagenauer, J., & Helbich, M. (2012). Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science*, *26*(6), 963–982. doi:10.1080/13658816.2011.619501

Hecht, R., Kunze, C., & Hahmann, S. (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS International Journal of Geo-Information*, *2*(4), 1066–1091. doi:10.3390/ijgi2041066

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(6), 550–557. doi:10.1016/j.isprsjprs.2010.06.005

Herold, M. (2009). Assessment of the status of the development of the standards for the Terrestrial Essential Climate Variables. *Technical report*. Rome: Technical report. Retrieved from http://www.fao.org/gtos/doc/ECVs/T09/T09.pdf

Herold, M., Woodcock, C. E., Mayaux, P., Belward, A. S., Latham, J., & Schmullius, C. C. (2006). A joint initiative for harmonization and validation of land cover datasets. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(7), 1719–1727. doi:10.1109/TGRS.2006.871219

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, *1*(1), 21–48. doi:10.5311/JOSIS.2010.1.3

Horanont, T., Basa, M., & Shibasaki, R. (2012). Towards thematic Web services for generic data visualization amd analysis. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume I-4*.

Hudson-Smith, A., Batty, M., Crooks, A., & Milton, R. (2009). Mapping for the Masses: Accessing Web 2.0 Through Crowdsourcing. *Social Science Computer Review*, *27*(4), 524–538. doi:10.1177/0894439309332299

Hudson-Smith, A., Crooks, A., Gibin, M., Milton, R., & Batty, M. (2009). NeoGeography and Web 2.0: concepts, tools and applications. *Journal of Location Based Services*, *3*(2), 118–145. doi:10.1080/17489720902950366

Hull, R., & Zhou, G. (1996). 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 - SIGMOD '96 (pp. 481–492). New York, New York, USA: ACM Press. doi:10.1145/233269.233365

Jokar Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, *27*(12), 2264–2278. doi:10.1080/13658816.2013.800871

Jokar Arsanjani, J., & Vaz, E. (2015). An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *International Journal of Applied Earth Observation and Geoinformation*, *35*, 329–337. doi:10.1016/j.jag.2014.09.009

Keune, H., Murray, A. B., & Benking, H. (1991). Harmonization of environmental measurement. *GeoJournal*, *23*(3), 249–255. doi:10.1007/BF00204842

Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., & Andrienko, G. (2010). Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. In *2010 14th International Conference Information Visualisation* (pp. 289–296). IEEE. doi:10.1109/IV.2010.94

Kottman, C. A. (1999). The Open GIS Consortium and Progress Toward Interoperability in Gis. In M. Goodchild, M. Egenhofer, R. Fegeas, & C. Kottman (Eds.), *Interoperating Geographic Information Systems* (pp. 39–54). Boston, MA: Springer US. doi:10.1007/978-1-4615-5189-8

Kuhn, W. (2007). Volunteered Geographic Information and GIScience. In *Workshop on Volunteered Geographic Information*. Santa Barbara, CA. Retrieved from http://ncgia.ucsb.edu/projects/vgi/participants.html

Laurini, R. (1998). Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. *International Journal of Geographical Information Science*, *12*(4), 373–402. doi:10.1080/136588198241842

Le Cozannet, G., Bagni, M., Thierry, P., Aragno, C., & Kouokam, E. (2014). WebGIS as boundary tools between scientific geoinformation and disaster risk reduction action in volcanic areas. *Natural Hazards and Earth System Sciences*, *14*(6), 1591–1598. doi:10.5194/nhess-14-1591-2014

Leung, D., & Newsam, S. (2010). Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2955–2962). IEEE. doi:10.1109/CVPR.2010.5540040

Mazzetti, P., Nativi, S., & Caron, J. (2009). RESTful implementation of geospatial services for Earth and Space Science applications. *International Journal of Digital Earth*, *2*(sup1), 40–61. doi:10.1080/17538940902866153

Mills, A., & Chen, R. (2009). Web 2.0 emergency applications: how useful can Twitter be for emergency response? *Journal of Information Privacy & Security*, *5*(3), 3. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=15536548&AN=47618127&h=kCXl1Anz8JPQpzMpYY5wn3f5TwdfQYE4/1ayM8jrXxYmhE2kqfR2zNOX//Jmyu/NrCqIB+CZuBD1TJz1Z57xGQ==&crl=c

Mohammadi, H., Rajabifard, A., & Williamson, I. P. (2010). Development of an interoperable tool to facilitate spatial data integration in the context of SDI. *International Journal of Geographical Information Science*, *24*(4), 487–505. doi:10.1080/13658810902881903

Mooney, P., Corcoran, P., & Winstanley, A. (2010). A study of data representation of natural features in OpenStreetMap. In *Proceedings of the 6th GIScience International Conference on Geographic Information Science*. Zurich, Switzerland.

Neill, C. J., & Laplante, P. a. (2003). Requirements engineering: the state of the practice. *IEEE Software*, *20*(6). doi:10.1109/MS.2003.1241365

OGC. (2014a). OGC Standards. *Open Geospatial Consortium*. Retrieved from http://www.opengeospatial.org/standards

OGC. (2014b). Open GeoSpatial Consortium Web Site. Retrieved from http://www.opengeospatial.org/

Okladnikov, I., Gordov, E., Titov, A., Bogomolov, V., & Martynova, Y. (2013). Application of web-GIS approach for climate change study. *EGU General Assembly*, *15*, 6682.

Olteanu, A., Mustière, S., & Ruas, A. (2006). Matching imperfect spatial data. In M. Caetano & M. Painho (Eds.), *7th international symposium on spatial accuracy assessment in natural resources and environmental sciences* (pp. 694–704). Lisbon. Retrieved from http://www.spatial-accuracy.org/system/files/Olteanu2006accuracy.pdf

OpenLayers. (2014). OpenLayers (version 3.1.1). Retrieved from https://github.com/openlayers/ol3/releases

OpenStreetMap. (2014). OpenStreetMap Map Features. Retrieved from http://wiki.openstreetmap.org/wiki/Map_Features

Papazoglou, M. P., Traverso, P., Dustdar, S., & Leymann, F. (2008). Service-Oriented Computing: a Research Roadmap. *International Journal of Cooperative Information Systems*, *17*(02), 223–255. doi:10.1142/S0218843008001816

Peng, Z.-R., & Tsou, M.-H. (2003). *Internet GIS: distributed geographic information services for the internet and wireless networks*. John Wiley & Sons, Inc.

Perger, C., Fritz, S., See, L., Schill, C., Van Der Velde, M., Mccallum, I., & Obersteiner, M. (2012). A Campaign to Collect Volunteered Geographic Information on Land Cover and Human Impact. In *GI Forum 2012: Geovizualisation, Society and Learning* (pp. 83–91). Herbert Wichmann Verlag.

Pultar, E., Raubal, M., Cova, T. J., & Goodchild, M. F. (2009). Dynamic GIS Case Studies: Wildfire Evacuation and Volunteered Geographic Information. *Transactions in GIS*, *13*, 85–104. doi:10.1111/j.1467-9671.2009.01157.x

Reips, U.-D., & Garaizar, P. (2011). Mining twitter: A source for psychological wisdom of the crowds. *Behavior Research Methods*, *43*(3), 635–642. doi:10.3758/s13428-011-0116-6

Ruiz, J. J., Ariza, F. J., Ureña, M. a., & Blázquez, E. B. (2011). Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science*, *25*(9), 1439–1466. doi:10.1080/13658816.2010.519707

Saalfeld, A. (1988). Conflation Automated map compilation. *International Journal of Geographical Information Systems*, *2*(3), 217–228. doi:10.1080/02693798808927897

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, 851. doi:10.1145/1772690.1772777

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., … Obersteiner, M. (2013). Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PLoS ONE*, *8*(7), e69958. doi:10.1371/journal.pone.0069958

See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A., … Obersteiner, M. (2014). Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, *103*, 48–56. doi:10.1016/j.isprsjprs.2014.06.016

Sencha. (2013). Sencha Ext JS (version 4.2.2). Retrieved from http://www.sencha.com/legal/GPL/

Sha, Z., & Xie, Y. (2010). Design of service-oriented architecture for spatial data integration and its application in building web-based GIS systems. *Geo-Spatial Information Science*, *13*(1), 8–15. doi:10.1007/s11806-010-0163-7

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, *6*(5). doi:10.1371/journal.pone.0019467

Simeoni, L., Zatelli, P., & Floretta, C. (2014). Field measurements in river embankments: Validation and management with spatial database and webGIS. *Natural Hazards*, *71*(3), 1453–1473. doi:10.1007/s11069-013-0955-9

Spence, P. R., Lachlan, K. a., Lin, X., & del Greco, M. (2015). Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication. *Communication Quarterly*, *63*(2), 171–186. doi:10.1080/01463373.2015.1012219

Stankut, S., & Asche, H. (2009). An Integrative Approach to Geospatial Data Fusion. In O. Gervasi, D. Taniar, B. Murgante, A. Laganà, Y. Mun, & M. L. Gavrilova (Eds.), *Computational Science and Its Applications – ICCSA 2009* (Vol. 5592, pp. 490–504). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02454-2

Stefanidis, A., Crooks, A., & Radzikowski, J. (2011). Harvesting ambient geospatial information from social media feeds. *GeoJournal*. doi:10.1007/s10708-011-9438-2

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., … Woodcock, C. E. (2006). Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps. In *European Communities*. Luxembourg.

Sui, D. (2007). Volunteered Geographic Information: A tetradic analysis using McLuhan's law of the media. In *Workshop on Volunteered Geographic Information*. Santa Barbara, CA. Retrieved from http://www.ncgia.ucsb.edu/projects/vgi/

Sui, D. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, *32*(1), 1–5. doi:10.1016/j.compenvurbsys.2007.12.001

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748. doi:10.1080/13658816.2011.604636

Sui, D., Goodchild, M., & Elwood, S. (2013). Volunteered Geographic Information, the Exa flood, and the Growing Digital Divide. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge* (pp. 1–12). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-4587-2

Sun, B. (2009). A multi-tier architecture for building RESTful Web services. *ibm.com/developerWorks/*, 1–8. Retrieved from http://www.ibm.com/developerworks/library/wa-aj-multitier/wa-aj-multitier-pdf.pdf

Tait, M. G. (2005). Implementing geoportals: applications of distributed GIS. *Computers, Environment and Urban Systems*, *29*(1), 33–47. doi:10.1016/j.compenvurbsys.2004.05.011

Tsou, M.-H., Yang, J.-A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., … An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, *40*(4), 337–348. doi:10.1080/15230406.2013.799738

Turner, A. J. (2006). *Introduction to Neogeography*. (O'Reilly Media, Ed.). Sebastopol, CA.

Twitter. (2014). Twitter Website. Retrieved January 28, 2014, from https://about.twitter.com

Vckovski, A. (1998). Guest Editorial Special Issue: Interoperability in GIS. *International Journal of Geographical Information Science*, *12*(4), 297–298. doi:10.1080/136588198241798

Vgi-net. (2013). vgi-net: a collaborative research project. Retrieved from http://vgi.spatial.ucsb.edu/

W3C, A. W. S. (2013). Web Services Activity Statement. Retrieved from http://www.w3.org/2002/ws/Activity.html

Wald, L. (1999). Some terms of reference in data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, *37*(3), 1190–1193. doi:10.1109/36.763269

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer*, *25*(3), 38–49. doi:10.1109/2.121508

Wiemann, S., & Bernard, L. (2010). Conflation Services within Spatial Data Infrastructures. In *13th AGILE International Conference on Geographic Information Science 2010* (pp. 1–8). Portugal.

Zielstra, D., & Hochmair, H. H. (2013). Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *Journal of Spatial Science*, *58*(2), 251–273. doi:10.1080/14498596.2013.801331

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, *2*(2), 6–32. doi:10.2202/1948-4682.1069

Appendixes

Appendix 1 – VGI initiatives by Elwood et al.[26]

#	Name	Availability	start date	coverage
1	43 Places	Av.	2006	Local
2	Aha	Av.	2009	Local
3	aka-aki	Not Av.	2007	Local
4	Belysio	Not Av.	2008	Local
5	Bing Maps	Av.	2005	Global
6	Birds and Climate change: building an early warning system	Av.	mid 2000s?	Regional
7	Bliin	Not Av.	2007	Local
8	Blummi	Not Av.	2008	Local
9	Brite Kite	Not Av.	2008	Regional
10	Buddy Cloud	Av.	2008	Local
11	Buddy Way	Av.	2008	Local
12	buzzd	Not Av.	2008	Local
13	Carticipate	Not Av.	?	Local
14	Center'd	Not Av.	2008	Local
15	Centrl	Not Av.	2007	Local

[26] Adapted from http://vgi.spatial.ucsb.edu/inventory (last accessed on September 1, 2013)

16	Christmas Bird Count	Av.	1997	Local
17	Citizen Science Inventory	Av.		Local
18	City sense	Av.	2008	Local
19	Cyclopath	Av.	2003	Local
20	Did you feel it?	Av.	2006	Local
21	DIY City	Av.	2008	Local
22	Dodgeball	Not Av.	2005	Local
23	Endangered Western Leopard Toad	Av.	2007	Local
24	Every Trail	Av.	2006	Local
25	EveryBlock	Not Av.	2008	Local
26	feedmap	Not Av.	2007	Regional
27	Find by click	Not Av.	2007	Local
28	Fire Eagle	Not Av.	2009	Local
29	Flagr	Not Av.	2006	Local
30	flaik	Av.	2005	Local
31	Flickr	Av.	2004	Global
32	Footprint History	Not Av.	2009	Local
33	Four Square	Av.	2009	Local
34	GeoCaching	Av.	2000	Local
35	Geocrowd	Av.	??	Local
36	GeographUK	Av.	2005	Regional
37	GeoMe	Not Av.	2008	Local
38	GeoNames	Av.	2005	Global
39	GeoSpot	Not Av.	2008	Local
40	GLOBE	Av.	1994	Local
41	Glympse	Av.	2009	Local
42	Google Earth	Av.	2005	Global
43	Google FluTrends	Av.	2009?	Local
44	Google Latitude	Av.	2008	Local
45	Google Maps	Av.	2005	Global
46	Groovr	Not Av.	2007	Local
47	GyPSii	Av.	2008	Local
48	HostIP	Av.	2006	Global
49	iFob	Av.	2008	Local
50	In Real Life (IRL)	Not Av.	2008	Local
51	KML Factbook	Av.		Local
52	Limbo	Not Av.	2008	Local
53	Loopt	Not Av.	2006	Regional
54	Map my ride	Av.	2005	Local
55	MapJack	Av.	2007	Local
56	MapQuest	Av.	1997	Regional
57	Mapufacture	Changed to Geocommons	2005	Regional
58	Meet Moi	Av.	2007	Local
59	Monarch Larva Monitoring Project	Av.	1997	Local
60	murmur	Av.	2003	Local
61	NASA World Wind	Av.	2003	Global
62	Nature Mapping	Not Av.	1992	Local
63	North American Bird Phenology Program	Av.	2003	Local
64	OpenAddresses	Not Av.	2007	Global
65	OpenCellID	Av.	2008	Local
66	OpenStreetMap	Av.	2004	Global
67	Ovi Maps	Changed name to Here	2009	Local
68	Panoramio	Av.	2005	Local
69	Platial	Not Av.	2006	Local
70	Plazes	Changed name to Here	2004	Local
71	Project BudBurst	Not Av.	2007	Local

72	Road Watch in the Pass	Av.	2005	Local
73	Roll'n'Zoom	Not Av.	2007	Global
74	Serve.gov	Av.	2008	Local
75	Stimulus Watch	Av.	2008	Regional
76	TomTom MapShare	Av.	2007	Local
77	Trackut	Not Av.	2008	Local
78	Trail Peak	Av.	2001	Local
79	Trapster	Av.	2008	Local
80	Travellr	Av.	2009	Regional
81	Twinkle	Not Av.	2008	Local
82	UMapper	Av.	2009	Local
83	UpNext	Not Av.	2008	Local
84	Urbantastic	Av.	2009	Local
85	US Fish Finder	Av.	2007	Regional
86	USGS National Map Corps	Av.	2001	Regional
87	Ushahidi	Av.	2008	Local
88	Waze	Av.	2008	Local
89	WHERE	Not Av.	2007	Local
90	Whereboutz	Not Av.	2008?	Local
91	Whrrl	Not Av.	2008	Local
92	WiGLE	Av.	2001	Local
93	Wikimapia	Av.	2006	Global
94	Wikipedia	Av.	2001	Global
95	Wild Style City	Not Av.	2009	Local
96	World Heritage Site	Av.	2001	Local
97	Yahoo! Maps	Av.	1997	Global
98	Yahoo! Placemaker	Not Av.	2009	Global
99	Zhiing	Not Av.	2007	Local
100	Zoom and Go	Av.	2002	Local

Appendix 2 – Inventoried VGI initiatives

#	Name	Availability	St. date	Coverage	Access	Link
1	Ancient-tree-hunt	Av.	2004	Local	No API	http://www.ancient-tree-hunt.org.uk
2	Mapmyrun	Av.	2005	Global	P. API	http://www.mapmyrun.com/
3	Twitter	Av.	2006	Global	P. API	http://www.twitter.com
4	Veloroutes	Av.	2006	Global	No API	http://veloroutes.org
5	Wikiloc	Av.	2006	Global	No API	http://www.wikiloc.com
6	MapMyFitness	Av.	2007	Global	P. API	http://www.mapmyfitness.com/
7	Mapmyride	Av.	2007	Global	P. API	http://www.mapmyride.com/
8	Picasa	Av.	2007	Global	P. API	http://picasa.google.com
9	Endomondo	Av.	2008	Global	N/A	http://www.endomondo.com
10	Let's do it	Av.	2008	Global	P. API	http://www.letsdoitworld.org
11	Citysourced	Av.	2009	Local??	P. API??	http://www.citysourced.com
12	Geo-wiki	Av.	2009	Global	No API	http://www.geo-wiki.org
13	Crowdmap	Av.	2010	Global	N/A	http://crowdmap.com
14	Instagram	Av.	2010	Global	P. API	http://instagram.com
15	AMO Portugal	Av.	2010	Local	No API	http://amoportugal.org
16	Oil Reporter	Not Av.	2010??	Regional	No plat.	http://www.intridea.com/oil-reporter#
17	GPSies	Av.	2011	Global	P. API	http://www.gpsies.com
18	Geopoll	Av.	NA	Global	N/A	http://research.geopoll.com
19	Mapmywalk	Av.	N/A	Global	N/A	http://www.mapmywalk.com/
20	Mapmyhike	Av.	N/A	Global	N/A	http://www.mapmyhike.com/
21	COBWEB	Un. dev.	---	Regional	---	http://devel.edina.ac.uk:50503/
22	CITCLOPS	Un. dev.	---	Regional	---	http://www.citclops.eu
23	Citi-sense	Un. dev.	---	Regional	---	http://www.citi-sense.eu/

24	Omniscientis	Un. dev.	---	Regional	---	http://www.omniscientis.eu/
25	Wesenseit	Un. dev.	---	Regional	---	http://www.wesenseit.eu/
26	Yardmap	Av.	N/A	Regional	No API	http://www.yardmap.org

Appendix 3 – Prototype source code

Appendix 3 index

Code structure

The code of the prototype presented in this Appendix 3 is divided in controllers, views and services, according to the following structure:

Prototype/
- **/app/**
  - **/ /controller/**
    - /InitMapController.js
    - /MainMapController.js
  - **/ /view/**
    - /FeatureInfo.js
    - /IndividualStats.js
    - /InitMap.js
    - /LayerListView.js
    - /MainMap.js
    - /Settings.js
    - /TagList.js
    - /Viewport.js
  - **/ /services/**
    - /panoramiotags.php
    - /panoramiotags.py
  - /app.js
  - /index.html

Controllers are used to catch events and take action on them. Views are meant to provide the tools for interaction and present information to the user. In this

application, services were used to facilitate the process of requesting specific data to specific initiatives.

The code for each file in this structure is presented hereafter.

InitMapController.js

```
var map;
var allTags = [];

Ext.define('VGI.controller.InitMapController', {
  extend: 'Ext.app.Controller',
  alias : 'widget.initmapcontroller',
  config : {
    refs : {
      initMap : 'initmappanel',
      settingsForm : 'settingsform',
      mainMapPanel: 'mainmappanel'
    }
  },

  init: function() {
    console.log('InitMapController controller initialized');
    initMapController = this;
    initMapController.control({
      'initmappanel': {
        'afterrender': initMapController.onMiniMapPanelAfterRender
      },
      'settingsform button[action=request]': {
        'click': initMapController.onRequestButtonClick
      }
    }, initMapController);
  },

  onMiniMapPanelAfterRender: function(componentDV){
    var center = ol.proj.transform([-9.133, 38.713], 'EPSG:4326',
'EPSG:3857');
    layers = [];
    var osm = new ol.layer.Tile({
      source: new ol.source.OSM({})
    });

    var locationFeatures = new ol.source.Vector();
    var locationVector = new ol.layer.Vector({
      name: 'Location',
      style: new ol.style.Style({
        image: new ol.style.Icon(/** @type {olx.style.IconOptions} */ ({
          anchor: [0.5, 32],
          anchorXUnits: 'fraction',
          anchorYUnits: 'pixels',
          opacity: 0.75,
```

```
          src: './resources/pin_red.png'
        }))
      })
    });
    layers.push(osm);

    var bBoxSource = new ol.source.Vector();
    var vectorBBox = new ol.layer.Vector({
      name: 'BBOX',
      source: bBoxSource,
      style: new ol.style.Style({
        stroke: new ol.style.Stroke({
          color: 'rgba(246, 99, 79, 1.0)',
          width: 2
        }),
        fill: new ol.style.Fill({
          color: 'rgba(246, 99, 79, 0.3)'
        })
      })
    });
  layers.push(vectorBBox);

  var view = new ol.View({
    projection: 'EPSG:3857',
    center: center,
    zoom: 16,
    minZoom: 2
  });

  map = new ol.Map({
    target: 'gis_map',
    renderer: 'canvas',
    layers: layers,
    view: view
  });

  map.on('click', function(evt) {

Ext.ComponentQuery.query('[itemId=settingsLong]')[0].setValue(ol.proj.transf
orm(evt.coordinate, 'EPSG:3857', 'EPSG:4326')[0]);

Ext.ComponentQuery.query('[itemId=settingsLat]')[0].setValue(ol.proj.transfo
rm(evt.coordinate, 'EPSG:3857', 'EPSG:4326')[1]);

  locationFeatures.clear();
  locationFeatures.addFeature(new ol.Feature({
    geometry: new ol.geom.Point([evt.coordinate[0], evt.coordinate[1]])
  }));
  locationVector.setSource(locationFeatures);
  map.addLayer(locationVector);
  });

  var defaultStyle = {
    'Point': [new ol.style.Style({
      image: new ol.style.Circle({
        fill: new ol.style.Fill({
          color: 'rgba(255,255,0,0.5)'
        }),
        radius: 5,
        stroke: new ol.style.Stroke({
```

```
          color: '#ff0',
          width: 1
        })
      })
    })]
  };

  var dragAndDropInteraction = new ol.interaction.DragAndDrop({
    formatConstructors: [
      ol.format.GPX,
      ol.format.GeoJSON,
      ol.format.IGC,
      ol.format.KML,
      ol.format.TopoJSON
    ]
  });

  map.addInteraction(dragAndDropInteraction);

  dragAndDropInteraction.on('addfeatures', function(event) {
    var vectorSource = new ol.source.Vector({
      features: event.features,
      projection: event.projection
    });
    map.getLayers().push(new ol.layer.Vector({
      source: vectorSource,
      style: defaultStyle
    }));
    var view = map.getView();
    view.fitExtent(
      vectorSource.getExtent(), /** @type {ol.Size} */ (map.getSize()));
    });
  },

  onRequestButtonClick: function(button) {
    Ext.ComponentQuery.query('[itemId=vgiDataPanel]')[0].disable();
    MainMapController.clearPanoramioInfoPanel();
    MainMapController.clearComponents();
    if (allTags){
      allTags=[];
    };

    lat4326 =
Ext.ComponentQuery.query('[itemId=settingsLat]')[0].getValue();
    lon4326 =
Ext.ComponentQuery.query('[itemId=settingsLong]')[0].getValue();
    coordinates3857 = ol.proj.transform([parseFloat(lon4326) ,
parseFloat(lat4326)], 'EPSG:4326', 'EPSG:3857');
    dist = Ext.ComponentQuery.query('[itemId=settingsDist]')[0].getValue();
    bbox3857 = [coordinates3857[0]-dist/2,coordinates3857[1]-
dist/2,coordinates3857[0]+dist/2,coordinates3857[1]+dist/2];
    bbox4326 = ol.proj.transform([bbox3857[0] , bbox3857[1] , bbox3857[2] ,
bbox3857[3]], 'EPSG:3857', 'EPSG:4326');

    minxy = ol.proj.transform([bbox3857[0] , bbox3857[1]], 'EPSG:3857',
'EPSG:4326');
    maxxy = ol.proj.transform([bbox3857[2] , bbox3857[3]], 'EPSG:3857',
'EPSG:4326');
    bbox4326 = [minxy[0],minxy[1],maxxy[0],maxxy[1]];
```

```
    Ext.ComponentQuery.query('[itemId=vgiDataPanel]')[0].doAutoRender();
    mainMap.getView().fitExtent(bbox3857, mainMap.getSize());

    var boundingBoxLayer = [
                            [bbox3857[0],bbox3857[1]],
                            [bbox3857[0],bbox3857[3]],
                            [bbox3857[2],bbox3857[3]],
                            [bbox3857[2],bbox3857[1]],
                            [bbox3857[0],bbox3857[1]]
                           ];

    var polygon = new ol.geom.Polygon([boundingBoxLayer]);

    for (var i = 0; i < map.getLayers().getLength(); i++) {
      var layer = map.getLayers().item(i);
      if (layer.get('name') == 'BBOX') {
        var bBoxSource = layer.getSource();
        bBoxSource.clear();
        bBoxSource.addFeature(new ol.Feature(polygon));
      };
    };

    for (var i = 0; i < mainMap.getLayers().getLength(); i++) {
      var layer = mainMap.getLayers().item(i);
      if (layer.get('name') == 'BBOX') {
        var bBoxSource2 = layer.getSource();
        bBoxSource2.clear();
        bBoxSource2.addFeature(new ol.Feature(polygon));
      };
    };

    if
(Ext.ComponentQuery.query('[itemId=settingsPanoramio]')[0].getValue()) {
      var panoramioPhotosFrom = 0;
      var panoramioPhotosTo = 20;
      var panoramioFeatures = new ol.source.Vector();
      var vectorPanoramio = new ol.layer.Vector({
        name: 'Panoramio',
        style: new ol.style.Style({
          image: new ol.style.Circle({
            radius: 5,
            fill: new ol.style.Fill({
              color: 'rgba(250,0,100,0.4)'
            }),
            stroke: new ol.style.Stroke({
            color: 'rgba(250,0,100,1)',
            width: 2
          })
        })
      })
    });
      initMapController.panoramioDataRequest(panoramioPhotosFrom,
panoramioPhotosTo, panoramioFeatures, vectorPanoramio);
    };

    if
(Ext.ComponentQuery.query('[itemId=settingsOpenStreetMap]')[0].getValue()) {
        initMapController.osmDataRequest();
    };
  },
```

```
  panoramioDataRequest: function(panoramioPhotosFrom, panoramioPhotosTo,
panoramioFeatures, vectorPanoramio) {

    Ext.data.JsonP.request({
      async: false,
      url: 'http://www.panoramio.com/map/get_panoramas.php',
      params: {
        set: 'public',
        from: panoramioPhotosFrom,
        to: panoramioPhotosTo,
        minx: minxy[0],
        miny: minxy[1],
        maxx: maxxy[0],
        maxy: maxxy[1]
      },
      success: function(result) {

      if (result.count != 0) {
        for (var i = 0; i < result.photos.length; i++){
          console.log('Adding photo number: ', i+1);
          Ext.Ajax.request({
            async: false,
            url:
'http://localhost/phd_thesis/services/panoramiotags.php?photo_url='.concat(r
esult.photos[i].photo_url),
            method: 'POST',
            success: function(response){
              panoramioFeatures.addFeature(new ol.Feature({
                geometry: new
ol.geom.Point(ol.proj.transform([result.photos[i].longitude,
result.photos[i].latitude], 'EPSG:4326', 'EPSG:3857')),
                upload_date: result.photos[i].upload_date,
                owner_name: result.photos[i].owner_name,
                photo_id: result.photos[i].photo_id,
                longitude: result.photos[i].longitude,
                latitude: result.photos[i].latitude,
                pheight: result.photos[i].pheight,
                pwidth: result.photos[i].pwidth,
                pheight: result.photos[i].pheight,
                photo_title: result.photos[i].photo_title,
                owner_url: result.photos[i].owner_url,
                owner_id: result.photos[i].owner_id,
                photo_file_url: result.photos[i].photo_file_url,
                photo_url: result.photos[i].photo_url,
                photo_tags: response.responseText,
                vgi_initiative: 'Panoramio'
    }));

    tags = response.responseText.split(",");
    for (var ii = 0; ii < tags.length; ii++) {
      allTags.push(tags[ii]);
      initMapController.tagsListPush(tags[ii]);
    };
    }
  });
  };

  Ext.ComponentQuery.query('[itemId=vgiDataPanel]')[0].enable();
};
```

```
},
            failure: function(result) {
              alert('Error requesting metadata from Panoramio initiative');
            }
          });
  vectorPanoramio.setSource(panoramioFeatures);
  mainMap.addLayer(vectorPanoramio);

Ext.ComponentQuery.query('[itemId=layerListPanelId]')[0].getRootNode().appen
dChild({
      text: 'Panoramio',
      checked: true,
      leaf: true
    });
  },

  calculateStats: function() {

    var stats = [];
    for (var iii = 0; iii < allTags.length; iii++) {
      if (!(Ext.Array.contains(Ext.pluck(stats, 'tag'),
allTags[iii].toLowerCase()))){
        stats.push({'tag':allTags[iii].toLowerCase(), 'freq':1});

      } else {
        for (var iiii = 0; iiii < stats.length; iiii++) {
          if (stats[iiii].tag == allTags[iii].toLowerCase()) {
            stats[iiii].freq = stats[iiii].freq + 1;
          };
        };
      };
    };

    var tagStore = Ext.create('Ext.data.Store', {
      fields: ['tag', 'freq'],
      data: stats,
      sorters: [{
        property: 'freq',
        direction: 'ASC' // or 'ASC'
      }],
    });

    var tagChart = Ext.create('Ext.chart.Chart', {
      animate: true,
      store: tagStore,
      axes: [{
        type: 'Numeric',
        position: 'bottom',
        fields: ['freq'],
        title: 'Frequency'
      }, {
        type: 'Category',
        position: 'left',
        fields: ['tag'],
        title: 'Tags'
      }],
      series: [{
        type: 'bar',
        axis: 'bottom',
```

```
        xField: 'tag',
        yField: 'freq'
      }]
    });

  a = Ext.ComponentQuery.query('[itemId=individualStatsPanelId]')[0];
  if (a.items){
    a.items.each(function(item, index, len) {
      this.remove(item, true); //and remove from DOM !
    }, a);
  };

  a.add(tagChart);
  a.doLayout();
  },

  osmDataRequest: function() {
    var vectorSource = new ol.source.ServerVector({
      format: new ol.format.OSMXML(),
      loader: function(extent, resolution, projection) {
        var epsg4326Extent = ol.proj.transformExtent(extent, projection,
'EPSG:4326');
        var url = 'http://overpass-api.de/api/xapi?map?bbox=' +
epsg4326Extent.join(',');
        Ext.Ajax.request({
          url: url,
          method: 'POST',
          success: function(response){

vectorSource.addFeatures(vectorSource.readFeatures(response.responseXML));
            vectorSource.forEachFeature( function(feature) {
              z = feature;
              for (key in feature.getProperties()) {
                if (key != 'geometry') {
                  tag = key + ":" + feature.get(key);
                  allTags.push(tag);
                  initMapController.tagsListPush(tag);
                };
              };
            });
            if
(!Ext.ComponentQuery.query('[itemId=settingsPanoramio]')[0].getValue()){
              Ext.ComponentQuery.query('[itemId=vgiDataPanel]')[0].enable();
            };
          }
        });
      },
      strategy: function(){
        return [bbox3857];
      },
      projection: 'EPSG:3857'
    });

    var vector = new ol.layer.Vector({
      name: 'OSM',
      source: vectorSource,
    });
    mainMap.addLayer(vector);
```

```
Ext.ComponentQuery.query('[itemId=layerListPanelId]')[0].getRootNode().appen
dChild({
      text: 'OSM',
      checked: true,
      leaf: true
    });
  },

  tagsListPush: function(newTag) {
    console.log(newTag);
    var tagsStore =
Ext.ComponentQuery.query('[itemId=tagListPanelId]')[0].getStore();
    var numTags = tagsStore.getCount();
    if (numTags == 0) {
      console.log('entrou1');
      tagsStore.insert(numTags, { tag: newTag});
    } else {
      console.log('entrou no else condition');
      if (tagsStore.find('tag', newTag, 0, false, false, true) == -1) {
        console.log('entrou2');
        tagsStore.insert(numTags, { tag: newTag});
      };
    };
  }
});
```

MainMapController.js

```
var mainMap;

Ext.define('VGI.controller.MainMapController', {
  extend: 'Ext.app.Controller',
  alias : 'widget.mainmapcontroller',

  config : {
    refs : {
      initMap : 'initnmappanel',
      featureInfo : 'featureinfopanel'
    }
  },

  init: function() {
    console.log('MainMapController controller initialized');
    MainMapController = this;
    this.control({
      'mainmappanel': {
        'afterrender': MainMapController.onMainMapPanelAfterRender
      },
      'taglist': {
        'selectionchange' : MainMapController.onTagListSelect
      }
    }, this);
```

```
  },

  onMainMapPanelAfterRender: function(componentDV){
    console.log('onMainMapPanelAfterRender event activated');
    var mainmapextent = ol.proj.transform([-9, -7, 41, 43], 'EPSG:4326',
'EPSG:3857');
        var mainmapcenter = ol.proj.transform([-8, 40], 'EPSG:4326',
'EPSG:3857');
          var raster = new ol.layer.Tile({
            source: new ol.source.BingMaps({
              imagerySet: 'Aerial',
              key: 'Ak-
dzM4wZjSqTlzveKz5u0d4IQ4bRzVI309GxmkgSVr1ewS6iPSrOvOKhA-CJlm3'
            })
          });

        mainLayers = [];

        var osm2 = new ol.layer.Tile({
            source: new ol.source.OSM({})
        });


Ext.ComponentQuery.query('[itemId=exportall]')[0].addListener('click',
function(e) {
          var exportFeatures = [];
          for (var i = 0; i < mainMap.getLayers().getLength()+1; i++) {
            if (mainMap.getLayers().item(i)) {
                layer = mainMap.getLayers().item(i);

            if (layer.get('name') == 'Panoramio') {
              var panoramioSource = layer.getSource();
              panoramioSource.forEachFeature(function(feature) {
                a = feature;
                    var clone = feature.clone();
                    clone.setId(feature.getId());  // clone does not set the
id
                    clone.getGeometry().transform('EPSG:3857', 'EPSG:4326');
                    exportFeatures.push(clone);
                  });
            };

            if (layer.get('name') == 'OSM') {
              var osmSource = layer.getSource();
              osmSource.forEachFeature(function(feature) {
                    var clone = feature.clone();
                    clone.setId(feature.getId());  // clone does not set the
id
                    clone.getGeometry().transform('EPSG:3857', 'EPSG:4326');
                    exportFeatures.push(clone);
                  });
            };
            };
        };

          var format = new ol.format.GeoJSON();
          h = exportFeatures;
          var geoJSONString = btoa(format.writeFeatures(exportFeatures));

Ext.ComponentQuery.query('[itemId=exportselected]')[0].setHref('data:applica
```

```
tion/octet-stream;charset=utf-8;base64,' +
encodeURIComponent(geoJSONString));
          window.open('data:application/octet-stream;charset=utf-8;base64,'
+ encodeURIComponent(geoJSONString) , '_self');
        });


Ext.ComponentQuery.query('[itemId=exportselected]')[0].addListener('click',
function(e) {
          var exportSelectedFeatures = [];
          selectedFeatures.getArray().forEach(function(feature){
            var clone = feature.clone();
            clone.getGeometry().transform('EPSG:3857', 'EPSG:4326');
            exportSelectedFeatures.push(clone);
          });

          var format = new ol.format.GeoJSON();
          h = exportSelectedFeatures;
          var geoJSONString =
btoa(format.writeFeatures(exportSelectedFeatures));

Ext.ComponentQuery.query('[itemId=exportselected]')[0].setHref('data:applica
tion/octet-stream;charset=utf-8;base64,' +
encodeURIComponent(geoJSONString));
          window.open('data:application/octet-stream;charset=utf-8;base64,'
+ encodeURIComponent(geoJSONString) , '_self');
        });

        var bBoxSource2 = new ol.source.Vector();
    var vectorBBox2 = new ol.layer.Vector({
      name: 'BBOX',
      visible: false,
      source: bBoxSource2,
      style: new ol.style.Style({
              stroke: new ol.style.Stroke({
                color: 'rgba(246, 99, 79, 0.8)',
                width: 1,
                lineDash: [8,6]
              }),
              fill: new ol.style.Fill({
                color: 'rgba(246, 99, 79, 0.01)'
              })
            })
    });

    mainLayers.push(vectorBBox2);

        var mainView = new ol.View({
            projection: 'EPSG:3857',
            center: mainmapcenter,
            zoom: 4,
            minZoom: 2
        });

        mainMap = new ol.Map({
            target: 'gis_mainmap',
            renderer: 'canvas',
            controls: ol.control.defaults({
              attributionOptions: /** @type {olx.control.AttributionOptions}
*/ ({
```

```
            collapsible: false
          })
            }),
            layers: mainLayers,
            view: mainView
        });

        var selectInteraction = new ol.interaction.Select();
        mainMap.addInteraction(selectInteraction);
        selectedFeatures = selectInteraction.getFeatures();

        mainMap.on('click', function(evt) {
          console.log('mainMap clicked');
          mainMap.forEachFeatureAtPixel(evt.pixel, function (feature, layer)
{
            MainMapController.featureInfo(feature, layer);
            });
        });

      var polygonStyle = {
      'Polygon': [new ol.style.Style({
          fill: new ol.style.Fill({
            color: 'rgba(0,255,255,0.5)'
          }),
          stroke: new ol.style.Stroke({
            color: '#0ff',
            width: 1
          })
        })]
      };

      var styleFunction = function(feature, resolution) {
        var featureStyleFunction = feature.getStyleFunction();
        if (featureStyleFunction) {
          return featureStyleFunction.call(feature, resolution);
        } else {
          return polygonStyle[feature.getGeometry().getType()];
        }
      };

      var dragAndDropInteraction = new ol.interaction.DragAndDrop({
        formatConstructors: [
          ol.format.GPX,
          ol.format.GeoJSON,
          ol.format.IGC,
          ol.format.KML,
          ol.format.TopoJSON
        ]
      });

      mainMap.addInteraction(dragAndDropInteraction);

      dragAndDropInteraction.on('addfeatures', function(event) {
        var vectorSource = new ol.source.Vector({
          features: event.features,
          projection: event.projection
        });
        var polygon = new ol.layer.Vector({
        name: 'Polygon',
        source: vectorSource,
```

```
        style: styleFunction
      });
        mainMap.getLayers().push(polygon);


Ext.ComponentQuery.query('[itemId=layerListPanelId]')[0].getRootNode().appen
dChild({
      text: 'Polygon',
      checked: true,
      leaf: true
    });
      });
  },

  featureInfo: function(feature, layer){
    MainMapController.clearPanoramioInfoPanel();
    var b = Ext.ComponentQuery.query('[itemId=featureInfoPanelId]')[0];
    if (layer) {
      if (layer.get('name') == 'Panoramio') {
        var panoramioFeatureItems = [{
          xtype: 'textfield',
              fieldLabel: 'Photo ID',
              name: 'photo_id',
              itemId: 'photo_id',
              readOnly: true,
              width: 350,
              value: feature.get('photo_id')
        },{
          xtype: 'textareafield',
              fieldLabel: 'Title',
              name: 'photo_title',
              itemId: 'photo_title',
              grow: true,
              width: 350,
              readOnly: true,
              value: feature.get('photo_title')
        },{
          xtype: 'textfield',
              fieldLabel: 'Upload date',
              name: 'upload_date',
              itemId: 'upload_date',
              readOnly: true,
              width: 350,
              value: feature.get('upload_date')
        },{
          xtype: 'textareafield',
              fieldLabel: 'Tags',
              name: 'photo_tags',
              itemId: 'photo_tags',
              grow: true,
              width: 350,
              readOnly: true,
              value: feature.get('photo_tags')
        },{
          xtype: 'image',
          name: 'photo',
              itemId: 'photo',
              width: 300,
              src: feature.get('photo_file_url')
        }];
```

```
        var panelPanoramio =
Ext.ComponentQuery.query('[itemId=featureInfoPanoramio]')[0];
        panelPanoramio.add(panoramioFeatureItems);

b.setActiveTab(Ext.ComponentQuery.query('[itemId=featureInfoPanoramio]')[0])
;
      };

      if (layer.get('name') == 'OSM') {
        f = feature;

        Ext.define('osmAttribute', {
            extend: 'Ext.data.Model',
            fields: ['key', 'value']
        });

        var osmStore = Ext.create('Ext.data.Store', {
          model: 'osmAttribute',
          proxy: {
                  type: 'memory'
              },
        });

        var listView = Ext.create('Ext.grid.Panel', {
              store: osmStore,
              columns: [{
                  text: 'Key',
                  dataIndex: 'key'
              },{
                  text: 'Value',
                  dataIndex: 'value'
              }]
          });

        for (key in f.getProperties()) {
          if (key != 'geometry') {
            var rec = new osmAttribute({
                    key: key,
                    value: f.get(key)
            });
            osmStore.add(rec);
          };
        };
        d = listView;
        var panelOsm =
Ext.ComponentQuery.query('[itemId=featureInfoOsm]')[0];
        panelOsm.add(listView);

b.setActiveTab(Ext.ComponentQuery.query('[itemId=featureInfoOsm]')[0]);
        b.doLayout();
      };
    };
  },

  clearPanoramioInfoPanel: function() {

  var a = Ext.ComponentQuery.query('[itemId=featureInfoPanoramio]')[0];
  var b = Ext.ComponentQuery.query('[itemId=featureInfoOsm]')[0];
```

```
  if (a.items){
    a.items.each(function(item, index, len) {
      this.remove(item, true);
    }, a);
  };

  if (b.items){
    b.items.each(function(item, index, len) {
      this.remove(item, true);
    }, b);
  };

  a.doLayout();
  b.doLayout();
  },

  clearComponents: function() {

    if (mainMap) {
      w =
Ext.ComponentQuery.query('[itemId=layerListPanelId]')[0].getRootNode();
      while (w.firstChild) {
          w.removeChild(w.firstChild);
      };
      for (var i = 0; i < mainMap.getLayers().getLength()+1; i++) {
        mainMap.removeLayer(mainMap.getLayers().item(0));
      };
    };
  },

  mapExport: function() {
      var exportFeatures = [];

      for (var i = 0; i < mainMap.getLayers().getLength()+1; i++) {

        if (mainMap.getLayers().item(i)) {
            layer = mainMap.getLayers().item(i);
        if (layer.get('name') == 'Panoramio') {
          var panoramioSource = layer.getSource();
          panoramioSource.forEachFeature(function(feature) {
                var clone = feature.clone();
                clone.setId(feature.getId());  // clone does not set the id
                clone.getGeometry().transform('EPSG:3857', 'EPSG:4326');
                exportFeatures.push(clone);
              });
        };
        if (layer.get('name') == 'OSM') {
          var osmSource = layer.getSource();
        };
        };
    };

    var base64 = btoa(new ol.format.KML().writeFeatures(exportFeatures));

Ext.ComponentQuery.query('[itemId=export]')[0].setHref('data:application/vnd
.google-earth.kml+xml;base64,' + base64);
  },

  onTagListSelect: function(component, records) {
    var selection = component.getSelection();
```

```
    selectedFeatures.clear();

        for ( var i = 0; i < selection.length; i++) {
            tag = selection[i].get('tag');
            for (var ii = 0; ii < mainMap.getLayers().getLength(); ii++) {
              layer = mainMap.getLayers().item(ii);

              if (layer.get('name') == 'Panoramio') {
                var panoramioSource = layer.getSource();
                var panoramioFeatures = panoramioSource.getFeatures();
                for (var iii = 0; iii < panoramioFeatures.length; iii++) {
                  var featurePanoramio = panoramioFeatures[iii];
                  if (featurePanoramio.get('photo_tags').match(tag)) {
                    var photoId = featurePanoramio.get('photo_id');
              if (selectedFeatures.getLength() == 0) {
                selectedFeatures.push(featurePanoramio);
              } else {
                var flag = 0;
                for (var iiii = 0; iiii < selectedFeatures.getLength();
iiii++) {
                  if (selectedFeatures.item(iiii).get('photo_id') ==
photoId) {
                    flag = 1;
                  };
                };
                if (flag == 0) {
                  selectedFeatures.push(featurePanoramio);
                };
              };
                  };
                };
              };

              if (layer.get('name') == 'OSM') {
                var osmSource = layer.getSource();
          osmSource.forEachFeature(function(featureOSM) {
            for (key in featureOSM.getProperties()) {
                  if (key != 'geometry') {
                    if ((key + ":" + featureOSM.get(key)) == tag) {
                      z = featureOSM;
                      if (selectedFeatures.getLength() == 0) {
                      selectedFeatures.push(featureOSM);
                      } else {
                        var flag = 0;

selectedFeatures.getArray().forEach(function(featureOSMToCheck){
                          if (featureOSM.getId() !=
featureOSMToCheck.getId()) {
                            flag = 1;
                          };
                        });

                        if (flag == 1) {
                          selectedFeatures.push(featureOSM);
                        };
                      };
                    } else {
                    };
                  } else {
                  };
```

```
                };
            });
                };
              };
          };
  }
});
```

FeatureInfo.js

```
Ext.define('VGI.view.FeatureInfo', {
    extend: 'Ext.tab.Panel',
    alias: 'widget.featureinfopanel',
    itemId: 'featureInfoPanelId',
    stateful: false,
    border: true,
    width: 600,
    height: 400,
    layout: 'fit',
    title: 'Features info',
    activeTab: 0,

  initComponent: function() {
    this.items = [{
      title: 'Panoramio',
      itemId: 'featureInfoPanoramio'
    },{
      title: 'OSM',
      itemId: 'featureInfoOsm'
    }];
    this.callParent(arguments);
  }
});
```

IndividualStats.js

```
Ext.define('VGI.view.IndividualStats', {
    extend: 'Ext.panel.Panel',
    alias: 'widget.individualstats',
    itemId: 'individualStatsPanelId',
    stateful: false,
    border: true,
    width: 600,
    height: 800,
    title: 'Tag Statistics',
    layout: 'fit',
    draggable: true,
```

```
  initComponent: function(config) {
    this.callParent(arguments);

  }
});
```

InitMap.js

```
Ext.define('VGI.view.InitMap', {
  extend: 'Ext.Panel',
  alias: 'widget.initmappanel',
  html: "<div id='gis_map'></div>",
  stateful: false,
  border: true,
  listeners: {
    resize: function () {
      var size = [document.getElementById("gis_map").offsetWidth,
document.getElementById("gis_map").offsetHeight];
      map.setSize(size);
    }
  },

  initComponent: function(config) {
    this.callParent(arguments);
  }
});
```

LayerListView.js

```
Ext.define('VGI.view.LayerListView', {
    extend: 'Ext.tree.Panel',
    alias: 'widget.layerlistpanel',
    itemId: 'layerListPanelId',
    require: ['VGI.store.LayerListStore'],
    xtype: 'check-tree',
    rootVisible: false,
    useArrows: false,
    stateful: false,
    border: true,
    width: 200,
    height: 400,
    title: 'List of layers',
    draggable: true,
    listeners: {
      checkchange: function (node, checked){
        mainMap.getLayers().forEach(function(layer){
```

```
          if (layer.get('name') == node.get('text')) {
            if (checked) {
              layer.setVisible(true);
            } else {
              layer.setVisible(false);
            };
          };
        });
      }
    },

    initComponent: function(config) {
      this.callParent();
    }
});
```

MainMap.js

```
Ext.define('VGI.view.MainMap', {
  extend: 'Ext.Panel',
  alias: 'widget.mainmappanel',
  itemId: 'mainMapPanelId',
  html: "<div id='gis_mainmap'></div>", // The map will be drawn inside
  stateful: false,
  border: true,
  width: 400,
  height: 400,
  title: 'main map',
  draggable: true,
  listeners: {
    resize: function () {
      var size = [document.getElementById("gis_mainmap").offsetWidth,
document.getElementById("gis_mainmap").offsetHeight];
      mainMap.setSize(size);
    }
  },

  initComponent: function(config) {
    this.callParent(arguments);
  }
});
```

Settings.js

```
Ext.define('VGI.view.Settings', {
    extend: 'Ext.form.Panel',
    alias: 'widget.settingsform',
```

```
    itemId: 'form1',
    title: 'Settings',
    storer: 'Settings',
    bodyPadding: 5,
    layout: 'anchor',
    defaults: {
        anchor: '100%'
    },

    initComponent: function(config) {

    this.items = [{
              xtype: 'textfield',
                fieldLabel: 'Latitude (Decimal Degrees)',
                name: 'Latitude',
                itemId: 'settingsLat',
                labelWidth: 160,
                value: '',
                listeners: {
                'change': function() {
                  if ((this.getValue == '') ||
(Ext.ComponentQuery.query('[itemId=settingsLong]')[0].getValue() == '') ||
(Ext.ComponentQuery.query('[itemId=settingsDist]')[0].getValue() == '')){

Ext.ComponentQuery.query('[itemId=requestButton]')[0].disable();
                  } else {

Ext.ComponentQuery.query('[itemId=requestButton]')[0].enable();
                  };
                }
            }
            },{
              xtype: 'textfield',
                fieldLabel: 'Longitude (Decimal Degrees)',
                name: 'Longitude',
                itemId: 'settingsLong',
                labelWidth: 160,
                value: '',
                listeners: {
              'change': function() {
                if ((this.getValue == '') ||
(Ext.ComponentQuery.query('[itemId=settingsLat]')[0].getValue() == '') ||
(Ext.ComponentQuery.query('[itemId=settingsDist]')[0].getValue() == '')){

Ext.ComponentQuery.query('[itemId=requestButton]')[0].disable();
                } else {

Ext.ComponentQuery.query('[itemId=requestButton]')[0].enable();
                };
              }
            }
            },{
            xtype: 'textfield',
            fieldLabel: 'BBox size (side in meters)',
                name: 'Distance',
                itemId: 'settingsDist',
                labelWidth: 160,
                value: '50',
                listeners: {
              'change': function() {
```

```
                if ((this.getValue == '') ||
(Ext.ComponentQuery.query('[itemId=settingsLat]')[0].getValue() == '') ||
(Ext.ComponentQuery.query('[itemId=settingsLong]')[0].getValue() == '')){

Ext.ComponentQuery.query('[itemId=requestButton]')[0].disable();
                } else {

Ext.ComponentQuery.query('[itemId=requestButton]')[0].enable();
                };
              }
            }
            },{
              xtype: 'label',
                text: 'List of VGI initiatives',
                style: {
                fontWeight:'bold'
              }
            },{
                  xtype: 'fieldcontainer',
                  defaultType: 'checkboxfield',
                  items: [
                      {
                          boxLabel  : 'Panoramio',
                          name      : 'Panoramio',
                          itemId    : 'settingsPanoramio',
                          checked   : true,
                          inputValue: '1',
                          id        : 'checkboxPanoramio'
                      }, {
                          boxLabel  : 'OpenStreetMap',
                          name      : 'OpenStreetMap',
                          itemId    : 'settingsOpenStreetMap',
                          checked   : true,
                          inputValue: '3',
                          id        : 'checkboxOpenStreetMap'
                      }
                  ]
            },{
              xtype: 'button',
              text: 'Request VGI data',
              itemId: 'requestButton',
              action: 'request',
              disabled: true
            },{
              xtype: 'label',
              text: 'Select a location on the map or type the Latitude and
Longitude to start'
            }];
    this.callParent(arguments);
  }

});
```

TagList.js

```
Ext.define('VGI.view.TagList', {
    extend: 'Ext.grid.Panel',
    alias: 'widget.taglist',
    itemId: 'tagListPanelId',
    stateful: false,
    border: true,
    width: 400,
    height: 800,
    title: 'Tags List',
    layout: 'fit',
    draggable: true,
    multiSelect: true,
    columns: [{text: 'Tag',  dataIndex: 'tag'}],

    initComponent: function(config) {
      console.log('TagList Info Panel rendered');
      var tagStore = Ext.create('Ext.data.Store', {
        fields: ['tag'],
        sorters: [{
          property: 'tag',
          direction: 'ASC'
        }],
      });
      this.store = tagStore;
      this.callParent(arguments);
    }
});
```

Viewport.js

```
Ext.define('VGI.view.Viewport', {
    extend : 'Ext.Viewport',
    layout: 'border',

    requires: [
               'Ext.layout.container.Border',
               'Ext.layout.container.Fit',
               'Ext.tab.Panel',
               'Ext.tree.TreePanel',
               'Ext.tree.plugin.TreeViewDragDrop',
               'Ext.form.field.Date',
               'Ext.form.Panel',
               'Ext.Img',
               'Ext.grid.*',
               'Ext.form.field.Time',
               'Ext.form.Label',
               'Ext.data.JsonP',
               'VGI.view.TopBanner',
```

```
                'VGI.view.InitMap',
                'VGI.view.MainMap',
                'VGI.view.LayerListView',
                'VGI.view.FeatureInfo',
                'VGI.view.Settings',
                'VGI.view.IndividualStats',
                'VGI.view.TagList'
            ],

   initComponent: function() {

        this.items = [{
                    xtype: 'panel',
                    region: 'north',
                    html: '<p><br><font size="32"><strong>UGsC-Interator
Prototype</strong></font></p>',
                    border: true,
                    height: 100
                },{
                    xtype: 'tabpanel',
                    region: 'center',
                    activeTab: 0,
                    border: true,
                items: [{
                    title: 'Location',
                    layout:'border',
                    items: [{
                      xtype: 'settingsform',
                      region: 'west',
                      width: 350,
                      border: true
                    },{
                      xtype: 'initmappanel',
                      region: 'center'
                    }]
                },{
                  title: 'VGI data',
                  layout: 'absolute',
                  disabled: true,
                  itemId: 'vgiDataPanel',
                  listeners: {
                    activate: function() {
                      initMapController.calculateStats();
                    }
                  },
                  items: [{
                    xtype: 'layerlistpanel',
                    x: 0,
                    y: 0,
                    dockedItems: [{
                          xtype: 'toolbar',
                          dock: 'top',
                          items: [{
                              text: 'Export all',
                              itemId: 'exportall',
                              refTarget: '_blank',
                          },{
                              text: 'Export selected',
                              itemId: 'exportselected',
                              refTarget: '_blank',
```

```
                        }]
                    }]
                },{
                  xtype: 'mainmappanel',
                  x: 200,
                  y: 0
                },{
                  xtype: 'featureinfopanel',
                  x: 0,
                  y: 400
                },{
                  xtype: 'individualstats',
                  x: 600,
                  y: 0
                },{
                  xtype: 'taglist',
                  x: 1200,
                  y: 0
                }]
              }]
              }];
    this.callParent(arguments);
  }
});
```

Panoramaiotags.php

```
<?php
  $photo_url = $_GET["photo_url"];
  $command = "python
C:\\xampp\htdocs\\phd_thesis\\services\\panoramiotags.py $photo_url";
  $output = exec($command);
  echo $output
?>
```

panoramiotags.py

```
import json
import csv
import sys
import os, urllib2, httplib
from bs4 import BeautifulSoup
import time

opener = urllib2.build_opener()

photo_url = sys.argv[1]
```

```
try:
    html = opener.open(photo_url)
    soup = BeautifulSoup(html.read())
    ul = soup.find(id='interim-tags')
    tags = ""
    if (ul<>None):
      for li in ul.findAll('li'):
        if(li.text.find("Show all tags")==-1):
          tag = li.text.strip()
          if tags == "":
            tags = tag
          else:
            tags = tags + "," + tag
    else:
      tags=""
    html.close
except urllib2.HTTPError, err:
    if err.code==404:
        tags="No photo found"
except httplib.BadStatusLine:
    tags="BadStatusLine rised for this photo"

print tags
```

app.js

```
VGIApp = Ext.application({
    name: 'VGI',
    appFolder: 'app',

    controllers: ['InitMapController',
                  'MainMapController'],

    autoCreateViewport: true
});
```

index.html

```
<html>
<head>
    <title>UGsC-Integrator</title>

    <link rel="stylesheet" type="text/css" href="lib/ext-
4.2.2/resources/css/ext-all.css">
    <link rel="stylesheet" type="text/css" href="lib/ol-3.1.1/css/ol.css">
```

```
    <script type="text/javascript" src="lib/ext-4.2.2/ext-
debug.js"></script>
    <script type="text/javascript" src="lib/ol-3.1.1/build/ol.js"></script>
    <script type="text/javascript" src="app.js"></script>

</head>
<body></body>
</html>
```