User Generated Spatial Content: An Analysis of the Phenomenon and its Challenges for Mapping Agencies.

Vyron Antoniou

Thesis submitted for the Degree of
Doctor of Philosophy (PhD)
University College London (UCL)

February, 2011

Author's Declaration

I, Vyron Antoniou confirm that the work presented in this Thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the Thesis.

All maps are produced according to the following copyright:
Ordnance Survey © Crown copyright 2010

Abstract

Since the World Wide Web (Web) became a medium to serve information, its impact on geographic information has been constantly growing. Today the evolution of the bi-directional Web 2.0 has created the phenomenon of User Generated Spatial Content. In this Thesis the focus is into analysing different aspects of this phenomenon from the perspective of a mapping agency and also developing methodologies for meeting the challenges revealed.

In this context two empirical studies are conducted. The first examines the spatial dimension of the popular Web 2.0 photo-sharing websites like Flickr, Panoramio, Picasa Web and Geograph, mainly investigating whether such Web applications can serve as sources of spatial content. The findings show that only Web applications that urge users to interact directly with spatial entities can serve as universal sources of spatial content. The second study looks into data quality issues of the OpenStreetMap, a popular wiki-based Web mapping application. Here the focus is on the positional accuracy and attribution quality of the user generated spatial entities. The research reveals that positional accuracy is fit for a number of purposes. On the other hand, the user contributed attributes suffer from inconsistencies. This is mainly due to the lack of a methodology that could help to the formalisation of the contribution process, and thus enhance the overall quality of the dataset. The Thesis explores a formalisation process through an XML Schema for remedying this problem. Finally, the advantages of using vector data in order to enhance interactivity and thus create more efficient and bi-directional Web 2.0 mapping applications is analysed and a new method for vector data transmission over the Web is presented.

Acknowledgments

The completion of this Thesis would not have been possible without the support of a large number of individuals.

First of all I would like to extend my sincere thanks to my supervisors Mordechai (Muki) Haklay and Jeremy Morley. Their remarks and feedback during my research were always to the point and challenged me to constantly improve and strengthen my work in several ways. I feel privileged to have been their student.

Many people provided valuable help during my research. I must mention Glen Hart, Tony Joyce, Jonathan Holmes and Les Mildon from Ordnance Survey for their time and valuable discussions. Thanks must be expressed to Barry Hunter from Geograph for sharing an API key with me and for his help in retrieving data from the Geograph database. Special mention must go to Claire Ellul for always finding the time for an interesting discussion regarding my research.

I would like to thank my fellow students in the Department of Civil, Environmental and Geomatic Engineering, Nicolas Zinas, Margarita Rova, Anna Bakare, Seong Kyu Choi and Toby Webb for their warm friendship and support. Many thanks also to UCL staff. Working in UCL would not have been that inspiring and creative if the UCL family was not so polite, supportive and encouraging.

I would also like to thank the Hellenic Army General Staff and the Hellenic Military Geographical Service for sponsoring the first two years of my research. Especially the Major General (Retired) Mallh Panagioth for believing in me and supporting my nomination for funding.

Last but most definitely not least I would like to thank my wife Eirini for her support and understanding and my new-born daughter who made me so happy the last four months of this hard effort.

List of Abbreviations and Acronyms

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interfaces
DOM	Document Object Model
DCLG	Department of Communities and Local Government
EXIF	Exchangeable Image File Format
GeoJSON	Geographic JavaScript Object Notation
GI	Geographic Information
GIS	Geographic Information Systems
GPS	Global Positioning System
HTML	Hyper Text Mark-up Language
IPRs	Intellectual Property Rights
ISO	International Organisation for Standardisation
ISO/TC	ISO Technical Committee
IT	Information Technology
JSON	JavaScript Object Notation
KML	Keyhole Markup Language
LoD	Level of Detail
NMA	National Mapping Agency
OECD	Organisation for Economic Co-operation and Development
OS	Ordnance Survey
OSM	OpenStreetMap
OA	Output Area
PC	Personal Computer
PDAs	Personal Digital Assistants
PGIS	Participatory GIS
POIs	Points of Interest
PPGIS	Public Participation GIS
RSS	Really Simple Syndication
SDI	Spatial Data Infrastructure
St. Dev.	Standard Deviation

SVG	Scalable Vector Graphics
UGSC	User Generated Spatial Content
URL	Uniform Resource Locator
VGI	Volunteered Geographic Information
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	Extensible Markup Language

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Introduction

1. Introduction

> *"The World Wide Web, abbreviated as WWW and commonly known as the Web, is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them by using hyperlinks"* (Wikipedia 2010).

Starting a Ph.D. Thesis about a series of fundamental changes in the Web, the phenomenon of user generated content and their combined impact on the Geomatics domain and the mapping agencies, it was deemed appropriate to begin with a quotation of the Wikipedia definition of what the World Wide Web is. Apart from the obvious reason of providing a much needed definition, this quotation provides also a way to shortly explain what this thesis is all about. Wikipedia is a prime example of the new era that the Web has entered. The concept behind the Wikipedia project is fairly simple, yet extremely efficient: a Web application that functions as an online encyclopaedia and allows anyone to submit a new or edit an existing article for literally any possible subject. The point that this first section of the Thesis is trying to raise, lies in the fact that Wikipedia's Web page that includes the definition of "World Wide Web" has been edited more than 4,500 times from October 2001 until August 2010. That is approximately 3 edits every 2 days. Beyond doubt, this single page is a sign of a considerable collaborative effort. Still concerns might be raised regarding the need for so many edits, how accurate this definition is, who are all these people that have contributed to the Web page, what is their educational background and so on. Both the collaborative effort and the concerns raised are not unique to Wikipedia project. In fact, these issues are part of a greater debate around the new phase that the Web has entered; a phase with new and uncharted characteristics, potentials, challenges and pitfalls.

This new phase of Web is affecting almost anything that has been using or relying on this medium. The Geomatics domain is no exception. Since the early days of the Web, Geographic Information (GI) experts have used the Web to publish maps and disseminate spatial information (see for example Plewe 1997, Kraak 2001, Peng and

Tsou 2003, Peterson 2003). The Web has changed the way that maps and geographical information has been presented and used, and consequently the way that cartographers '*design, produce and deliver*' maps (Cartwright 2008, p.199). Today, the medium itself is in the middle of a major change that inevitably affects any kind of information that uses the Web, including GI. In this context, the efforts to gather, handle, share and visualise spatial content over the Web have started to change dramatically.

However, before examining the impact of those changes in Geomatics, let us first discuss the change itself.

1.1 The evolution of the Web and the emergence of Web 2.0

The advances in Information Technology (IT) over the past few years have been more than impressive. It was not a long time ago that high-specification hardware, specialised software and access to data, was a privilege only of governmental agencies or large enterprises. However, the proliferation of cheap hardware, high bandwidth and low cost hosting services, enable almost anyone with access to the technology and understanding on how to operate it, to upload content on the Web. Moreover, recent developments in mobile technology have enabled Web access via mobile devices (mobile phones and Personal Digital Assistants - PDAs). These innovations have expanded considerable both the quantity of Web users and the time they spent online (Tapscott 2009). In an in depth analysis Friedman (2006) argues that the combination of hardware proliferation, the extensive investment in infrastructure during the dot-com bubble and the emergence of standards and protocols, which enabled undisturbed communication over the Web, played a key role in the formation of a '*global platform for collaboration*' (page 92). This change led us to what is currently known as Web 2.0. The term 'Web 2.0' was coined by O'Reilly vice president Dale Dougherty in 2004 (O'Reilly 2005) in an effort to define the new strategy followed in building Web applications. O'Reilly (2005) trying to clarify what the new buzzword means, analyses the characteristics of Web 2.0 and describes the design patterns and business models of the new Web 2.0 era.

1.2 Characteristics of Web 2.0

1.2.1 Collective intelligence

One of the striking characteristics of Web 2.0 is the ability to harness collective intelligence. Collective intelligence is a wide term that has always been present in human societies but only relatively recently has this behavioural model reached the Web, appearing as the most important principle of Web 2.0. Collective intelligence can be broadly defined as *"groups of individuals doing things collectively that seem intelligent"* (Malone et al. 2009). Powered by the advent of communication technology, and innovative Web 2.0 techniques, people that connect to the Web can work in collaboration with others to tackle common problems. Early examples of such behaviour are the open source software communities. Applications like Linux, Firefox and Apache are impressive results of collective intelligence and collaboration of users in an effort to build an open source operating system, a Web browser and a Web server respectively. In Web 2.0 though, this characteristic has been transformed into the driving force behind the major popular Web applications. Key players of the Web (such as Amazon, eBay, Yahoo! etc.) having realised the power behind massive collaboration and the boost that this can result in their aims, either entice users to participate in their efforts or directly harness the intelligence created through user participation. Amazon for example, urges its users to comment on its products and then taps this participation to provide informative product overviews back to its users. In fact, O'Reilly (2007) supports that the ability to harness the collective intelligence created by users' participation has been the key factor that enabled Web 1.0 players to survive the dot-com bubble and lead the Web 2.0 evolution.

1.2.2 User generated content

A slightly different flavour of collective intelligence is user generated content. The importance of the user generated content phenomenon is twofold. On the one hand the content generated provides an immense pool of information. Early enough, the focus turned into tapping this pool and understanding and exploiting the value chains and the

business models that this new phenomenon introduces (see for example an OECD study on the subject, OECD 2007).

Despite the fact that personal web pages were early indications of user generated content, in the Web 2.0 era the phenomenon has grown to new dimensions. The underlying philosophy of Web 2.0, apparently, demands that the user ceases from being simply the consumer of information, or just part of a "*lets work together to solve a problem*" group, that helped open source software initiatives come into life, and instead dictates that the user should be promoted into a key partner in creating, sharing, consuming and disseminating information on the Web. Bruns (2008) supports that in Web 2.0 applications these activities have become so interconnected that the distinction between content producers and content users is constantly becoming more difficult. The author explains how the traditional content production process with its distinct phases of content production, distribution and consumption has been replaced by a hybrid process. Both production and usage have been condensed at a single point and take place at the meeting point of the Web platform with the participants. Thus the participants now have a fluid role (both producers and users) in this new content generation process; this process has been termed '*produsage'* and its participants as '*produsers'* (Bruns 2008). Today's produsers populate the Web with all kinds of information which actually is the underlying data flow that fuels the expansion of Web 2.0.

1.2.3 Long tail

On the other hand, equally important is the fact that this phenomenon has largely changed the nature of the information flow on the Web and particularly the privilege of controlling the published content and the level of its diversification.

The traditional method of the Web data/information flow has been from the publisher to the reader, that is, from the data/information provider to the user that consumed it. In the Web 2.0 applications though, this one-way direction of information has been replaced by a bi-directional flow through interactive environments. This has been greatly helped by a new breed of dedicated interactive tools and programming techniques (such as wiki

software and Asynchronous JavaScript and XML - AJAX) that have been engineered during the Web 2.0 evolution to keep constant data flow from users to Web applications. Thus, users are able now to generate content primarily because Web 2.0 applications provide the tools that enable users to interact with existing content or create and publish new on the Web. In most of these cases the traditional data publisher had full control over the published content. As the content generation shifted from the publisher to the users so did the authority and the control over the content. A particularly interesting case in point is CNN. CNN's Web pages used to include articles solely from journalists employed or invited by CNN. Initially, the visitors of CNN's Web pages were able to read the articles and possibly post their opinions or their objections as a letter or an email to the editor. Through this process CNN managed to have full control (and full responsibility) of the publications and their contents. When CNN decided to enter the Web 2.0 era, it provided to its users the necessary tools to start creating content themselves, and thus abolishing the power that had over the content published through its Web pages. The first step was made by providing the users the ability to post comments directly to articles. These comments that appear almost instantly at the end of the articles gave the sense that there was a broader discussion between CNN's journalists and the public about the articles' subjects. The first impression was so positive that CNN extended this bi-directional flow of information by establishing a dedicated Web application (www.i-report.com) that hosts users' videos and reports posted from around the globe. Interestingly enough, through this bi-directional flow, CNN's task of reporting the news from around the globe has been partially transposed to its users.

The emergence of users' value in Web 2.0 did not only lead to increased quantity of content available on the Web, it also gave a boost to its diversification. Since anyone can publish anything about everything it is expected that in the Web sphere even the most unusual particulars of each subject of every aspect of our society can find a place and potentially an audience. What was not expected though, and caught many Web 1.0 enterprises by surprise, was that a business model, and apparently a quite successful one, could actually be build on top of this phenomenon known as '*The Long Tail*'. The term was coined by Chris Anderson (Anderson 2004) in an effort to analyse and explain the economics and behaviour of Web 2.0 companies and their customers respectively.

The basic idea is that revenues generated by obscure content (for example unpopular books or music) that is available only online, can surpass revenues generated by mainstream ones. Such content cannot be found in a physical store because the logistical cost (including its physical storage) does not make it profitable since the actual audience that can physically reach that store is not able to generate enough income. More generally, through Web 2.0 applications that enable their users to create and interact even with obscure content, it is possible to gain eventually a critical participation mass that can well surpass the participation generated through mainstream content. Anderson (2006) provides a more in depth analysis of the subject.

1.2.4. Interactivity

The underlying basis of the three issues discussed earlier (i.e. collective intelligence, user generated content and long tail), and consequently perhaps the most important Web 2.0 characteristic is the fundamental change in another ubiquitous, yet elusive, Web 2.0 characteristic: interactivity. The assertion of interactivity's elusiveness is based on the combined fact that O'Reilly (2007) fails to distinctively recognise the importance of interactivity on the formation of Web 2.0 although there are implicit references to it and the difficulty highlighted by researchers (see for example Richards 2006 and Cover 2006) in their efforts to find references that define interactivity in its own right.

In an effort to delineate the nature of interactivity researchers have suggested that interactivity is a communication activity (Rafaeli 1988, Birdsall 2007) and thus an action that originates from the users that participate in a communication environment. Examining the issue from another point of view, Sundar (2004) describes interactivity as a technology-led phenomenon and thus as a technological attribute that the developers/authors need to plant into their applications. Moreover, researchers (Kiousis 2002, McMillan 2002) have recognised interactivity as a mixture of the two above mentioned approaches. Similarly, Hoffman and Novak (1996) contend that there can be both a level of person interactivity when the applications' environment provides inter-personal communication channels and machine interactivity when users interact directly with the application (Figure 1a).

The important point here is that the user-machine level of interaction takes place in a conscious effort to add or change the content of the application. Building on this conceptualisation, Richards (2006) further completes the description of interactivity's nature by placing it as a mediator among environment (i.e. context), content and user (Figure 1b). More importantly, Richards contends that there is a direct link between interactivity and user content generation. In an effort to describe the '*generative power of interactivity*' (p. 532), the author supports that '*interactivity is not just about exchange of communication but also generation of content*' (p. 533). Similarly, Cover (2006) notes that an interactive in nature communication blurs the line between author and audience, a position that does not differ much from the produsage concept (cf. Bruns 2008 earlier). Jarrett (2008) identifies that one of the endogenous interactivity features is its 'creative capacity'. The fundamental factor acknowledged by Richards that affects the relationship between interactivity and content creation is the relative position of user and content: when users are positioned or are able to position themselves in a proactive role with regards to creation of content, then user production becomes possible.

Additionally, in an independent approach of the subject, Newhagen (2004) also relates interactivity with content generation but through a different process. The author supports that content generation can take place (as a reaction) when there is a mental mismatch between the content presented and the user's conceptualisation of the subject. The important thing here is that interactivity enables this content generation to be hosted inside the content presented. In contrast, when content's interactivity is absent this reaction can either be suppressed or directed outside the content. A simple, yet helpful, spatial example of this would be to imagine a user presented with a map of his/her neighbourhood with obvious inaccuracies or omissions. A possible user's reaction would be the intention to correct the map. The presence of an interactive map can enable the reaction to be targeted directly at the content. On the contrary, a static (non-interactive) map will either suppress the reaction or divert the reaction to another context (e.g. send an email to the map's author).

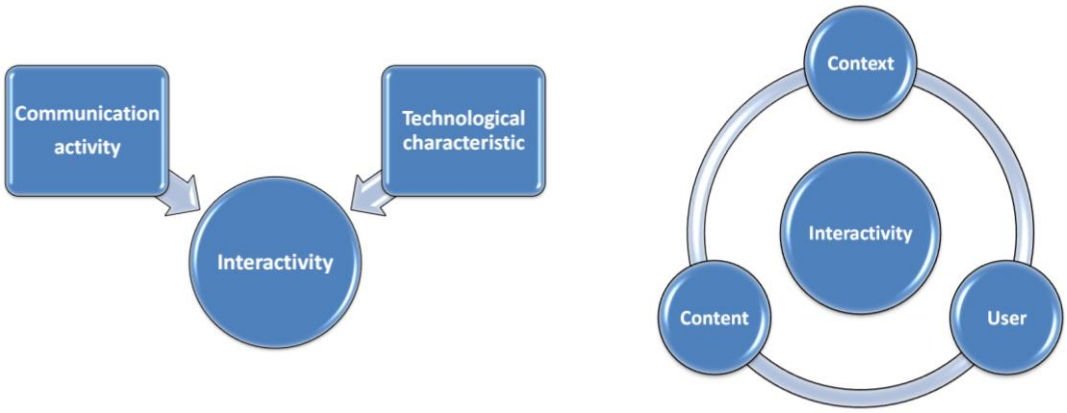


Figure 1. The nature of interactivity; a) as a combination of communication and technological efforts, b) interactivity as a mediator among context, content and users.

To realise the role and importance of interactivity in Web 2.0, apart from investigating its nature, it is equally important to examine how interactivity affects the users' *attitude* towards Web applications on the one hand and the level of their *participation* on the other. As explained below, these two factors are fundamental for the user content generation phenomenon in Web 2.0. Research has shown that an increased level of interactivity has positive effects on user engagement, with an effect on areas such as feelings of satisfaction (Rafaeli 1988), increased effectiveness and time saving (Cross and Smith 1996). Along the same line, Teoa et al. (2003) supported that interactivity has a direct impact on user attitude and application's usability. They have empirically demonstrated that high levels of interactivity have positive effect on the information delivery channels of a Web application while at the same time improve the information retrieval efficiency of the users. Also, it has been shown that interactivity assists users when they are presented with decision-making needs. Most importantly though, their research revealed that there is increased users' interest for interactive applications (in contrast with boredom for non-interactive ones) even thought the content presented was exactly the same. In other words, in a spatial context, users' attitude is fundamentally different when they are presented with an interactive map, in contrast with a non-interactive one of the same content, and this positive attitude is further enhanced as the map content's interactivity levels increase. In a sense, this reconfirms the assertion of Preece et al. (1994) that interactivity positively affect user's attitude on the Web and the findings of Ghose and Dou (1998) that supported that interactivity as a design feature is possible to improve Web usability. User attitude towards a Web application is a key factor. Long standing research on the users' approval of and their engagement with

technology has shown that user's attitude is the major factor that influences the user's intention to use a system (Ajzen 1989). In turn, the user's intention to use a system is the best predictor of the actual use of the system as asserted in the seminal research of Davis et al. (1989) about the Technology Acceptance Model (TAM) that explains the factors and principles that govern user's acceptance and their participation in new digital technologies (Figure 2).

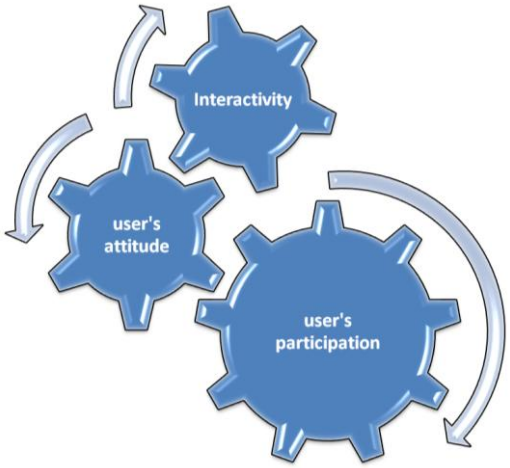


Figure 2. The role of interactivity in the improvement of user's attribute and of user's participation.

1.2.5 Amateurs and experts

Finally, perhaps the most controversial characteristic is the debate about amateurs and experts. In the aftermath of Web 2.0 evolution, issues related to the publishing of uncontrolled and unedited content on the Web by any user equipped with a computer and an internet connection, gathered increased focus. Even the supporters of this new environment acknowledge that with the plethora of authors and content available in the Web there are quality issues that have to be effectively addressed. Web 2.0 enabled millions of users to express their questionable knowledge about any subject on the same basis as experienced authors, researchers or scholars, a phenomenon that supporters of Web 2.0 welcome as the democratisation of the Web. Nevertheless, it is argued that the quality issue is not a new problem or it is associated only with the advent of Web 2.0 but also exists in traditional and well established sources of information. For example,

according to a controversial[1] investigation in Nature (Giles 2005) an average article in Wikipedia has almost the same level of accuracy as the famous Britannica that employs experts to publish its content. Moreover, Tapscott and Williams (2008) state that even without the presence of experts for guiding and examining the quality of the content published, in the Web 2.0 environment the network itself is filtering the quality of the content. Additionally, new methods and practices emerge that help machines and humans to describe content's quality such as 'tagging' and 'folksonomies'. Tagging (loosely defined) is the act of assigning specific keywords to describe content (see also Section 3.3), while the term 'folksonomy', which is built on top of tags, is used to describe user-defined categorisation of content. The overall outcome of such practices though is severely criticised as shallow and misleading (Lanier 2006). Additionally, Keen (2007) points out a number of flaws and deficiencies in the new principles introduced by Web 2.0 such as easily fabricated content or popularity of websites. Keen also supports that while untrained authors continue to publish amateurish content, users progressively become accustomed to low quality content while at the same time experienced and up to now respected authors are marginalised.

***

In this new controversial context, a number of Web 2.0 applications have flourished and managed to attract the interest and the participation of users resulting in the creation of huge volumes of user generated data. For example, YouTube provides a Web platform that enables users to discover, watch, share and comment on videos. YouTube users are watching 2 billion videos a day while at the same time there are hundreds of thousands of videos uploads daily: for every minute, 24 hours of video is uploaded (YouTube 2010) turning YouTube into one of the largest (if not the largest) video collection in the world. As discussed above, another interesting example of the magnitude of users' contribution is Wikipedia. Wikipedia currently (August 2010) has about 91,000 active[2] contributors and hosts approximately 16 million articles in more than 270 languages, making it one of the largest (if not the largest) reference Websites in the World. The

[1] Britannica has strongly objected the findings of the investigation (http://corporate.britannica.com/britannica_nature_response.pdf), but Nature insists in its original position (http://www.nature.com/nature/britannica/index.html).
[2] Users who edited at least 10 times since they arrived.

presentation of such examples shows the success stories of some of the Web 2.0 applications but at the same time they are an important illustration of a wider point. Web 2.0 has been a fertile environment that allowed otherwise lay user to create immense volumes of data; data that have been embraced by other users. It would be quite simplistic to attribute this embracement only to untrained or naïve users that obediently accept or accustom themselves to low quality content. A more fair judgment would be to recognise that there is also content of substantial value created through the Web 2.0 processes. Discovering and tapping that value is one of the most challenging issues on the Web.

1.3 The influence of Web 2.0 on Geomatics

Given the acceptance and expansion of the Web 2.0 characteristics, it is reasonable to argue that this will be the new environment where the next generation of geo-application will emerge. Web 2.0 formulated a new reality for GI of which "*the research community is still trying to grapple its meaning and significance*" (Sui 2008, p.1). The better understanding that GI experts have about this environment, its underlying philosophy, its characteristics, its rules and its principles the better will cartographic and GI Science principles be adopted so that a new generation of geo-applications will be able to exploit all the advantages offered by this new reality. In parallel, by understanding and evaluating the impact of the Web 2.0 on Geomatics and particularly i) the nature of the user generated content phenomenon in a spatial context, ii) the interactivity and the bi-directional flow of data and iii) the quality of the content generated, mapping agencies will be in position to ride the Web 2.0 wave and claim their own share of success stories in the new environment.

The first step towards this direction will be to initially examine what are the Web 2.0 influences on Geomatics. This will give a better understanding and it will provide the necessary insights to recognise potential areas of research and the challenges that need to be addressed.

1.3.1 Web-based geo-applications and user generated content

Researchers have already focused on how cartography and Geographic Information Systems (GIS) can facilitate collaboration in the forms of participatory GIS (PGIS) or Public Participation GIS (PPGIS) (see for example MacEachren 2000 and Sieber 2006). As seen earlier though, today's Web 2.0 users populate the Web with any kind of information (spatial content included) not necessarily constrained by any underlying context. The evolution of Web 2.0 has revealed a form of spatial collaboration of those numerous users who individually upload freely multiple kinds of spatial content trying to describe in detail their neighbourhoods, home towns or vacation places. As a result, numerous Web mapping applications have been created that allow users to upload, digitise, update or annotate a great variety of spatial content. Google My Maps, Wikimapia, Panoramio or OpenStreetMap (OSM) are just a few examples of a new reality in Geomatics. As it will be discussed later on (see Section 2.2.2), notable researchers have acknowledged the power of user generated content in Geomatics and focus their research interests on the challenges that this new reality presents and the roles that this phenomenon could play to the future of GI Science. Consequently, research interest has been attracted in issues regarding the most effective ways to exploit the phenomenon of the user generated content in the field of Geomatics.

As discussed, another Web2.0 characteristic that is closely related to user generated content is the 'Long Tail'. The "Long Tail" concept argues that small otherwise unpopular products can draw widespread attention when overcoming physical barriers by offering them online. This may be the explanation to the phenomenon of so many mapping websites focusing on content that official map authors would never have bothered to map (for example walking paths, cycling network and so on). Moreover, these online products can target specialised audiences that cover relatively small parts of a market, incapable of generating enough revenues to entice mainstream companies. The expression of this trend in Web mapping is local maps. Local maps generated by users are mainly aimed at small audiences interested in confined areas. The most common examples of such maps can usually be found in efforts to map university campuses like, Stanford University (ucomm.stanford.edu/cgi-bin/map/), MIT (openlayers.org/gallery/mit.html) or UCL (crf.casa.ucl.ac.uk/exploreMap.aspx). These

kinds of maps are far more detailed than the maps provided by other sources (such as Google Maps or Bing Maps) but the number of users interested in them is relatively small. Yet, as it will be discussed later on, the combination of all these patchy efforts has the potential of providing datasets of significant value.

An interesting characteristic of this process though is that many current Web mapping applications have significantly low 'mapping value'. This means that some map applications are used to present data that could actually be presented with a simple table on a web page. Such maps can be characterised as early indications of what Skupin and Fabrikant (2007) describe as spatialisation; an effort to visualise non-spatial information using spatial metaphors.

1.3.2 Spatial data on Web 2.0 and interactivity

Online geo-applications have not been left untouched by the change in the interactivity levels of Web 2.0. Yet, this change has not reached the full extent (i.e. users, context and content) of geo-applications. Interactivity is mainly confined in a personal communication level or to new content addition. Both of these types of interaction take place in the broader context of the geo-application (e.g. the exchange of comments over a geo-located photos or the upload a GPS file for further processing) and not inside the map's content (i.e. interact directly with the map's spatial features). It is still uncommon to encounter Web 2.0 geo-applications that provide interactivity at the content level (such examples are the OSM online editing applications or the polygon creation process in Wikimapia).

This reality is attributed to the fact that the Web 2.0 geo-applications are still heavily based on raster formats. Web-based geo-applications from its early days (Putz, 1994) to Google Maps is raster-based and despite considerable efforts made from a number of proprietary or international bodies to introduce and establish a vector-based format to build content-level interactive Web geo-applications, little has changed. The direct link between vector data and interactivity has been supported by numerous researchers (see

for example MacEachren and Kraak 1997, Neumann and Winter 2005, Bertolotto and Egenhofer 2001 and Bertolotto 2007 to name a few).

Up to now, Web vector formats suffered from setbacks that prevented wide implementation (see Section 2.3.2 for more on that). In contrast, the ease of creating, storing and handling raster data made it the de facto encoding for Web mapping. Yet, intrinsic characteristics of the encoding such as inflexible content, limitations in interaction and inability to explicitly store map feature objects has been brought to the surface mainly because of the explosion of user generated spatial content in Web 2.0. Indeed, despite the fact that since 2005 there have been impressive developments in Geomatics paradoxically, Web geo-applications are still struggling to provide a friendly environment for one of the fundamental elements of desktop GIS: vector data. As a result many of the emerging Web 2.0 geo-applications are considerably affected by low map interactivity, problematic feature manipulation and poor cartography.

Guided by the experience of Web 2.0 where the flow of user generated data surged when there was a fundamental change in the interactivity levels of the applications through new programming techniques and tools, it stands to reason to support that a similar surge in the flow of spatial data will occur when geo-applications will become more interactive in all possible levels, particularly though in the level of the spatial content.

1.3.3 User generated spatial content and quality

As seen, the proliferation of publishing methods in Web 2.0 inevitably allowed numerous amateur authors to create and publish content on Web applications. In many cases the content generated by users was previously exclusively handled by professional authors. Maps available on the Web were affected in a very similar way. Data gathering and handling, map composition and web mapping used to need considerable expertise. By contrast, in the Web 2.0 era, the fact that web maps have become ubiquitous indicates that even people untrained in mapping procedures (and therefore amateurs in terms of Geomatics training), can easily create and publish their own maps.

Interestingly, by examining the map topics that these two groups of authors (experts and amateurs) are choosing to focus on, a clear demarcation line is revealed. Experts in cartography and GIS when engaged in Web mapping applications, usually deal with more theoretical and thus much more complex applications than putting pins on a base map as the majority of lay users do. The majority of scholars and GI experts focus their research efforts and professional expertise on subjects such as spatial data infrastructures, geo-visualisation, spatial analysis, quality, and usability issues. Peterson (2003a) and Peterson (2008) are two indicative collections of papers for further study on expert's research topics. In contrast, the majority of amateurs' selected topics are substantially more impulsive such as summer vacations, fishing spots or Angelina Jolie's new tattoo (Sui 2008) and often take place in applications that enable socialisation among users. Accordingly, the tools used to support each group of topics are quite different. Scarcely any amateurish application is built with something more than XML-encoded files and JavaScript while the more advanced of that kind may fetch data from a database This is because now the software and data available (such as pre-generated raster tiles) take care of all the crucial questions that a traditional map maker had to answer (e.g. projection, symbolisation, naming etc.) (Goodchild 2008b). In contrast, experts' solutions usually incorporate technical specifications, sophisticated GIS software, map servers and spatial databases that have to be coordinated by the map author in accordance with the cartographic principles. Consequently, the quality of each group's mapping products is different. Relative to the concern regarding the themes chosen and the infrastructure used for the mapping products that have emerged from the Web 2.0 evolution, is the scepticism regarding the questionable quality of the spatial data used and the lack of experience, on behalf of the users, regarding the special nature of the GI and its impact on the overall quality of the user generated content. Although there are examples of notable mapping efforts that emerged from Web 2.0 applications such as OpenStreetMap (OSM), still the majority of GI that is available online is provided by amateurish, low quality and of questionable accuracy maps.

* * *

Although it seems that Geomatics follow closely the trends introduced by Web 2.0, there are some fundamental differences between these two domains. Firstly, in Web 2.0,

most of the time, the user generated content is created from scratch or reflects the users' knowledge of a subject. This is in contrast with the reality in Geomatics. The basic tool used as a backdrop to generate spatial content is mapping products by companies such as Google, Yahoo! or Microsoft or national mapping agencies (NMAs) like the Ordnance Survey (OS). Thus, users have a substantial starting point on which to build their maps – that is, the topographic maps or satellite images that have been created or processed respectively by experts with well-established and trusted methods. Moreover, GPS devices are frequently used to capture spatial data equip amateur users with "professional" tools. Still, concerns exist regarding the positional accuracy of spatial content present in Web 2.0 (see Goodchild 2007a and 2008c for example) or regarding quality elements such as completeness and attribution.

1.4 Research issues

The abundance of Web 2.0 geo-applications available today consist of potential sources of spatial content that can be used to update or enrich mapping products. At the same time, there is the inability of mapping agencies around the world to keep up-to-date their spatial databases. Mapping and map updating programs are experiencing serious delays in many countries. In that context the value of the crowdsourced spatial content available in the Web is something that mapping agencies cannot afford to discard and thus questions are raise about the nature of the phenomenon and the potentials that can emerge for mapping agencies.

Definitely, the phenomenon of user generated spatial content is a multi-dimensional one (there are social, legal and technological aspects, to say the least). In parallel, the phenomenon can be examined from different angles. The aim of the Thesis is not to cover every aspect. Instead there are some characteristics of the phenomenon that have been selected for further research and analysis. The general understanding of the phenomenon at a national level (using empirical methods in contrast with the existing theoretical assumptions), the methodologies to enhance the productivity of the phenomenon in terms of content generation and the analysis of the content's quality are the main issues that this Thesis chose to focus on.

The point of view selected is that of a mapping agency. By the term 'mapping agency' is meant any institution public or private that aims to collect spatial content and provide mapping products to a broad (regional, national or global) audience. For example, such public national-level mapping agencies could be the Ordnance Survey of Great Britain or the Hellenic Military Geographical Service of Greece. On the other hand, Google, Microsoft and Yahoo! belong to the private mapping agencies aiming to a global audience. It is important to note that the analysis conducted here is clearly more focused on the phenomenon itself and not on the requirements of the mapping agencies. Therefore, the role of the mapping agencies in this Thesis can be resembled with the metaphor of an interested and thus a critical bystander who is not directly interacting with the phenomenon but has a considerable interest to understand, evaluate and possibly engage with it.

Given the fact that this phenomenon is a new reality for mapping agencies it is understandable that there are issues that need to be clarified regarding whether and how they should approach the phenomenon of user generated spatial content:

- *What is the nature of the phenomenon and what are its main characteristics?*

This knowledge is needed to understand the phenomenon and evaluate its overall spatial value. It is also needed to document how this phenomenon is realised in the world of Web 2.0 and monitor how users are behaving and the spatial data that they are creating. This will enable to evaluate the processes followed and improve the creativity of the crowd.

- *What is the quality of the user generated spatial content?*

Gaining insights regarding the quality of the spatial content generated by lay users is important both for the engagement of mapping agencies with it and for the future evolution of the phenomenon itself. On the one hand, this knowledge will be used by the Web 2.0 geo-applications to improve their data generation processes and thus raise the quality level of the content offered. On the other, the consumers of such content (either simple or institutional users) will understand the merits and the potential uses of the data at hand.

- Are there any challenges that are unique to the phenomenon that need to be addressed, and what are the best solutions?

As this is a totally new phenomenon in the Geomatics domain, it is reasonable to expect that there is a series of challenging issues that need to be faced. These challenges need to be surfaced promptly and efficient solutions and methodologies should be developed to confront them.

In the course of the Thesis, and after the review of the relative literature, these questions and the issues raised here will be further clarified, grouped and re-submitted as the core research objectives of this Thesis along with the description of the methodology adopted to answer them.

1.5 Contribution

The contribution of this Thesis springs from two sources.

On the one hand, it is the results generated during the research, the discussion of the findings and the answers offered to the research questions, and the conclusions in which this process leads. The subject of the research is a novel and fairly unknown phenomenon and thus a firm and empirically proven knowledge base needs to be built. The core of this research is oriented to the fulfilment of this aim as its findings contribute to this knowledge base valuable insights regarding the overall nature of the user generated content phenomenon and its relationship with space and geography. Moreover, the Thesis analyses the major types of geo-applications through which the phenomenon is realised, namely the social networking photo-sharing Websites and the vector-based ones, highlights their particularities and concludes on a typology based on their attitude against space. Furthermore, insights are provided for the evolution and the trends of the phenomenon as well as the contributors' behaviour. Finally, there is an empirical study of the spatial data quality of user generated spatial content (focusing on positional accuracy and attribution of vector data).

This series of experiments provide initially a basic understanding of the phenomenon, but more importantly sketch the potentials of the phenomenon in the Geomatics domain. Particularly from the standing point of a mapping agency, the knowledge base provided allows the evaluation of the user generated spatial content on its merits and consequently enables mapping institutions to evaluate the gains and the added value that a potential use of crowdsourced data can offer. Finally, as the experiments reveal both the valuable and the erroneous aspects of the phenomenon their results can equip mapping agencies with the necessary information to be proactive against potential pitfalls and challenges that a possible engagement with user generated spatial content hides.

Relative to this final point is the second source of the Thesis' contribution. Throughout the research a number of challenges has been surfaced that need to be thoroughly studied and efficiently addressed. Highlighting, discussing, analysing and providing solutions for the demanding issues of this new phenomenon, advances the research on the subject and prepares its adoption from the Geomatics domain.

The first challenge lies in the erroneous processes that are endogenous to the user generated content phenomenon. The freedom and flexibility in content creation generates errors that affect the contents' quality. A methodology for improving quality through the formalisation of the whole process and the introduction of a specifications-based user contribution process, that also takes advantage of the availability of interactive content, is presented. This methodology will greatly enhance the overall user generated data quality while at the same time it will leave unaffected the openness of a Web 2.0, crowdsourcing geo-application or the excitement and sense of freedom that lay geographers feel when contributing to such applications. Another major challenge recognised during this research was the need for improvement of the content's interactivity of the Web 2.0 geo-applications. The crucial point here is that any mapping agency that aims to become involved with spatial content generated on the Web, has to invest in the development of interactive geo-applications, able to foster and enhance user participation that will lead in substantial flow of user generated spatial content. In other words, as explained earlier, effective transmission methods for vector data have to be engineered to supply the much needed interactivity to the Web 2.0 geo-applications'

content. This research offers a methodology for vector data transmission over the Web, applicable to multi-scale datasets (which practically is the rule for mapping agencies), that overcomes long standing problems that prohibited the widespread use of vector data in Web mapping applications. For this group of challenges practical solutions will be presented.

The next group of challenges is revealed from the discussion and analysis of the empirical results. Here, a more theoretical approach for confronting the new challenges will be presented. One particularly important issue revealed in the course of the Thesis is the nature and the specific characteristics that a Web 2.0 geo-application should have to be productive in terms of content generation but most importantly, to be efficient in terms of spatial coverage to act as universal source of spatial content in a broad level (e.g. in a national level when it comes to national mapping agencies). Finally, the results of this research can be used to support the convergence of user generated spatial content with well established mapping procedures (such as map update, change detection, map auditing, enriching existing databases or creating new products) still followed by mapping agencies. Before entering this phase though, it is important for mapping agencies to understand the environment and the circumstances under which an engagement with the Web 2.0 world should take place. As the spatial data quality is paramount for mapping agencies, the need for strong reassurances for the user generated data quality is a major challenge. Therefore, a theoretical approach for a constructive way of quality information sharing is presented.

1.6 Thesis structure

This Thesis comprises 8 Chapters. This is the first one and it has provided an introduction on the evolution and the basic characteristics of Web 2.0. Next the influence of this evolution on the Geomatics domain has been examined and an outline of the research issues and the contribution of the Thesis have been presented.

Chapter 2 provides the literature review around the subjects that have been singled out as important for the scope of this Thesis. Thus, there are three different sections: the

phenomenon of user generated content itself, the factor of content-level interactivity and the issue of content's quality. Regarding the phenomenon itself the literature review provides the initial response of the research community and highlights both the scepticism and the potential value of the phenomenon. Regarding the interactivity issue, the existing methods for spatial data transmission over the Web are examined and their limitations are highlighted. This paves the way to focus on the development of a methodology that is able to overcome the existing limitations and provide the necessary content-level interactivity. The Chapter ends with the section about quality. The basic principles of spatial data quality and the methods applied for its examination are presented. Finally, the particular relation between crowdsourced spatial content and quality is examined.

Chapter 3 starts by clarifying and formalising the research questions. Then, the methodologies followed to provide adequate answers are presented. A series of experiments is designed and analysed so as to explore the dimensions of each particular issue. There are two basic experiments. The first one focuses on the analysis of the phenomenon using as sources of the spatial content the geo-tagged photos of four popular photo-sharing Web 2.0 applications. The second one is concerned with the analysis of the vector based datasets contributed to OSM. In each case a number of sub-experiments are conducted.

Chapter 4 and 5 present the results of the experiments for the geo-tagged photos and vector-based datasets respectively.

In Chapter 6, helped by the experiments' findings, a series of challenges around the user generated content phenomenon is presented along with the practical solutions.

Chapter 7 then completes the three previous Chapters as it presents a full discussion of the empirical findings and the challenges met. Moreover, in this Chapter, the discussion expands to another set of challenging issues that are generated when the phenomenon of user generated spatial content is considered from a broader point of view. Here, theoreticall suggestions are presented.

Chapter 8 is the final of the Thesis and consolidates the conclusions of the research. The knowledge gained by the experiments, their findings, the discussion, the challenges and their solutions is summarised in this Chapter. Finally, based on the experience gained, a discussion about the future of this new, unexplored and rapidly evolving phenomenon is presented along with issues that need further research. In the course of the Thesis a number of such issues have been recognised but not analysed in depth. These issues are forming a series of suggestions for further work.

Chapter 2

Literature review

2. Literature Review

2.1 General

As discussed, the Geomatics domain has been influenced by innovations that emerged during the evolution of the Web 2.0. Interestingly, there are also some spatially-related factors that played a catalytic role to the Geomatics domain. The first milestone was the removal of the selective availability of the Global Positioning System (GPS) signal by the U.S. President William Clinton (Clinton 2000). The importance of this act and the consequent proliferation of everyday GPS-enabled devices have been acknowledged by a number of authors (see for example Goodchild 2007a, Goodchild 2007b, Cartwright 2008, Haklay et al. 2008, Elwood 2008a, Goodchild 2008a). Indeed, the integration of an accurate positioning system in everyday devices such as mobile phones or in low cost, autonomous hand-held GPS receivers, enabled users to gather spatial information effortlessly and spontaneously. In turn, this signalled the beginning of an unprecedented spatial data flow by the users to scattered geo-applications all over the Web. A second factor that greatly contributed to the advance of the Web-based geo-applications' functionality so to reach the functionality levels of the rest of the Web 2.0 applications was the introduction of a programming technique known as AJAX (Miller 2006, Plewe 2007, Haklay et al. 2008). AJAX is a well known methodology that has played a key role in the evolution of Web 2.0 by helping programmers to build desktop-like applications over the Web. AJAX is based on the coordinated exchange of XML data fragments using JavaScript. XML is a text-based format endorsed by the World Wide Web Consortium (W3C), which allows interoperable communication among computers. When combined with JavaScript, which is a browser scripting language, through asynchronous requests by the browser to the server, applications' response times are minimised enhancing user interaction and usability. Particularly in regards with the geo-application though, AJAX's role was even more crucial. By implementing AJAX methods it was made possible to overcome long standing obstacles related with the transmission of raster data over the Web (due to their volume) or poor users' interaction with the mapping applications. A third important factor was the development, publication and free access of Application Programming Interfaces (APIs). APIs are

programming interfaces that allow developers or programming savvy users to program the core functionality of a Web service. Particularly for the Web-based geo-applications the APIs enable developers to mix spatial data with other types of data and applications to create a wide variety of applications known as mash-ups (Miller 2006) or to develop their own spatial-related independent applications. The APIs provided by major vendors such as Google, Yahoo! and Microsoft are the capstone in the proliferation of the Web geo-applications as they gave to the users the power to literally put a map in any Web page with little effort and in a very short time. This resulted in Web users becoming more and more familiarised with the subject matter of space, location, geography and maps (Goodchild 2008a) and thus enabled users to realise the value, the importance and the potentials of GI. Consequently, this generated an upwards spiralling helicoid of need for accessible spatial data, ubiquitous map availability and up-to-date maps on the Web as well as in everyday devices.

Before the proliferation of Web mapping APIs, cartographic experts mapped geographic entities, and subsequently printed or published on the Web, content that was of some accepted significance. This was dictated by the fact that GIS and mapping software was too expensive to use for obscure purposes. There was little investment in creating maps for smaller audiences and non standardised map products. With the appearance of map APIs the scenery changed dramatically. The free APIs coupled with ease of use and cheap hardware turned non-experts, and up to now simple map users with little programming experience, into map authors. The diversity of users' interests and hobbies were provided the means to be published on the Web with minimal effort and at considerably reduced cost. Goodchild (2008b) describing the proliferation and the diversity of the Web geo-applications, provides an analysis of the driving forces of the phenomenon: the diffusion in the power of mapping from institutions to individuals is attributed to the fact that there was a transformation in the economies of scale regarding the cost of mapping. The scale economies before Web 2.0 dictated that mapping products should target as broad an audience as possible to cover the cost of production. The proliferation of GPS-enabled devices, ready to use mapping APIs over geo-data of global coverage and the explosion in user generated spatial content, have significantly reduced the entrance costs to Web mapping which has resulted in the appearance of highly diversified mapping content (Goodchild 2008b).

The convergence of the above mentioned factors created a new environment in Web mapping. This environment enabled the rest of the Web 2.0 principles to flourish leading to the appearance of a new phenomenon described by Goodchild (2007a) as Volunteered Geographic Information (VGI) or by Turner as 'neogeography' (Turner 2006). Turner defines that '*neogeography means new geography and consists of a set of techniques and tools that fall outside the realm of traditional GIS*' (p. 2). In this novel reality a new breed of geospatial applications emerged that cover a wide range of applications. For example, there are applications that urge their users to contribute several points of interest (POIs) in the http://garminpoi.co.uk, walking paths in http://www.everytrail.com, photos, tags and descriptions in http://www.geograph.org.uk, geographical names in http://wikimapia.org or even complete topographic maps in http://www.openstreetmap.org. It is indicative that more than 50,000 Web 2.0 geo-applications and mash-ups were available in the first two years since the publication of the Google Maps API alone (Tran 2007).

From another point of view this wealth of Web 2.0 geo-applications is expected to need a constant flow of spatial content to be sustainable and evolving. Thus, user participation and especially content creation is the vital point. As discussed in Chapter 1, interactivity is an important factor that generally leads to user contribution and is abundant in the Web 2.0 applications (Section 1.2.4) but it is considerably restricted when it comes to intrinsic content interactivity in the Web-based geo-applications (Section 1.3.2). Therefore the need to research methods for bridging this gap is considered crucial for the evolution of the phenomenon as interactivity and content generation have a cause-and-effect relationship (as explained in Section 1.2.4).

Another important aspect is the overall value of the phenomenon. It is understandable that the value of user generated spatial content is closely related to the quality of the data created. The availability of high quality data is not guaranteed; on the contrary quality is one of the main concerns and calls for further research on the subject are raised by scholars (see Sections 2.2.1.1 and 2.4) to further understand the phenomenon. Interestingly, content-level interactivity can prove to be a valuable asset towards the improvement of the data created (see Section 6.2).

Therefore, content generation, interactivity and data quality are among the most important factors affecting the user generated spatial content phenomenon. The literature review that follows concentrates on these three issues.

2.2 User generated spatial content

As the phenomenon of VGI started to draw attention among scholars, a growing debate about the term and specifically the word "volunteered" began (Elwood 2008a). Researchers (Obermeyer 2007, Sieber 2007, Williams 2007, Elwood 2008b, Bishr and Mantelas 2008) suggest that the term "volunteered" can be misleading regarding the particularities of the generated data and the intentions of the data providers. In a sense "volunteered" implies a noble and altruistic gesture as if the users donate the data, personal or not, to the world for any use, known or not to the data provider. Acknowledging the issues raised, a more general but still precisely descriptive term is used in this research: User Generated Spatial Content (UGSC) (Antoniou et al 2009b, Brando and Bucher 2010).

2.2.1 UGSC scepticism

Moving a step further, a more substantial line of criticism than the name of the phenomenon emerges when considering the nature of the phenomenon itself.

2.2.1.1 Quality

The most compelling issue is perhaps the quality of UGSC or the credibility (Flanagin and Metzer 2008) of such content.

As the phenomenon of UGSC has a strong social aspect, researchers have recognised the close relationship between the social factors that motivate users to participate in content creation with the quality of the content itself. For example, Goodchild (2007a)

raises the issue of users' motivation for content creation. The author relates the motivation with the overall behaviour as the danger that users will not forever behave in an altruistic manner is raised. Consequently, malicious contributions, like those that are very common and largely anticipated to the rest of the Web today, will start to appear in the context of UGSC. For instance, is there any guarantee that malicious or selfish users are not going to tamper an areas' dataset to promote their agenda (e.g. the removal of a Roma's camp by a real estate agent). Similarly, Coleman et al. (2009) explain that as there is a considerable range in the motivation of the users that participate in the creation and sharing of spatial content on the Web, the quality of the data can range significantly. According to the authors, the understanding of the participants' motivation and nature can give valuable insight about the resulting content quality.

Further to these social-related issues are factors that relate to the users' expertise and the familiarity with the subject matter of issues like space, geography and GI creation and thus can considerably affect the quality of the data created. Goodchild (2008b) explains that the fact that contributors or *neo-geographers* (Turner 2006) lack any cartographic background or the skills that professional geographers and surveyors have is affecting the quality of the data and the mapping products that are created through this process. Moreover, Flanagin and Metzer (2008) and Goodchild (2007a) raise concerns regarding the credibility of UGSC since data often come from multiple sources which may have vague origins. This results in objective difficulties in assessing the credibility of information at hand or in the understanding of possible misuses of data that was never intended for particular purposes.

On the other end of the spectrum though, even in this open and non-authoritative context, measures to safeguard the content quality can be taken. Most of the suggestions towards this direction originate from the accumulated experience of other cases of citizens' science where there are tested methods that can apply different degrees of quality control over user generated content; methods that can be applied in UGSC as well. For example, the practice of adequate training before data capture as in Christmas Bird Count and the Project Globe (Goodchild 2007b, Flanagin and Metzer 2008) can apply to more spatially explicit projects as is the case of The National Map Corps (Bearden 2007). Moreover, there are examples of social approval as in Wikimapia and

of professional editing before adopting the information as in the "people's map" initiative in Scotland (Rideout 2007).

Additionally, crowd participation can prove a strong quality improving factor. For example, Flanagin and Metzer (2008) suggest that judging by the developments in popular Web sites that favour user generated content, the existing UGSC systems will gain in terms of data quality and credibility if they manage to achieve substantial increase in their popularity and usage. Similarly, social filtering will possibly prove to be a sufficient mechanism that will manage to contribute to the improvement of the overall quality of UGSC (Flanagin and Metzer 2008, Goodchild 2008b).

Nevertheless, it is interesting to note some research issues raised relative to the subject of UGSC quality:

'But largely missing at this point are the mechanisms needed to ensure quality, to detect and remove errors, and to build the same level of trust and assurance that national mapping agencies have traditionally enjoyed' (Goodchild 2007b, p.31);

or

'Yet, to date researchers have barely begun to examine the credibility of VGI. Pressing questions in this pursuit include whether users and professionals will accept systems populated largely by volunteered input as credible and, if so, for what purposes and with what effects?... what technical and sociotechnical tools can help users and professionals navigate VGI systems appropriately?....At the same time, however, problems of knowing what VGI systems and sources to trust will likely continue to affect usage of these systems.' (Flanagin and Metzer 2008, p. 144).

Such statements eloquently show that research on the subject of UGSC quality is still in its infancy.

2.2.1.2 Sustainability

Long term sustainability of UGSC is also under investigation. It is a fact that many popular spatial Web applications (such as Wikimapia, OSM or Geograph) entice their users to participate to achieve, what is presented as a noble cause, the mapping our world. However, researchers are questioning if this participation will end when the cause is achieved and thus transforming the impressively growing datasets into neglected, out-of-date archives (Goodchild 2008b). In the same spirit Sui (2008) questions if this phenomenon is sustainable or a passing fad.

Yet, despite the justified scepticism, based on the observations of the up to now evolution of Web 2.0 and the social networking and the proliferation of the spatially enabled devices the consensus seems to be that UGSC will be an enduring phenomenon (Goodchild 2008b, Craglia et al. 2008). Moreover, Flanagin and Metzer (2008) suggest that the growing fiscal interest for Web mapping applications that host user participation will eventually lead to methodical popularisation of such applications and consequently to further engagement of on-line users.

Of course not all of the currently existing Web 2.0 spatial content generating applications will continue to exist for decades ahead. Some are expected to become obsolete; they will gradually fade and finally disappear. Some of the remaining will merge either out of survival purposes or to pursue other more demanding aims. But the future of UGSC should be assessed as part of the more general issues of user participation and user generated content on the Web (Goodchild 2007a). Both phenomena are expected to be enhanced rather diminished by the ongoing evolution in IT and social trends.

2.2.1.3 Digital divide

Social disparities appear in different flavours. In the digital word of IT the disparities recorded have been named as *digital divide* in an effort to describe the different levels of access that people have to digital technology. As expected Web 2.0 has not been an exception (see for example Cox 2008), neither has the phenomenon of UGSC.

Interestingly though, for the particular case of UGSC this digital divide affects both the production and the consumption of digital information. Regarding the former issue for example, Haklay (2010) has showed that the spatial content contribution to OSM is considerably lower in deprived areas compared to the more affluent ones. For the latter issue though there is a considerably greater awareness. Scholars and researchers (see for example Goodchild 2007a, Craglia et al. 2008, Sui 2008, Maue and Schade 2008, Goodchild 2008b to name a few) acknowledge the problem of inequality in accessing UGSC but also point out that given the right tools and incentives this proliferation of GI in the Web, might very well be a step towards the bridging of the digital gap.

2.2.1.4 Intellectual Property Rights (IPRs)

Another source for scepticism is the issue of intellectual property rights (IPRs) of the spatial content available in the Web. Although the presence of spatial data seems ubiquitous the fact is that Web mapping services often pose strong restrictions that disallow any third party to commercially use their content (see for example Appendix A for indicative excerpts that describe types of IPRs from different sources of spatial content).

At this point it should be stated clearly that the examination of IPR issues related to UGSC is out of the scope of this research, not least due to the legal knowledge needed to analyse this issue. Yet, not examining the impact of IPRs in the exploitation of UGSC is far from an arbitrary simplification. There are many cases where IPRs are not an issue at all. Such cases can be found when mapping agencies gather spatial content from in-house Web mapping applications and so they can freely use the content to improve their spatial repository (e.g. when spatial content is submitted to a central military mapping agency by Units that operate on the field or Web mapping applications from commercial companies like Google, TomTom and Navteq or national mapping agencies like OS). In other cases the use of UGSC is permitted by the provider to a certain extent. For example, data that is licenced under certain versions of the Creative Commons Licence (e.g. Attribution-Share Alike 3) is free for any use as long as the outcome of that use is published under the same or similar licence scheme. Finally, and perhaps most

importantly, in a field undergoing major changes in the licensing and pricing policies, not least due to initiatives such as the European Commission Directive of INSPIRE (EU 2007), the free/non-commercial data distribution might very well be the common practice in Europe and most of the developed countries.

* * *

UGSC was bound to be a matter of controversy (the interested reader can see also Elwood 2008b). All these different aspects of consideration and scepticism regarding the phenomenon of UGSC were expected given the extent, the importance, the implications and the novelty of the phenomenon in the Geomatics domain. A step further, it can be supported that this is a constructive and thus welcomed process that enables initially the introduction and then the gradual assimilation of technological and social advances in the body of a scientific domain. Additionally, through this scrutiny the value and the potential of UGSC in Geomatics started to emerge. As it is explained in the following Section, after the initial scepticism the value of UGSC has been acknowledged as it can improve GI in several ways by providing adequate flow of spatial data that describe our world in more detail.

2.2.2 UGSC value

Although humans are naturally aware of space, the proliferation of GI in the Web and in everyday devices (e.g. GPS navigators or mobile phones) is radically changing the relationship between the public and the subject matter of geography. Phelan (2008) eloquently described the degree of that change by saying that "*...we will be the last generation to know what it means to get lost*". Indeed, this abundance of GI can be the starting point that could make scholars, professionals, mapping agencies and lay users alike to recognise the value of UGSC and work on the new challenges that this phenomenon is presenting to the Geomatics' domain.

2.2.2.1 Extended field of scope

Earlier (see Section 2.1) the discussion focused on the range of Web geo-applications and particularly on the types of spatial data that are currently available on the Web. Another aspect of this discussion could be the range of scope of those applications and consequently the fields that UGSC covers. For example, Goodchild (2007b, 2008b) provides examples of air pollution measurements, traffic and congestion recordings, intelligence and homeland security data collection and soil mapping. In a sense, this argument perhaps falls under the broader discussion of citizen science and the role of humans as sensors (Goodchild 2007a). Indeed, as data collecting devices become more and more accessible the vision of having on the planet a 6 billion-sensors network that could potential record anything of importance comes closer. Moreover, it has been proven that initiatives even if they are based solely on citizens' participation they can provide the necessary volume and quality of the data collected to serve specific aims (e.g. Christmas Bird Watch).

2.2.2.2 Cost

Researchers (Goodchild et al. 2007, Goodchild 2008b, Budhathoki et al. 2008) have raised the point of the high costs needed for the collection of the necessary spatial data to produce a topographic map. Although the technologies involved (such as satellite imagery and aerial photographs) are quite expensive, the fact remains that they can provide only part of the data needed to complete a topographic map sheet (e.g. geographical names cannot be remotely sensed). In contrast the phenomenon of UGSC can, if efficiently tapped, be the answer for reduced costs. The 'volunteered' mode of spatial data contribution from a broad network of lay users is considerably more cost-effective than to use (expensive) professionals for data collection. Moreover, from the point of view of a mapping agency it is considerably more cost-effective to invest in coordinating such a network (i.e. provide the proper mechanisms for content generation, quality assurance, user incentives etc.) than to plan and execute the whole project by itself and of course by using solely its own means (Shirky 2005).

Additionally, scale economies have so far dictated that collection and administration of spatial data could be achieved only in a centralised way under the protection, guidance and funds of governmental agencies or big enough companies resulting into spatial products that could have multiple uses. As the local production and consumption of spatial information is gaining ground, the collapse of scale economies push towards the exploitation of locally produced spatial content (Goodchild 2008b).

2.2.2.3 Correct, enrich, complete and update existing datasets

Closely related to the previous issue is a challenging point raised by Goodchild (2007a) who states that the arguments made by Estes and Mooneyhan (1994), about a mistaken popular notion of a well mapped world, are still true. He argues that in fact, mapping and map updating programs are suffering serious delays in many countries mainly due to the increased cost of mapping. For example, the author refers to the U.S. Geological Survey policy of not updating its map series on a regular basis and instead preferring a more patchy way of updating their database at a national level. Similarly, McDougall (2009) describes how the role and the output of national mapping agencies have been considerably downsized while the need for spatial data has increased. Based on these arguments, it can be suggested that the UGSC model of data creation fits the requirements of NMAs. Early efforts towards this direction can be traced in the USGS's National Map Corps program (Bearden 2007).

In parallel, as discussed in the previous Section, the inability of traditional methods of spatial data collection to effectively capture and attribute data that are not detectable remotely, considerably enhances the role of the phenomenon in completing existing datasets. Finally, the use of UGSC can help to enrich longstanding geographic products like gazetteers which traditionally were based on the use of knowledge that local people had (Goodchild 2008b) or to correct existing datasets. Importantly, local citizens are characterised by researchers as the most suitable sensors for identifying errors or notifying for changes that take place at their local area (Goodchild 2008a, Goodchild 2008b, Craglia et al. 2008). This special relationship between UGSC and local knowledge is discussed in Section 2.2.2.5

2.2.2.4 The contribution of UGSC in SDIs

The U.S. Executive Order 12906 (Clinton 1994) defines a National Spatial Data Infrastructure (SDI) as the '*the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilisation of geospatial data*' (p.1) and indentified the avoidance of wasteful duplication of effort and the promotion of effective and economical management of resources as a primary motivation for creating one (Clinton 1994). Many mapping agencies have been involved in the development of national or regional SDIs. Two prominent examples are the NSDI in the U.S. and the INSPIRE initiative in Europe.

The creation of such an infrastructure follows a top-down approach (Budhathoki et al. 2008, McDougall 2010) and it is developed by governmental agencies to be used by other institutional agents, or in other words, they are designed by experts to be used by experts Craglia (2007). Researchers (Coleman et al. 2009, Budhathoki et al. 2008) have highlighted the analogy between the bottom-up way of spatial data collection from lay users and the intense top-down effort put by NMAs in building SDIs on the one hand and the Raymond's (1999) example of the cathedral and the bazaar that refers to the open source software development on the other. Interestingly though, the difference here is that the focus is not on the competitiveness of the two approaches rather on possible convergence of these two strains of spatial data creation (Craglia 2007). For example, Goodchild (2007a) recognise the fact that the nature of the UGSC seems as a suitable mechanism to fit the development of SDI. Moreover, Budhathoki et al. (2008) and McDougall (2009, 2010) argue for a re-conceptualisation of both user's role in the context of an SDI and the functionality and openness of the model of the SDI itself (Figure 3). The call is for a more user-centric SDI model that will take advantage of the user participation and the value of UGSC.

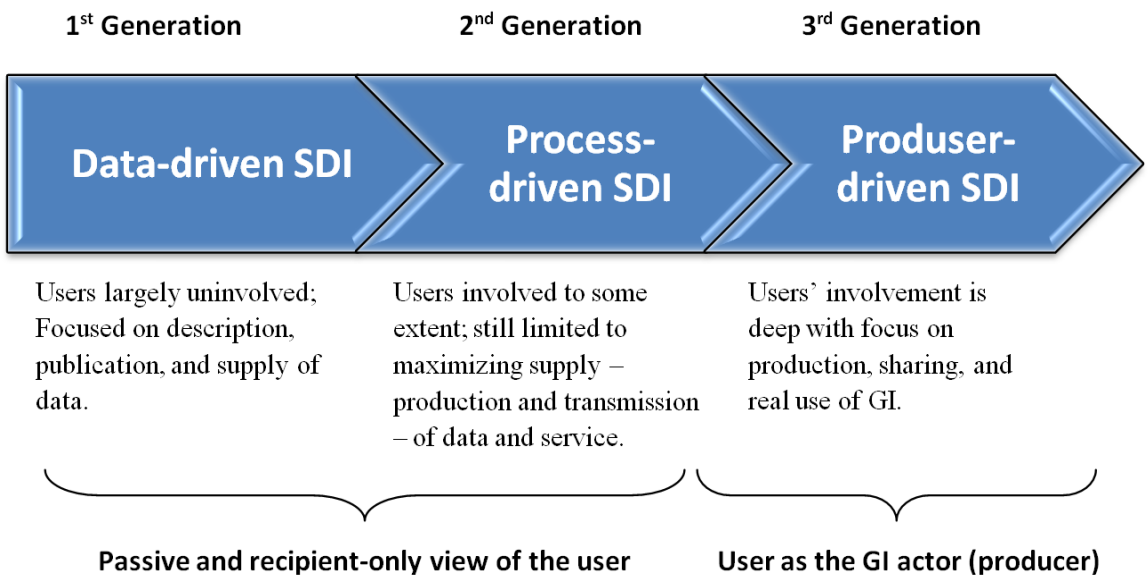


Figure 3. The evolution of SDIs: from a data-driven to a user-centric SDI.

(source Budhathoki et al. 2008)

Interestingly, it is recognised that this trend can possibly challenge the traditional role that mapping agencies have. Their institutional functionality and their authoritative control on the spatial data available in SDIs might come under fresh examination.

2.2.2.5 Local knowledge

The vision of empowering local citizens to generate spatial content for their neighbourhoods, access information and actively participate in a collaborative manner in decision making is not new (Schroeder 1996, Tallen 1999, Tallen 2000, Elwood 2002). PPGIS has been a heavily debated issue by scholars and researchers (the interested reader can see Sieber 2006 for a literature review on the subject). The cornerstone of this effort is the local knowledge of space, of geography, of human activity and its consequences that citizens with little or no experience with GIS can bring in a collaborative GIS platform. Moving a step further Tallen (1999) calls for a '*resident-generated*' GIS as a framework that will be '*constructed by rather than for local neighborhood residents*' (p.534). Tallen supports that this would be beneficial in two different ways. On the one hand, it will facilitate neighbourhood interaction as it will provide an effective communication channel for ideas, problems, expectations and opportunities regarding the neighbourhood. On the other, it will provide the research

community with unknown data regarding the inner functionality of neighbourhoods and small communities. This same vision of user participation has now re-emerged through the phenomenon of UGSC. Due to the nature of UGSC phenomenon there is an intrinsic relationship between the spatial information generated and the local knowledge with which users are predominantly geared (Elwood 2008a, Goodchild 2008b, Heipke 2010). UGSC has been recognised as a source that can describe effectively and in unprecedented way local activities and life at local levels that usually go unrecorded by the mainstream methods of spatial data collection. Goodchild (2007a, p.220) focuses on the UGSC's "*... potential to be a significant source of geographer's understanding of the surface of the Earth*", while other researchers (Sieber 2006, Miller 2006, Tulloch 2008) support that UGSC can enhance the social aspects of the GI science.

It is interesting that in both cases (i.e. PPGIS and UGSC) the goals and benefits from local users' participation who unveil, study and thus understand the everyday human behaviour in a local/community level, remain the same. On the other hand, what also remains the same are some of the challenges such as issues of ditital divide and marginalised citizens. Also, in the strain of citizen participation through PPGIS there was an active interest in improving the users-GIS interaction (Haklay and Tobon 2002). This challenge has returned today in the context of UGSC with the form of enhancing user participation, improving interaction with the spatial content and ultimately advancing content generation.

2.2.2.6 Creation of new products

Similarly to the issues raised in the previous Section, the close relationship of UGSC and users' local knowledge can be a valuable factor for creating new spatial products. For example, Craglia et al. (2008) consider UGSC as a key feature of the necessary developments towards the Digital Earth vision. Also, Hudson-Smith and Crooks (2008) present a series of innovative Web-based mapping products based on user generated content by converging advances in neogeography and social networks. Furthermore, Hanke (2007), the co-founder of Keyhole (now Google Earth), suggested during the O'Reilly Where 2.0 conference in 2007, that this is an opportunity for all of us to build "*a map of the world that I think will be more detailed, more comprehensive, more*

inclusive than any map of the world that has ever been created." Hanke did not refer simply to satellite imagery, but to "*…a map of user annotations, of descriptions, of images, of movies, of sound*". These approaches are indicative of the enthusiasm that stems from the evolution of the UGSC phenomenon and the consequent proliferation of GI.

2.2.2.7 Timely data

A particularly valuable characteristic of UGSC is the limited time needed to be collected and even more importantly the time needed to be published on the Web (and thus become accessible by everyone) compared with the traditional spatial data publishing procedures. The importance of this characteristic is paramount when it comes to emergency situations where early warnings are needed. For example, in the case of natural disasters, the availability of immediate response from the local people on the ground is of high importance in the delivery and coordination of help from response units (see for example the case of Hurricane Katrina in U.S.A.) as other sources of spatial data input (e.g. satellite images) might not be at hand for a substantial period of time especially at the immediate aftermath of such an event (Goodchild. 2007a).

* * *

Although UGSC is a fairly new phenomenon, its value and potentials have started to emerge. This explains the early interest of mapping agencies for understanding and potentially embodying such practices and the resulting spatial content in their mapping procedures. Examples of the early engagement between UGSC and mapping agencies can be found in the sponsoring of Geograph (http://www.geograph.org.uk/) from OS (Geograph 2006), the development of the MapReporter Web application from NavTeq to collect user generated data (Navteq 2010), the National Map Corps (Bearden 2007) from USGC or the "people's map" initiative in Scotland (Rideout 2007). Furthermore, in a workshop organised by EuroSDR, regarding the use of crowdsourcing for updating National Databases, the NMAs of Great Britain (Havercroft 2009), Switzerland (Guélat 2009) and France (Viglino 2009) have presented their early attempts to understand, explore, analyse and tap the UGSC phenomenon.

2.3 Interactivity and vector data transmission methods[3]

As discussed, geo-applications that aim to engage user participation have been transformed from one-directional to bi-directional communication channels (Goodchild 2007b) that allow users to have access, create and interact with their content or content that other users have already submitted, turning them into produsers (see also Section 1.2.2). In order for the new geo-applications to be able to offer the advantages of the content-level interactivity it is crucial to improve the means used to serve spatial content to the users. As this section discusses, the raster-only Web maps are not able to provide intrinsic interactivity. On the other hand, vector data encodings that are natively interactive face strong limitation when it comes to being transmitted over the Web. Thus, research into the area of interactivity needs to focus on the development of effective vector data transmission methods. This can make the emerging breed of Web maps considerably more informative and interactive and thus enable them to effectively foster further user participation and allow the flow of more UGSC. The issue is discussed in this Section.

Helped by the evolution of Web 2.0, mapping applications and spatial information is now ubiquitous on the Web. One of the clear observations though for Web maps is that the lion's share is delivered in raster-based data formats (Plewe 1997, Cecconi and Galanda 2002, Peng and Tsou 2003, Zaslavsky 2003, ESRI 2006) and thus the spatial entities presented are not interactive. This ubiquitous presence of raster maps stems mainly from the fact that transmission methods for raster data over the Web are well established and easily implemented.

Nevertheless, there are specific cases where raster images are inadequate. Researchers (Bertolotto and Egenhofer 2001, Bertolotto 2007) have pointed out the limitations of raster-only mapping applications and that in many cases the Web mapping applications

[3] This Section is adapted from:
Antoniou, V., Morley, J. & Haklay, M.M., 2009. Tiled Vectors: A Method for Vector Transmission over the Web. In J. D. Carswell, A. S. Fotheringham, & G. McArdle Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 56-71.

require the user to be able to interact directly with the cartographic entities presented on the map. Interactivity is also especially important in Exploratory Spatial Data Analysis (ESDA) (Dykes 1997, Cook et al. 1997, Andrienko and Andrienko 1999, Zhao and Shneiderman 2005), where a highly interactive environment can significantly enhance presentation, synthesis, analysis and exploration (MacEachren and Kraak 1997). In Web 2.0 though, the need for content interactivity became a key factor in the overall functionality of an application. Yet, despite the need for Web geo-applications able to host vector data, vector maps suffered from setbacks that prevented wide implementation. These problems stem from the voluminous nature of vector data and range from limitations inherent in each of the formats introduced to intrinsic disadvantages of vector encoding (such as on-the-fly generalisation and efficient transmission over the Web). To tackle the latter issue, many efforts introduced are trying to apply, with limited success in real-life applications, progressive transmission methods to vector data, similarly with the progressive methods followed for raster data (see section 2.3.2.2 for more on that).

A productive way to understand the implications that a Web 2.0 geo-application with limited interactivity in the content level (i.e. in the map presented to the users) has on the phenomenon of UGSC is to analyse its limitations.

2.3.1 Limitations of raster-only maps

The raster encoding provides clear advantages and therefore it is not surprising to see that it dominates the area of Web geo-applications. On the surface raster-only maps are hugely successful and indeed major commercial players have demonstrated that all that is needed to deliver spatial information to a very large number of users is to serve only pre-prepared raster tiles. However, a more detailed examination reveals that there are several limitations posed by the use of raster-only maps.

The term ‘interactive map’ seems to be used in an uncritical way. Web maps with basic functionality limited to zoom and pan easily gain the title of interactive map. However, true interaction requires that individual elements represented are responsive (Neumann

and Winter 2004). Thus, a truly interactive map has to provide the ability to interrogate potentially every object monitored by the mapping application. Given the nature of their encoding, raster maps cannot be natively interactive and any hosting of object level interaction is cumbersome. Raster-based maps try to achieve interactivity by bypassing the problem either by overlaying vector data representing features of interest or by tracking the "active" thematic layer of the application and the mouse coordinates corresponding to user's clicks. These parameters are then transmitted back to the server and any information associated with the object is returned to the user. This is an important limitation as it can cause considerable delays when there is intense spatial data input as in the case of collaborative GIS (Haklay 2006). A raster-only map hinders the application from direct accessing the elements that compose the map without further communicating with the server, and thus paying the price of network latency and server processing for any request relative to the content of the map. Due to the additional interaction with the server, the level of interactivity is diminished. Web maps need to be natively interactive to propagate high levels of interactivity to their users.

Apart from the limited interactivity to the end users, a major drawback for raster maps is that they serve inflexible content that creates obstacles in the communication process between the author and the users. A map is a graphic representation of geographical objects (Robinson et al. 1995) which the cartographer chooses to show. When the map is served in a raster format, objects lose their existence because the content of each entity is embodied into the pixel-based structure of the raster and cannot be changed. As MacEachren (1995) points out, by examining the communication process of cartography it can be seen that there are a number of filters and obstacles that information has to pass through, from the map author to the map and then to the map user. In this context MacEachren suggests that '*we can improve map communication if we can reduce the filtering or loss of information at various points in the system*' (p.5). Embodying individual spatial entities in a raster file which then is presented to the user is an intense form of filtering both at the author's and the user's level that hinders the communication and is not much different conceptually from the process of printing the map on paper. Moreover, Wood (1994) argues that if the map offers the option to change its content instantly it will make both a quantitative difference in the number of things that a user can make visible and a qualitative one in the thinking mechanism of the user. Going a

step further, Andrienko and Andrienko (1999) suggest that maps are not only a communication medium but also act as tools '*to support visual thinking and decision making*' (p. 357). It is clear that the role of map as a medium for communication, thinking and decision making is hindered by the lack of interactivity in the content level. Web geo-applications can meet all these challenges by modelling and transmitting to the user discrete spatial entities. Additionally, as Duce et al. (2002) point out, with the advent of XML technologies the practice to transmit images to render entities with semantic content is likely to decrease. Indeed, there are early efforts of modelling and transmitting XML-based spatial data over the Web (see for example Neumann and Winter 2003, Zaslavski 2003, Antoniou and Tsoulos 2006, Antoniou and Morley 2008). A step further, Antoniou et al. (2008) suggested that the case of using only the strong points of the raster data overlaid with vector data provides the means for spatial entities to host directly scripting, animation and attribute data allowing instant interaction between the user and the map.

Another limitation of the static nature of raster-only Web maps is that they considerably limit the functionality that a Web geo-application can provide to the user (ESRI 2006). Thus, there is a need to introduce a new layer between the map and the user to host the missing functionality. Major commercial map providers (e.g. Google, Microsoft, or Yahoo!) have themselves experienced the limitations of the raster-only Web maps when it came to enhance the functionality of their mapping applications by building more task oriented services. Such an example is Google's routing service where the Web mapping application responds to the user's inputs and calculates the shortest path. The only data returned to the client is the path in vector format avoiding the need to refresh the whole map while offering the user the ability to customise the path interactively by adding route constraints. The alternatives would be either to create and serve another layer of transparent raster image with the path or overlay the vector path on the raster image, rasterise the layers into a final image and serve it to the client. While both options are technically feasible, they are slow in response time, resource hungry and deprive users from the route customisation option.

Finally, there are cases where it is necessary for a Web application to exploit the intrinsic characteristics of vector data. Buttenfield (2002) presents a number of GIS

modelling and analysis tasks such as power, transportation and telecommunication routing models that rely on specific vector properties. The author also raises the need for vector delivery to support methods that will allow map content to be kept up to date in real time applications - which is the case of the majority of critical applications. In such applications it is common to incorporate and disseminate real time data (for example a current location of a vehicle) and this mandates the ability of the Web mapping application to handle vector data. In other cases there is need to handle and manipulate directly cartographic features while maintaining their topological and metric properties (Bertolotto and Egenhofer 2001).

2.3.2 Limitations of vector maps

Given the discussion on the limitations of raster-only maps, an obvious question raised is why is there so limited presence of vector maps as they can fill these gaps and provide the much needed interactivity. The answer lies to the variety of intrinsic limitations that vectors themselves have. Interestingly, the significance of some of those limitations is further enhanced when the vector data are used in the Web environment. Consequently, this further deters the adoption of vector maps for Web 2.0 applications.

A prime example is the lack of browsers' native support for vector data. The omission of vector data from the Hyper Text Mark-up Language (HTML) specifications had as a consequence the development of browsers that did not provide native methods for parsing and rendering vector-encoded content. Up until recently, this generated the need for a plug-in to be present on the client's machine or one to be downloaded and installed by the users for the content to be properly rendered. An interesting de facto exception is the Flash format that has been introduced in 1996 by Macromedia (now acquired by Adobe). Flash is a format with continuing success among users and developers. Flash is used predominately for Web development and particularly for animated Web advertisements. Although in the context of Web mapping Flash has a number of limitations (Neumann 2002 and Held et al. 2004 provide extensive evaluations of the format against general and cartographic criteria), researchers have experimented in developing Web mapping applications. For example, Steiner et al. (2001) used Flash as

a platform to implement dynamic and linked exploratory geospatial data analysis methods. Importantly, although a browser plug-in is needed to properly render Flash, most users have it preinstalled with their Web browsers and from the user perspective it seems that Flash is natively supported. Nevertheless, from the developers' perspective, the fact that Flash is a proprietary format makes investing or committing to such a format not always a welcomed option (see for example Apple's attitude towards Flash).

Things changed dramatically the last few years, with the appearance of Scalable Vector Graphics (SVG), an open XML-based vector format supported by World Wide Web Consortium (W3C) (W3C 2001). Despite the fact that SVG was natively supported by some browsers (e.g. Firefox, Safari, Opera and Chrome) and it was embraced by developers and researchers as a new and promising vector format, able to give a boost to high quality and interactive Web mapping applications (Peng and Tsou 2003, Neumann and Winter 2003, Peng and Zhang 2004, Dunfey et al. 2006, Williams and Neumann 2006a), the format did not manage to gain widespread acceptance mainly because of Microsoft's denial to natively support it in its browser. Yet, factors like the increased interest for interactivity (see for example the discussion in Sections 1.2.4 and 1.3.2) and the proliferation of touch-screens mainly in mobile devices but also in personal computers (PCs) and tablet PCs as well, or the critical remarks of pioneering figures like Sir Tim Berners-Lee on Microsoft's slow progress in supporting SVG (Svensson 2008) played a role in making Microsoft natively support SVG in its new browser. Although such corporate decisions are out of the scope of this research, the native support of an open vector format from all major Web browsers creates a highly promising environment for vector data on the Web.

Apart from these external to vector data factors, there are more endogenous reason that restrict their broader adoption. One of the most intensively researched subjects around vector formats for Web mapping is dynamic generalisation. Because vector data is voluminous (for example a sample dataset in 1:1250 scale from OS MasterMap for an area of 25km^2 in the suburbs of London is 706 MB in GML or 315 MB in shapefile format), when the data is viewed at a small scale there is an advantage in providing data which is generalised to the appropriate scale. Also, as Buttenfield (2002) points out, despite advances in broadband access, technological improvements have led to more

detailed data collection resulting in increased file sizes and thus, they need longer to transmit. Mackaness (2008) suggests that data redundancy, storage efficiency, exploratory data analysis, data integration and paper map production are among the reasons to provide dynamic generalisation.

The need for vector data on the one hand and the objective inability to send huge amounts of data over the Web on the other, drove researchers in the quest of efficient vector data transmission methods with limited success up to now (see Section 2.3.2.2 for more on that).

Finally, the open nature of XML technologies becomes a drawback which poses a strong barrier when it comes to preserving intellectual property rights (IPR) of the map producer. Because the XML specifications use plain text to encode information, which is easily accessible to humans and machines, an XML based vector format means that the client's machine will have access to raw spatial information. This fundamental characteristic of XML leaves little space for protecting the intellectual rights of spatial information. This is a significant issue for data providers as they might be unwilling to migrate to technologies that could lead to the compromisation of their IPR (but see also Section 2.2.1.4 for more on the issue of IPRs).

Summarising the current environment for vector data on the Web it is clear that the main obstacle for a broader adoption of such data is the introduction of efficient vector data transmission methods over the Web. Indeed the problem has drawn the attention of researchers but so far no applicable ways to achieve efficient transmission of vector data has been developed. But before examining the methodologies proposed for vector data transmission over the Web it will be constructive to briefly examine the available methods of raster data transmission. This is mainly because the vector methods have tried to imitate the successful solutions applied to the raster data.

2.3.2.1 Raster data transmission

Many progressive transmission techniques for raster data have been introduced. Bertolotto and Egenhofer (2001), Yang et al. (2005) and Bertolotto (2007) offer reviews

of those techniques and describe in brief their main characteristics. In general, highly sophisticated compression algorithms and interleaving transmission methods made raster formats suitable for data delivery over the Web. The efficiency of these methods enabled the Geomatics community to build the majority of Web mapping applications using raster data and thus easily deliver spatial information to the users.

The evolution of the Web and the pursuit of enhanced responsiveness and increased usability for Web applications lead to the development of new programming techniques like AJAX and a new method of raster data transmission that uses tiled raster images and different levels of detail (LoD). According to this method the highest LoD of the mapping area is divided into a number of quadrants (tiles). Each of these quadrants is further sub-divided into new tiles that form the next LoD. This process continues until the lowest LoD is reached. Although this can lead to a huge number of tiles for a detailed or large dataset, the storage, indexing and handling of raster files is straightforward, especially as data storage becomes cheaper.

When a map of a given LoD is requested a number of tiles are sent to the user. The tiles are loaded into the Web browser window as a matrix, and from the user's perspective it seems to be a continuous image. For any consequent request such as pan, zoom or managing layers a new request is made by the application that runs on the client's Web browser and the server transmits to the client only the tiles that are needed in addition to the ones currently in the client's cache. This method makes the application considerably faster and more responsive since the use of AJAX techniques reduces the application's response time by communicating with the server without the user actually noticing it.

The tile-based method for raster data delivery has been successfully implemented by major mapping providers like Google, Yahoo! and Microsoft. In fact, the explosion of mapping applications on the Web and the consecutive phenomena of map mash-ups, neogeography and VGI have been based on the efficiency and ease of raster data delivery using the tile-based technique.

2.3.2.2 Vector data transmission

The volume of vector data and the difficulty of transmitting it over the Web has been a long standing problem for Web mapping and Web GIS. The success of the progressive transmission methods for raster data turned the focus of research towards the development of similar techniques, tailored to vector encoding. The aim of progressive transmission is to alleviate the user from long waits for the complete dataset to be downloaded before accessing the data (Bertolotto and Egenhofer 2001, Zhou and Bertolotto 2004, Yang et al. 2005).

Efficient methods of progressive transmission have been introduced for a particular case of vector data: triangular meshes. Triangular meshes are usually used to describe digital terrain models or the surface of 3d objects. On the contrary, progressive transmission of cartographic vector data over the Web remains problematic despite numerous efforts (see Bertolotto and Egenhofer 2001, Bertolotto 2007 and Yang et. al 2007 for a review). According to progressive transmission methods a coarser map version is sent initially to the user, and depending on the user's requirements, consecutive data packets are sent to improve the map. The coarser map versions can either be generated dynamically at the time of the request (on-the-fly) or can be pre-calculated through the process of generalisation. On-the-fly generalisation is an unsolved problem for cartography. A number of researchers have focused on dynamic generalisation (Cecconi and Galanda 2002, Lehto and Sarjakoski 2005, Jones and Ware 2005) to enhance progressive vector transmission but since there are no formalised cartographic generalisation principles and applicable generalising operators (Weibel and Dutton 1999), automated dynamic generalisation still remains a challenge. Moreover, the existing generalisation algorithms are time-consuming and thus not applicable for real-life Web application. Another disadvantage of dynamic generalisation is that it produces inconsistent results in terms of retaining topological and geometric attributes and thus often need an a posteriori evaluation of their consistency. A topologic consistent approach has been proposed for polygons and lines by Yang et al. (2007) with the exception of isolated polygons that have area smaller than a given threshold and lines that belong to the smallest category (for example first-order streams in a river network).

On the other hand, off-line generalisation and the creation of different LoDs is the norm for mapping agencies. In this case, generalisation is time-insensitive and is usually performed interactively by expert cartographers with the help of specialised software (Cecconi and Galanda 2002). Bertolotto and Egenhofer (2001) presented a method for pre-computing and storing multiple map representations suitable for progressive transmission to the user but without further implementation. Although the maintenance of different LoD is cumbersome, off-line generalisation yields topologically and geometrically accurate products.

The use of progressive data transmission relies on the fact that users can perform preliminary operation even on a coarser version of the map or they can assess the suitability of the map requested and possibly change their request without waiting for the whole dataset to be downloaded. The process is shown in Figure 4, where A, B, C, D1 and D2 are various time periods of the process. Period A is the time during mouse move, B is the time that the user waits for the coarser version of the map to be loaded, C is the time the server needs to extract the data. D1 is the time the user observes the map while it becomes more detailed and D2 is the time that the user observes the fully detailed map. Progressive transmission enables users to start observing the map before the whole map is downloaded.

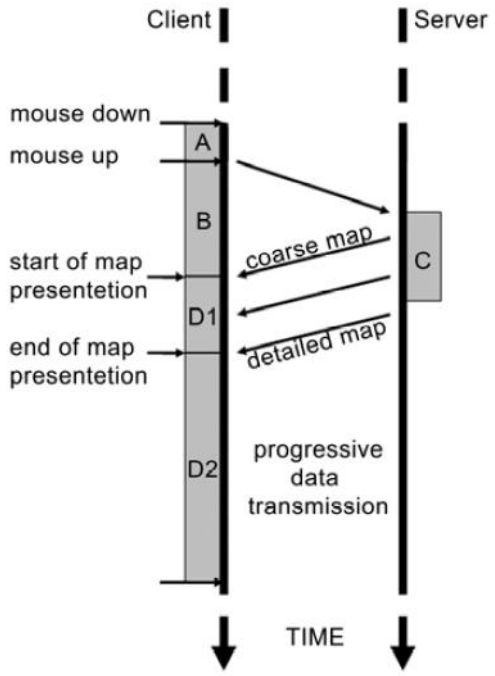


Figure 4. Steps and time periods in the progressive transmission of vector data

In Section 1.2.4 the focus was on the importance of interactivity for the evolution of Web 2.0 and in Section 1.3.2 in the effect that this has on UGSC. It was explained how interactivity plays a catalytic role in user participation and serves as an accelerator for content generation for both spatial and non-spatial Web 2.0 applications. Here the focus turned on the existing efforts for providing interactivity to a Web 2.0 geo-application by presenting vector data to the users. It turns out that there are fundamental obstacles when it comes to efficient transmission of vector data over the Web. Despite the efforts for the creation of an efficient vector data transmission method, the fact remains that there is no practical implementation of them in real life applications. This is mainly because the time that these solutions need is well beyond the average response time of a Web application. Also, specialised database structures or topologically and cartographically incorrect results hinder further development and implementation of these efforts. Thus this issue remains an unsolved challenge for the advance of Web 2.0 geo-applications.

2.4 Spatial data quality and UGSC

2.4.1 General

The scope of this section is not to provide an extensive literature review on the subject of spatial data quality. Quality in Geomatics is a subject that has drawn the interest of researchers for quite a long time now. The interested reader can find a collection of papers in Shi et al. (2002), Devillers and Jeansoulin (2006) and an extensive literature review in Van Oort (2006). Furthermore, the reader can turn to the International Organisation for Standardisation (ISO) that provides a series of Standards and Technical Specifications that focus on the concepts, the descriptors, the evaluation and the reporting mechanisms of spatial data quality. For example, ISO 9000:2005 provides the basic definitions and explanations around the quality concept in general. A more spatially-related approach on quality comes from the ISO Technical Committee 211 (ISO/TC 211) that caters for quality issues on spatial data. The ISO/TC 211 has published the following Standards and Technical Specifications:

- **ISO 19113 - Quality principles**. The Standard provides a set of terms and definitions regarding the quality of spatial data and the principles for describing and reporting the quality of spatial data.

- **ISO 19114 - Quality evaluation procedures**. The Standard provides a set of procedures for evaluating and reporting the quality of spatial data.

- **ISO 19138 - Data quality measures**. The Technical Specification extends the ISO 19113 by providing a set of data quality measures.

Nevertheless, a discussion of the fundamental principles of quality and quality management concepts that are relevant to the subject of this Thesis will be provided. Then the focus will turn in examining how these concepts have been applied to the phenomenon of UGSC and examine whether UGSC creates new challenges in the subject matter of quality. Finally, a review of the up-to-date efforts to evaluate the quality of UGSC sources will be discussed.

2.4.2 Spatial data quality: definitions and concepts

Chrisman (2006) provides an eloquent review of the evolution of spatial data quality from the early narrow-minded conceptualisation that quality fully coincides with positional accuracy up to the recent developments and standards on spatial data quality. Chrisman explains that along with the evolution of the concepts around quality, there was also a major shift regarding who is responsible to interpret the quality of a spatial dataset. At the beginning, the author of a spatial product (i.e. mostly paper maps) was responsible to inform the users whether the product complied with pre-defined standards and thus whether it was acceptable to be used for certain purposes. This was judged by examining the product against specific thresholds: if the product tested within the threshold it complied with the standard. Today, with the flexibility and the ease in the dissemination of digital information, a threshold-based treatment of spatial data quality is not practical. The producer of a spatial dataset simply cannot foresee all the possible

uses of its data. Therefore the producer is not making any judgments about the usability of its products, but rather reports the results of a series of tests that the products undergo. These tests include the examination of the final product against its specifications. Consequently, there is a need for data users to play a more active role in the evaluation of the spatial data quality especially under the prism of the intended use. In other words the responsibility of the judgment shifts to the users who need to determine the *fitness-for-purpose* of a specific spatial dataset.

Not much different to the latter conceptualisation is the approach suggested by the International Organisation for Standardisation. According to the ISO 9000:2005 Standard (ISO 2005d), quality is the "*degree to which a set of inherent characteristics fulfils requirements*" (p.7). As the terms *characteristics* and *requirements* are vague the specification provides further explanation for both. *Characteristics* (or more commonly known as quality elements) are defined as distinguishing features of a product that can be either inherent or assigned, and can be either qualitative or quantitative (see Sections 2.4.2.1 and 2.4.2.2 for more on these elements).

On the other hand, *requirement* is defined as a need or an expectation that is stated, obligatory or generally implied, where "*generally implied means that it is custom or common practice for the organization, its customers and other interested parties, that the need or expectation under consideration is implied*" (p. 7, ISO 2005d). As far as the producer is concerned, these requirements are realised with the help of specifications and guidelines. But these requirements also bring into the frame the importance of how the final user of the product has conceptualised the functionality and the quality of the product.

These two different conceptualisations of quality (i.e. the producers' and the users') are known as internal and external quality and are further discussed in the next Section.

2.4.2.1 Internal and external spatial data quality

Any spatial dataset is an abstract model of the real world. The creation of this model is based on a set of specifications, put forth by the data producer, that describe the

abstraction process from the real world to the model. This model, more commonly referred to as '*universe of discourse*' (p.14, ISO 2005a), once created, is used to examine the conformance, and thus the internal quality of a dataset (Devillers and Jeansoulin 2006). On the other hand there is the concept of external quality that corresponds to the degree of conformance between the users' needs and requirements and the spatial dataset provided by the author (Devillers and Jeansoulin 2006). In simple terms, the internal quality reveals the quality of the final product from the point of view of the producer, whereas the external quality reveals the quality of the final product from the point of view of the user (Figure 6).

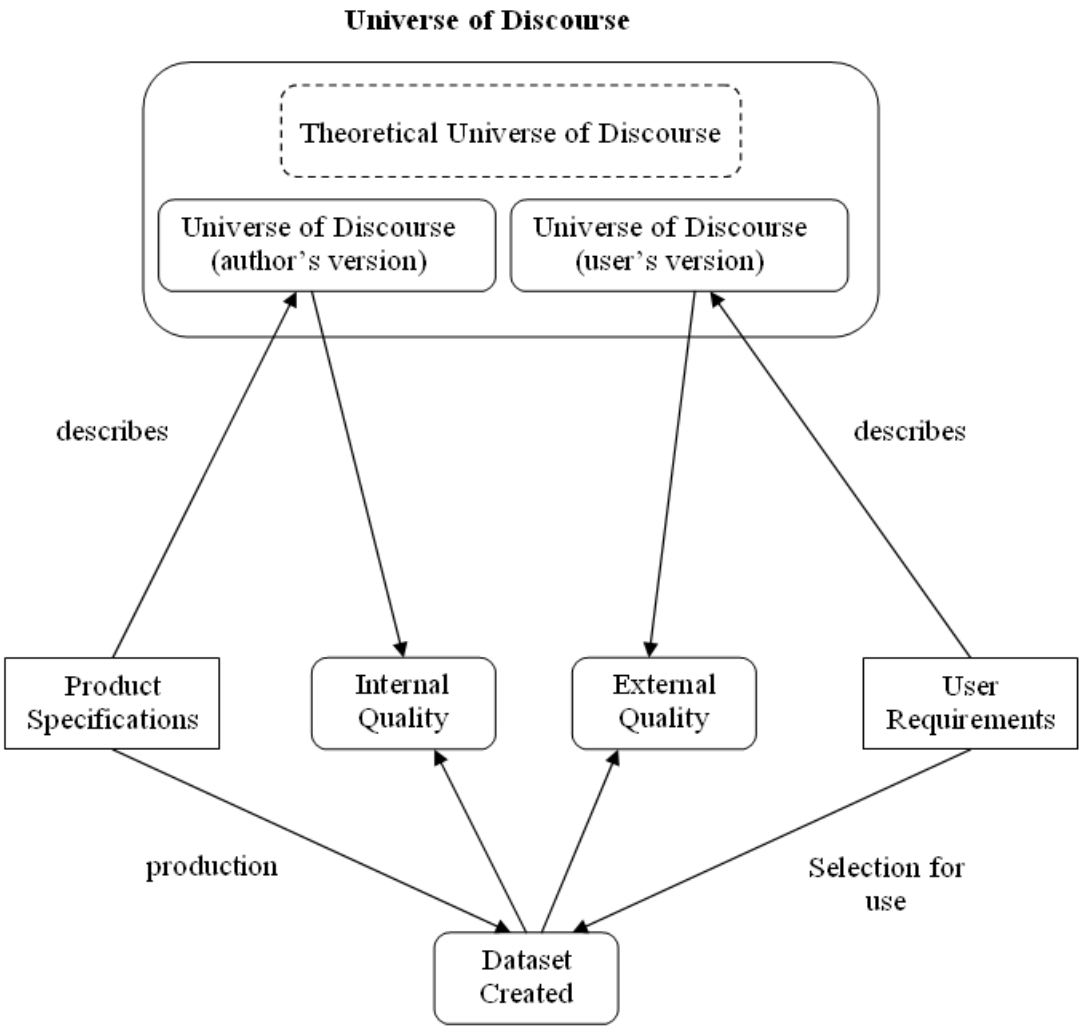


Figure 5. The concepts of internal and external quality.
(source ISO 2005b)

2.4.2.2 Spatial quality elements

Of high importance to both internal and external quality are the method followed and the means used for the quality evaluation of a spatial dataset. The subject matter of these evaluations is the conformance of a dataset against a set of spatial quality elements (i.e. the product characteristics). There have been many efforts (Table 1) to successfully define these quality elements for spatial datasets. The importance of these elements is easily understood as they are the descriptors of different aspects of a dataset's quality and thus function as the components of the overall spatial data quality.

Quality Element	Aronoff (1989)	USA SDTS (1992)	ICA (1995)	CEN TC287 (1998)	ISO (2005)
Lineage	E	E	E	E	E
Positional accuracy	E	E	E	E	E
Attribute accuracy	E	E	E	I	E
Logical consistency	E	E	E	E	E
Completeness	E	E	E	E	E
Semantic accuracy			E	E	
Usage, purpose, constraints	E	E		E	E
Temporal quality	E	E	E	E	E
Variation in quality		I	I	E	I
Meta-quality		I	I	E	I
Resolution	E	I	I	I	I

Table 1. List of Quality Elements from different sources (E = explicitly recognised as a quality element, I = implicitly recognised as a quality element).
(based on Van Oort (2006); updated for the ISO 2005b)

The literature around spatial data quality provides extensive analysis of these elements (see for example ISO 2005b, Van Oort 2006, Devillers and Jeansoulin 2006, Servigne et al. 2006) and their further analysis on sub-elements. Here a basic presentation of the explicitly stated elements in the ISO Standards is provided as these quality elements will be used in the course of this Thesis:

Completeness: refers to the presence of data that fall out of the scope of the universe of discourse and thus there are errors of commission and to the absence of data that fall within the scope of the universe of discourse and thus there are errors of omission.

Logical consistency: refers to the degree of the dataset's adherence to the logical rules (conceptual, domain, format and topological consistency) provided by the product's specification.

Positional accuracy: refers to the geometric accuracy (absolute, relative or gridded accuracy) of the position of the captured features.

Temporal accuracy: refers to the accuracy of the temporal attributes and temporal relationships of the captured features.

Thematic accuracy: refers to the accuracy of the attributes recorded for each captured feature with the exception of the positional and temporal attributes.

Purpose: refers to the rationale for creating the dataset.

Usage: refers to the known application(s) that have used the dataset. The uses can origin either by the dataset's producer or the dataset's users.

Lineage: refers to the history of a dataset including the collection and compilation processes followed.

2.4.2.3 Spatial data quality evaluation process and methods

Up to now the discussion was about the basic definitions and principles of quality and the elements used to describe the quality of a dataset. Now the focus turns to the process and the methods followed to perform quality evaluation of a spatial dataset. The process for performing a spatial data quality evaluation, as described by the ISO (ISO 2005c), follows a series of basic steps (Figure 6).

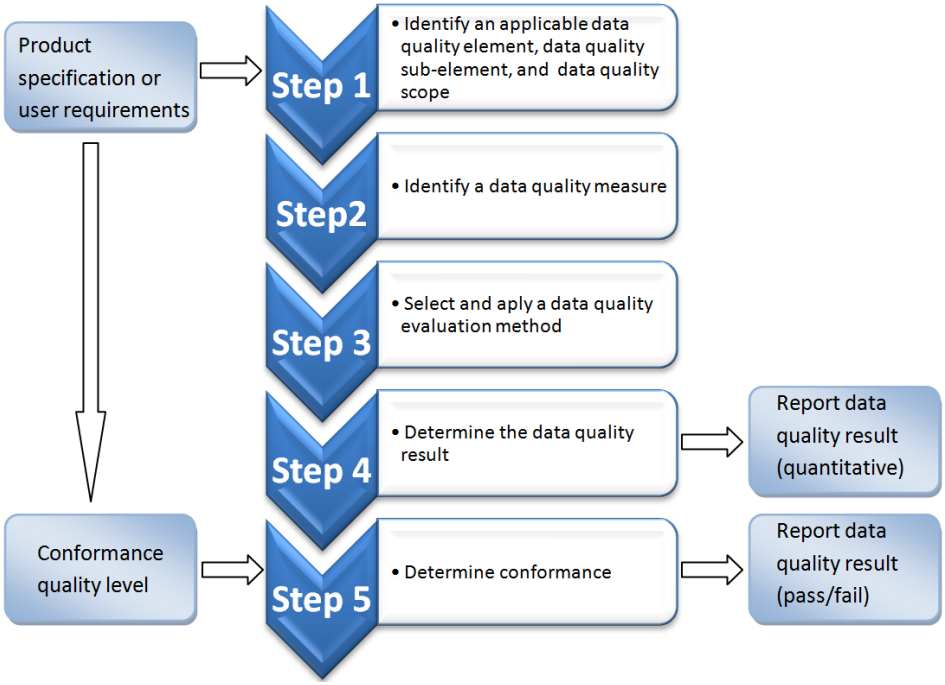


Figure 6. Spatial Data Quality Evaluation
(source ISO 2005c)

Initially, an applicable data quality element, its sub-elements and the data quality scope (i.e. which features this evaluation concerns) are identified. Next, the data quality measure (e.g. the number of excess features in a dataset) is selected and the actual measurement of the dataset's conformance is implemented with the help of an evaluation method. Finally, the measurement's results are recorded and if the user's conformance level of acceptance is available, a conformance judgment (i.e. pass or fail) can be provided.

Central to this quality evaluation process is the evaluation method followed (step 3 in Figure 6 and Figure 7) that can be either direct or indirect. The latter method uses available information for the data (e.g. lineage or known uses) to give an estimation about the overall quality of the dataset. In contrast, the direct evaluation method evaluates quality through direct comparison with internal or external data. For example, all the data needed for a topological consistency examination of polygon closure are available in the polygon datasets and thus such an examination is internal. On the other hand, a dataset's completeness evaluation requires an external dataset (reference data), against which the evaluation will take place.

Figure 7. Classification of data quality evaluation methods
(source ISO 2005c)

Finally, two more parameters have to be considered during the quality evaluating process of a spatial dataset. The first is relevant to the data population used in the evaluation. Either the full dataset will be examined or a proper sampling method needs to be adopted for the selection of the sampling data (the interested reader can see Annex E of ISO 19114:2005 for more on the sampling methods). The second depends on the update frequency of the dataset under evaluation. In a static or infrequently updated dataset the process is straightforward as the dataset is used for the evaluation with no further action. In the case of dynamic dataset though, where the data receive updates frequently (as is the case with the popular UGSC sources), a copy of the dataset is acquired and the evaluation takes place as if the dataset was static. This benchmark procedure can be repeated periodically and the evaluation results refer to the copy creation date. Alternatively, a continuous process can be adopted where the evaluation focuses on the impact that the updates have to a dataset of known quality. This method requires the evaluation procedure to be embedded in the data creation process.

2.4.3 Quality issues for UGSC

As explained thus far, *quality* is an issue that has drawn the attention of the scholars and of the Geomatics industry as it is directly affecting the use of spatial data. In parallel, it has been explained how the evolution of Web 2.0 has affected the Geomatics domain resulting in the emergence of UGSC. A variety of data is now available on the Web that are directly or indirectly associated with location and thus their examination under the

prism of quality assessment is a legitimate topic (Goodchild 2008c). In fact, well above the legitimacy of such a discussion is the need that arises for understanding the quality and thus the fitness for use of UGSC. Many researchers (Elwood 2008b, Goodchild 2008b, Sui 2008, Ather 2009, Haklay 2010, Antoniou et al. 2010a, Grira et al. 2010) point out that the need to understand the quality of UGSC will become increasingly pressing as the growth in the UGSC volume will keep rising. Goodchild (2008b) explains that the neo-geographers that spring out of the Web 2.0 culture have very little knowledge of the basic principles of GI, cartography or spatial accuracy and consequently, the vast majority of the popular Web 2.0 geo-applications show no or little concern about accuracy and data quality. *Or is it also the other way around?* Can it be possible that the Web 2.0 geo-applications available are leading their users in an environment that is ignorant of the quality's importance, and thus the blame of limited quality awareness should be put more on these geo-applications and less on the lay users? For example EveryTrail (www.everytrail.com) is an application that allows its users to share GPS trails. Figures 8a and 8b show two different GPS trails that contain fuzzy geometric segments. Are the users or the GPS devices to be blamed for these obvious geometric errors? It stands to reason to expect that the geo-application should be intelligent enough to recognise such patterns, highlight them as potential errors, and provide the means to the users to correct them. In the course of this Thesis it will be supported that there is a need to engraft the Web 2.0 geo-applications with the culture of spatial data quality.

(a) **(b)**

Figure 8. Fuzzy segments in the trails' geometry

Given the abundance of GI in the new Web 2.0 environment Goodchild (2008c) concluded a discussion about quality with the question "*What should spatial accuracy assessment mean in a world in which everyone is a potential user of geospatial data?".* Although this question is successfully raising the issue of the spatial data quality in the context of Web 2.0, it is failing to cover the whole picture of the challenges regarding the issue. A more inclusive question would be: "*What should spatial accuracy assessment mean in a world in which everyone is both a potential producer and a potential user of geospatial data?*". Still, such a debate would be part of a broader discussion regarding the issue of quality evaluation and management of UGSC. More specifically, on the one hand, the interest should focus on the successful implementation of the accumulated knowledge and experience about spatial data quality on UGSC. On the other, and more importantly, the particularities in the quality management that spring from the new collaborative environment of spatial data generation should be identified and analysed. This will provide valuable insights on UGSC and enable the modification, completion or adjustment of the existing practices in quality management to accommodate the new breed of spatial data.

An eloquent example of the possible particularities in UGSC quality management can be found when examining the ISO 19113:2005 (ISO 2005b) International Standard *"Data quality elements, together with data quality sub-elements and the descriptors of a data quality subelement, describe how well a dataset meets the criteria set forth in its product specification and provide quantitative quality information"* (p. 4). In this description it is assumed that there is a rigid product specification that enables the data quality evaluation through the examination of the data quality elements. However, this assumption does not hold true in most of UGSC sources as the users first create the content in an unconstrained environment (Antoniou et al 2010, Grira et al 2010. Brando and Bucher 2010) and then emerges the challenge to assess its quality.

Nevertheless, the fundamental principles of spatial data quality are still applicable to UGSC. For example, the quality elements of completeness, positional accuracy or attribute accuracy are as relevant for UGSC as for the data created through formal procedures by mapping agencies. For example, Ather (2009) in an effort to examine the positional accuracy of OSM data in four test sites in the city of London UK, used the

methodology described by Goodchild and Hunter (1997). Moreover, as is the case of traditionally created spatial data, the successful management of the UGSC uncertainty requires the identification of the sources that introduce this uncertainty (Grira et al 2010) and thus deteriorate the overall data quality.

The interesting point here is that there are uncertainty sources for UGSC that are quite different from the well-documented error sources encountered in the traditional spatial data creation processes. In that context, efforts have been made to understand the role of space in the completeness of UGSC datasets. An evaluation of OSM road network of England (Haklay 2010) and Germany (Zielstra and Zipf 2010) using reference data from OS and TeleAtlas respectively has shown similar results. Urban areas receive a better coverage than rural areas. Also, the role of socio-economic factors (Haklay 2010) in the completeness of OSM data for England was examined. By using deprivation indexes the author showed that poor and marginalised areas receive less coverage by OSM users and thus the data completeness is negatively affected. Basiouka (2009) examined whether the number of contributors plays a role in the positional accuracy of OSM data (see also Section 5.4.6 for more on that). Goodchild (2008c) notes that statistically it is expected that the positional accuracy will increase in accordance with the user's participation (as the final position will be the average of many measurements), but this quality improvement process is not applicable when spatial dependencies disallow independent partial corrections or when it comes to discrete data such as attributes. The importance of attribution in UGSC and the effect that a wiki-based process of data capturing (as is the case in OSM) has on the attribution accuracy will be discussed in the course of the Thesis. Another interesting case in point is the local knowledge or *space of familiarity* (Goodchild 2009) that has been recognised as a highly influential factor for UGSC quality (though still in a theoretical level). While for the production process of most NMAs the importance of local knowledge is almost non-existent (an exception to that is the geographical names datasets where local knowledge is a valuable tool), for UGSC its importance seems to be paramount (Goodchild 2007a, Flanagin and Methzer 2008). Researchers (Goodchild 2009, Grira et al 2010) have supported that it is expected that the data quality will rise in accordance with the level of the user's familiarity with the collected data; or in other words, poor user familiarity with the geography of the space in scope could result in data of low quality.

Early theoretical suggestions have been described in an effort to address these, particular to UGSC, challenges. For example, Coleman et al (2009) suggest a user evaluation based on their purposes and Elwood (2008b) describes a reputation-based rating system for the contributors. Furthermore, Goodchild (2009) investing on the local knowledge factor, describes a mechanism that classifies users into a geographic hierarchy of expertise based on their location. Such a scheme could be used to evaluate user contributions as each new entry would be examined by users that are familiar with the area. Section 6.3 presents a mechanism that provides a way to build inside a Web 2.0 geo-application the necessary formalisation that will enforce the contribution of high quality data. Along the same vein, Brando and Bucher (2010) describe a system that will incorporate both reference data and product specifications to ensure the data quality during the contribution phase.

Apart from these new influential factors affecting the UGSC quality, another issue that is of high importance for UGSC has drawn the interest of researchers: communicating the UGSC quality (see a further analysis on the subject along with proposed solutions in Section 7.5). The task of communicating quality for the spatial data that are created by mapping agencies through traditional methods, has been addressed with the help of the metadata which is a well accepted and fairly established mechanism (see for example the ISO 2003 and ISO 2007 Standards). In contrast, in the context of UGSC there has not been a similar mechanism to effectively share quality information for the data created. Consequently a communication gap between data producers and data users exists (Goodchild 2008c, Maue and Schade 2008, Grira et al. 2010). To bridge that gap, Goodchild (2008c) suggests a move to *metadata 2.0* that will be user-centric and application-specific. The quality information could be gathered by the end-users of the data using Web 2.0 tools like wikis, and it could be referred to specific usages and applications. Of course, this suggestion assumes that the same willingness that exists for the data collection by lay users will also be present for the quality information gathering. Similarly, Antoniou et al. (2010a) and Gira et al. (2010) recognising the problem suggest a more active users' involvement. Interestingly enough, there seem to be a consensus that this user involvement can be triggered and supported by the introduction of new *interactive* tools that will facilitate the contribution of high quality

content. Grira et al. (2010) have described an interactive communication process that will visualise spatial quality and prompt users to contribute to the quality documentation. Section 6.2 presents an interactive mechanism that initially enables users to visually understand the underlying quality of the data that are either contributing or using and then trigger their active participation for data quality improvement through a formalised process.

Finally, and notwithstanding the acceptance of the metadata mechanism, there has been scepticism over its ability to communicate quality information to untrained users (Goodchild 2007c). The suggestions and solutions discussed earlier for the UGSC generates exactly the opposite limitation. They are not sufficient for communicating the overall quality of a dataset from a UGSC source to mapping agencies. The mechanisms to facilite quality improvement discussed earlier are based on further user involvement as they provide interactive tools to engage users to the quality improvement task. Despite the fact that this would be a major leap forward for UGSC usability, still it does not bridge the gap that exist when there is a need to communicate the data quality from the crowd-based sources to mapping agencies (for more on that see also Section 7.5).

2.5 Summary

During the literature review the discussion focused on three seemingly distinct but in fact closely related subjects:

i) The nature of UGSC. The initial reaction of scholars and researchers on the subject matter of user generated spatial content was presented. This included both the scepticism that naturally followed this rapid change on the Geomatics world and the benefits that could possibly stem from the evolution of the phenomenon. The fact that the phenomenon is going through the early steps of its evolution has as a result the literature to be mostly in the sphere of conceptual and theoretical analysis and assumptions. Little is the knowledge provided by empirical studies and thus a need to bridge the gap emerges by corroborating or rejecting the various hypotheses presented.

ii) Content-level interactivity. Among the driving factors that can boost participation and content generation for this Web-based phenomenon, this Thesis has chosen to focus on the content-level interactivity. As explained, interactivity can play a vital role in the expansion of the phenomenon by further engaging users to interact directly with spatial data. This requires the spatial content to be able to natively support such type of interaction, something that cannot be achieved by the raster-based geo-applications available today. On the other hand, vector encoding that can natively support interactivity is facing serious obstacles when it comes to Web environment. Therefore, the focus turns into the exploration of efficient vector data transmission methods over the Web. Solving this problem will instantly provide the means to build content-level interactive geo-applications.

iii) UGSC quality. After examining the UGSC phenomenon, analysing its possible disadvantages and highlighting its strong points, what followed was the research on a method that could further enhance content generation. Of high importance is the issue of that content's quality as it was made clear that mapping agencies are in need of quality patchworks to nurture both their existing mapping products and to develop new ones. From the rich literature in the subject of spatial data quality, the review focused on the principles that govern the evaluation of spatial data quality following mostly the well-accepted ISO guidelines. On the other hand, the existing knowledge and research on the quality of UGSC was presented. This served two aims. Firstly, it was made possible to realise the strong interest of the research community on the subject and become accustomed with the current findings. Secondly, allowed to realise the gaps in knowledge and outline the areas where further research is needed.

Chapter 3

Methodology

3. Methodology

3.1 Research objectives

In Section 1.4 a list of research issues was presented. These issues emerged during the initial phase of the research as a first reaction towards the phenomenon of UGSC. The literature review that followed helped to recognise the more important of these issues and then to further clarify and group them into more specific research objectives.

3.1.1 Objective One: Understand the nature of the UGSC phenomenon

The first objective is to understand the nature of the UGSC phenomenon. The aim is not to describe the phenomenon in a conceptual or theoretical level, but to empirically examine its nature and discover its distinctive characteristics forming a knowledge base around the issue of UGSC. This knowledge will be fundamentally useful both for mapping agencies when they will need to form a strategy for their possible engagement with UGSC and the developers of Web 2.0 geo-applications (including again mapping agencies) to work towards further enhancement and improvement of the user participation and the content generation.

3.1.2 Objective Two: Evaluate the quality of UGSC

The second objective is to evaluate the quality of the spatial content generated by the users. For a mapping agency the spatial data quality is paramount. Consequently, a possible engagement with a crowd-based data source can possibly damage both the quality of the products offered to its users and the reputation of the agency. Thus, the empirical examination of the current quality levels of UGSC is necessary.

3.1.3 Objective Three: Highlight the challenges of UGSC and provide solutions

The third objective is a multiple one. Both from the literature review and from the experience gained during the empirical study of the first two objectives a number of challenges related with the nature of UGSC have been recognised regarding how to:

1. *Recognise and improve erroneous processes during content generation.*
2. *Achieve efficient content-level interactivity for a Web 2.0 geo-application.*
3. *Describe the fundamental characteristics of a Web 2.0 geo-application that could serve as UGSC source.*
4. *Efficiently share UGSC quality information*

The aim then is to analyse these challenges and provide the necessary solutions.

3.2 Methodology overview

To meet these objectives, two different types of sources of UGSC were used as case studies: the social photo-sharing Web applications of Flickr, Panoramio, Geograph and Picasa Web on the one hand and the vector-based datasets of OSM on the other. By doing so, this research covers a considerable part of the neo-geographic data spectrum available today on the Web.

It is easily understandable that the analysis of those two different types of sources dictated the adoption of different methodologies. The methodologies had to be able adjust to the degree of evolution, the nature and the particularities of each source.

Regarding photo-sharing Web applications, the analysis started when the relative literature on the subject was mainly at a conceptual and theoretical level with little or no support from empirical studies. Therefore, the aim was to develop a methodology to empirically analyse the phenomenon. The methodology had to be able to provide the means to examine the phenomenon from as many different perspectives as possible

(namely the spatial distribution of the data, the intensity of the phenomenon, its connection with the population distribution and analysis on user behaviour) in national and local level. Furthermore, the methodology adopted had to be able to falsify or corroborate a typology of the selected sources, revealed during the early steps of the analysis, to provide a better understanding of the phenomenon and enable the building of solid knowledge around the basic principles that characterise it. The outcome of this methodology can serve as the starting point for further work to access the potentials and the challenges that lay ahead and also to reach safe conclusion that can motivate further research.

Regarding vector-based application of UGSC, the overall methodology adopted was oriented towards the examination of the quality elements of OSM data. This is mainly because the analysis started at a point where the evolution, the impact and the potentials of OSM had already been subjected to the lights of academic and corporate publicity. This fact generated the need to perform the necessary analysis either by focusing on well established characteristics of OSM as a UGSC source that had not been examined yet or by trying to complement or develop on top of the findings of ongoing and existing research. For the former part the focus was on the analysis of the positional accuracy of the OSM data for England as there was no solid research on the specific subject. On the latter part, though, the quality element of completeness had been investigated (see Haklay 2010) mainly in terms of spatial coverage. Pushing the research a step further, the effort here was to examine the quality of OSM mainly by investigating the completeness and the validity of the attributes assigned to the spatial entities, through the attribution process adopted from OSM users, without leaving the process itself out of scope. Finally, in an effort to produce insights into the nature of the OSM data generation process, a comparison and a joint analysis of the results provided by the positional accuracy and the spatial completeness research were undertaken.

These two methodologies are further analysed in the Sections 3.3 and 3.4 respectively.

3.3 Geo-tagged photos

Before getting into more details on the methodology used for investigating the fundamental research objectives posed; let us discuss the nature of photo-sharing Web applications. This will enable to better understand the context of the research.

As stated earlier, the sources chosen for the analysis are Flickr, Panoramio, Picasa Web and Geograph. The reason for choosing these Web applications as potential sources of spatial content is twofold. Firstly, these photo-sharing applications are among the most popular websites and secondly, APIs are provided for them which allow access to their data. Apart from the evident commonalities among the chosen sources there are also some differences that make each source unique compared with the other sources. Flickr is a social networking website that allows users to publish and share photos of any content. Users can also comment, add tags (i.e. words that describe the content of the photo) and descriptions to photos. In contrast with Flickr where this functionality is only available online[4], Picasa Web is a similar Web application which is also supported through a desktop application (i.e. Picasa) that allows users to perform those actions for their photos off-line and upload the content when they wish to do so. Panoramio is a Web application that urges users to freely upload photos to describe places that they like or want to annotate. Finally, Geograph motivates its users to submit photos for every square kilometer (km^2) of UK and Ireland and thus declares a more precise aim to attract user participation and geographic coverage. Photo-sharing Web applications emerged around 2004 and their social impact along with the phenomenon of user generated content and the increased presence of geographic information in such applications have motivated researchers to consider them as a source of geographic information. However, from the selected sources, the former two are more socially-oriented whereas the latter two are more spatially-related. ('Ludicorp', the initial version of Flickr, was launced in February 2004, while Flickr and Geograph were launced in March 2005. Panoramio was launced in October 2005 and Picasa in June 2006).

The initial hypothesis is that the photo-sharing Web sources could be categorised into spatially implicit and spatially explicit ones, according to their overall attitude towards

[4] This was true at the time of the research as now Flickr also offers a desktop application.

space. Spatially explicit applications like Geograph and Panoramio, urge their contributors to interact directly with spatial features. In other words, spatially explicit applications ask from their contributors to focus their attention into capturing spatial entities in their photos. In that sense, photos of people, photos inside buildings or object's close-ups (Figure 9) are not preferable for such sources. Nevertheless, there is no established way to reject such photos or stop their contribution. At the same time, these applications encourage their users to contribute photos (and thus captured content) which are spatially distributed. For example, Geograph is based on the attempt to capture at least a photo for each km^2 in the UK. In contrast, Flickr and Picasa Web are more socially oriented, as they are aiming to allow people to share their photo albums with no explicit reference to space, and thus are regarded as spatially implicit Web applications. The support of geo-tagged photos is one of the many interesting features that spatially implicit applications have but spatial information is neither one of the core features nor is it the main motivation of their users, in contrast with what takes place in spatially explicit applications, in which the users are explicitly expected to use geography and location as a motivational and organisational factor.

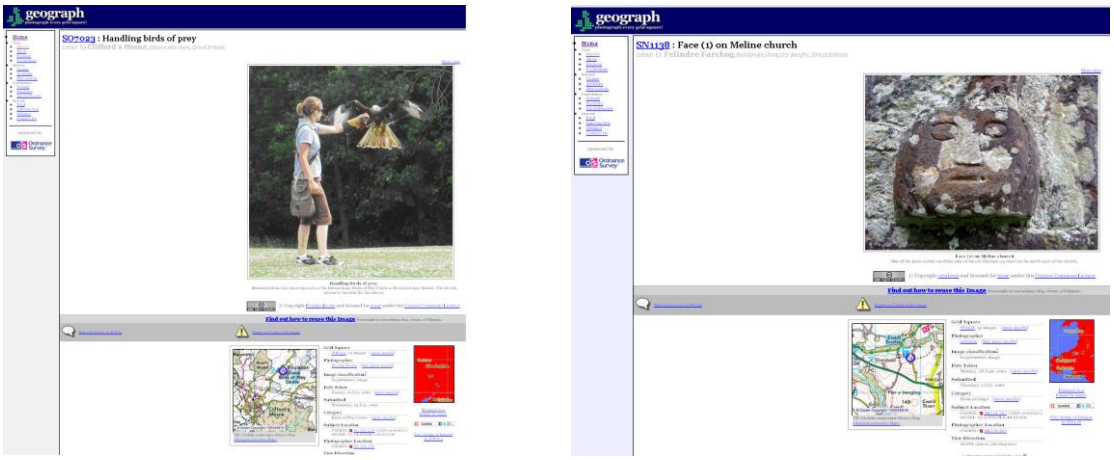


Figure 9. Geograph's photos with no spatial interest.

In all the selected Web applications users can upload photos, add titles, tags, descriptions and comments to photos, form groups and socialise with other users (Figure 10). In addition, and more importantly, geographic information can be added to the photos uploaded through a process commonly known as geo-tagging (i.e. a photo is associated with a pair of co-ordinates). The geo-tagging can be achieved through various ways that have direct impact on the accuracy of the location recorded. In

general, there are three main ways used for associating geo-location to a photo. The first one is purely manual and requires from the person that wants to geo-tag a photo to pinpoint a place on a map. Instantly, the photograph is associated with the co-ordinates of that specific location. This process though creates an important ambiguity. It is unclear whether the place pinpointed corresponds to the capture location (i.e. where the person was standing when the photo was taken) or to the photo's theme location (i.e. the actually location of the object depicted in the photo). Figure 11 shows an example of that ambiguity where a photo search about the London's Battersea factory in Panoramio Website reveals that there are two groups of photos: those positioned at the actual place of the landmark (red square) and those positioned at the capture locations.

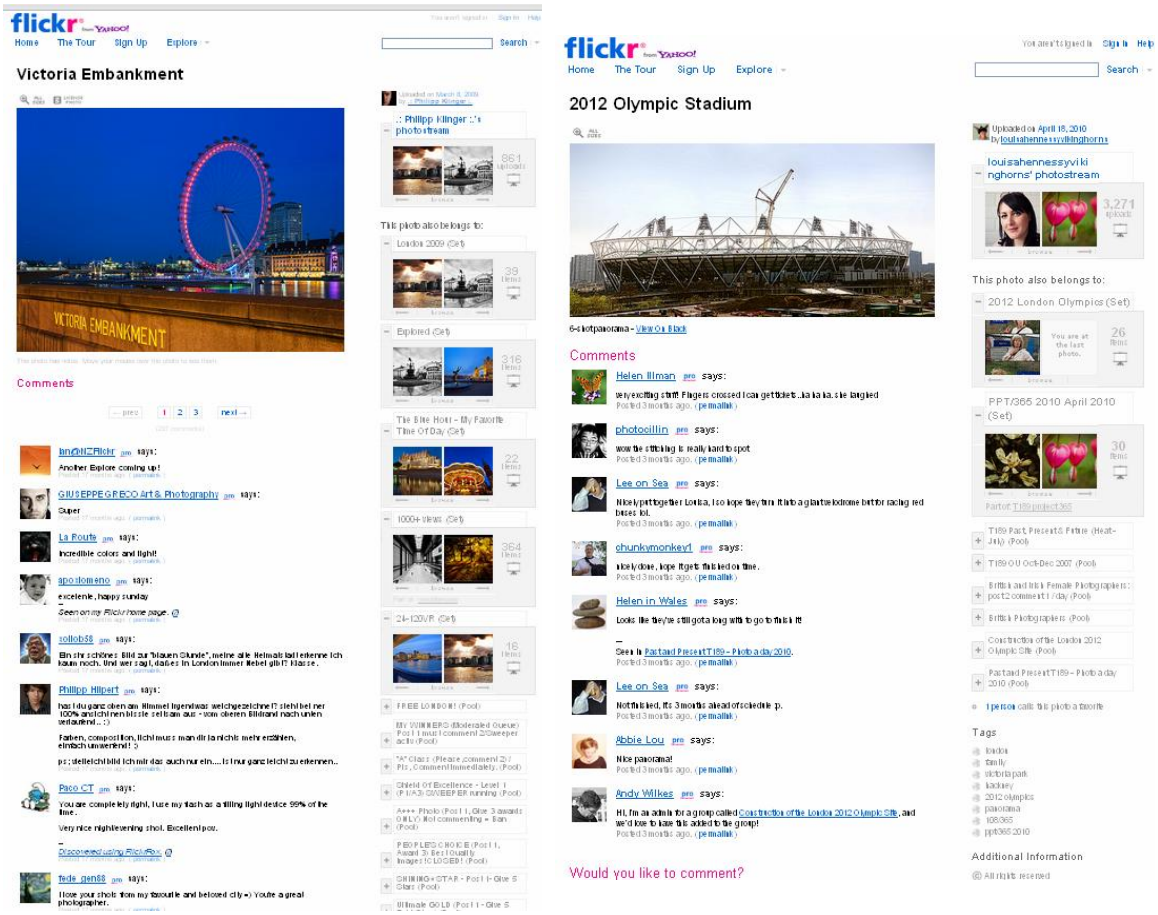


Figure 10. Flickr users commenting on published photos.

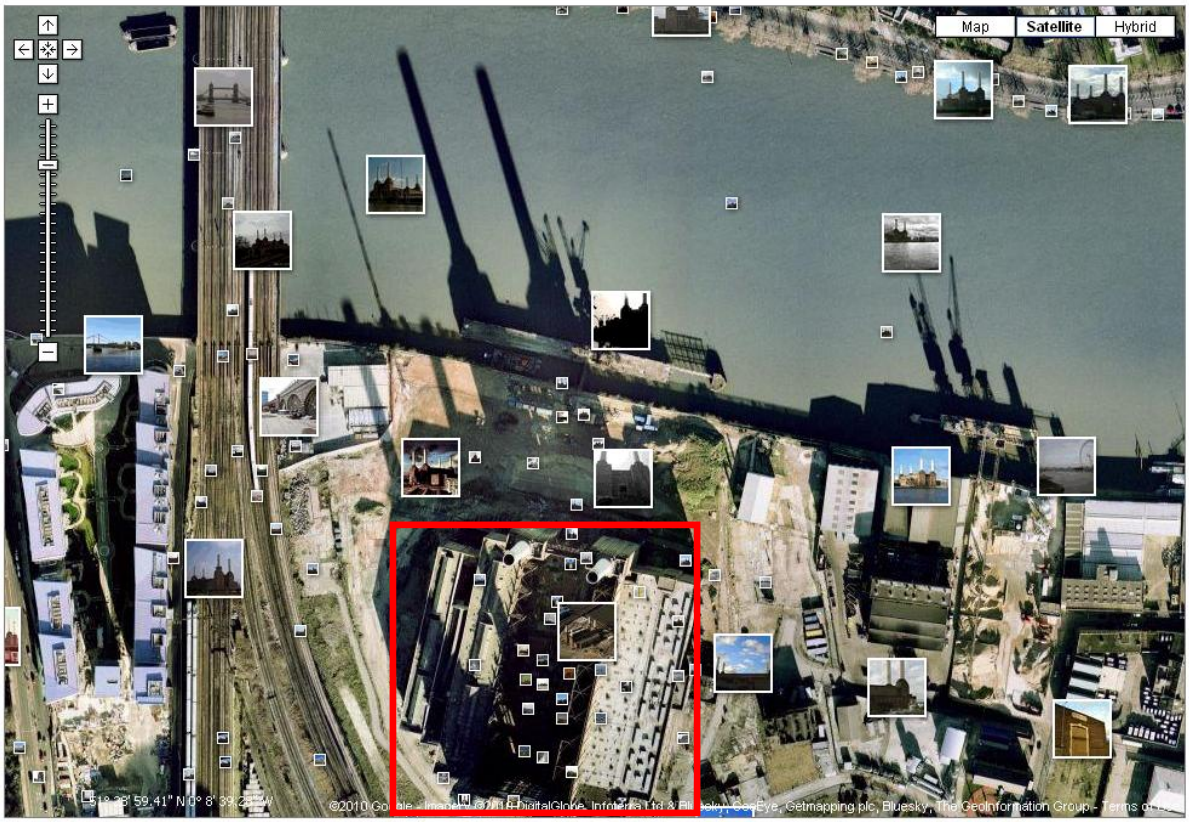
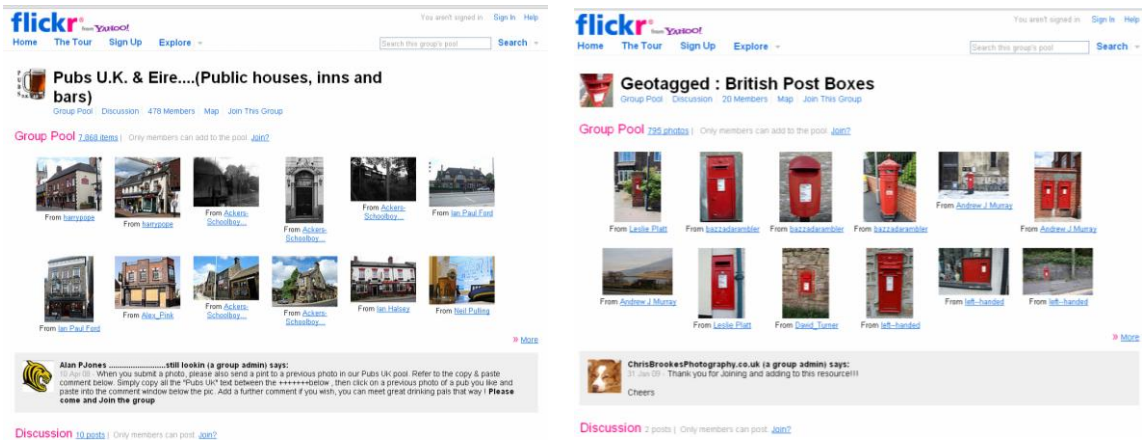


Figure 11. The spatial distribution of Battersea photos in Panoramio Website.

The second method of geo-tagging is a semi-automatic one. The user might be equipped with a camera and a hand-held GPS receiver. By doing so, the user can record the photo capture location using the GPS and at a later time to manually associate the photograph taken with the stored GPS position. As the GPS-enabled devices become ubiquitous, it stands to reason to suggest that this method will gradually be eclipsed. Nevertheless, for quite some time this method was a popular way to geo-tag images and therefore a significant part of the Web applications' photo pool have been geo-tagged using this method. Finally, the third method is a completely automatic one that takes place when the photo capturing device (either a camera or a mobile phone) is either GPS-enabled or is able to calculate its position by triangulating the signal from the mobile network cells (for the case of mobile phones). In those cases the location of the device at the time of photo taking is automatically recorded in the header of the image that stores the photo's metadata, known as Exchangeable Image File Format (EXIF). Consequently, the photo is automatically geo-tagged and when the user uploads such a photo to the photo-sharing Web application, the later is able to automatically locate the photo on a map.

The geo-tagging process is particularly important since all the information that a photograph bears is automatically correlated with a specific location. It is expected that a portion of such photographs will include information that has no particular spatial interest (such as close-ups of persons or objects). On the other hand though, a great variety of the photo's elements can be of some value for GI retrieval purposes. This information might relate both to what is actually captured by the photo and to any attributes contributed by the users. For example, a photograph might show the construction of a new building, or give some kind of indication that a new road is or will be opened at a specific place. Additionally, the photograph's caption, its tags or a short description attached by a user, might include spatial information such as place-names or administrative boundaries that can be further exploited. Interestingly, as such Web applications provide a fruitful environment for user interconnection (a process loosely defined as social networking), user groups can be formed that are spatially oriented. For example (Figure 12), the groups might focus on taking photos and commenting construction sites or a variety of points of interest such as pubs, coffee shops, post offices etc.



(a) **(b)**

Figure 12. Flickr Groups about a) Pubs and b) Post Boxes.

While the photo-sharing websites received some attention for GI retrieval purposes (Liu et al. 2008, Purves and Edwards 2008, Dykes et al 2008, Popescu et al. 2008), there has been no systematic analysis of their potential role as reliable and universal sources of spatial content. Therefore, the current research aims to shed light on the breadth and

depth of the spatial data provided by such sources. This effort was conducted from the point of view of a mapping agency and therefore, the focus was on the ability of such sources to provide the necessary data to assist mapping agencies in updating existing map products, creating new ones or facilitating the established map production procedures.

3.3.1 Methodology for photo-sharing websites

The main steps of the methodology adopted for the examination of the photo-sharing Websites were:

1. The gathering and the creation of the auxiliary datasets
2. The development of a strategy for the collection of the actual data through the sources' APIs.
3. The development of a Web application that would be able to implement the data collection, and
4. The collection and analysis of the data.

3.3.1.1 Auxiliary datasets

To implement the data collection process two vector datasets provided by Ordnance Survey (OS), through the EDINA service, were used. A third dataset, Great Britain's population surface was constructed for the needs of the analysis:

a. The Great Britain boundary

The Great Britain boundary dataset includes the coastal boundaries of Great Britain and it has been derived from Ordnance Survey's Strategi data (scale 1:250000). The total area covered is 230,535km^2.

b. The Great Britain National Grid

The Great Britain National Grid has been created by progressively dividing the space into square tiles. As described by OS (2010a), the largest unit of the Great Britain National Grid is the 500km square and it is named by a letter from A-Z (not including I)

(Figure 13). Each one of these tiles is further divided into 100km tiles named by two letters; the first letter is inherited by the 500km tile in which each 100km^2 tile belongs and the second is a letter form A-Z, once again excluding letter I. Each 100km^2 tile is further divided into 10km^2 tiles where each one of these tiles is named according to its position in a Cartesian system numbered from 0-9 in each of the X and Y axis. Finally, each 10km^2 tile is divided into 1km squares which are the smallest units of the Great Britain National Grid. Again, a Cartesian system is used to number each of these tiles.

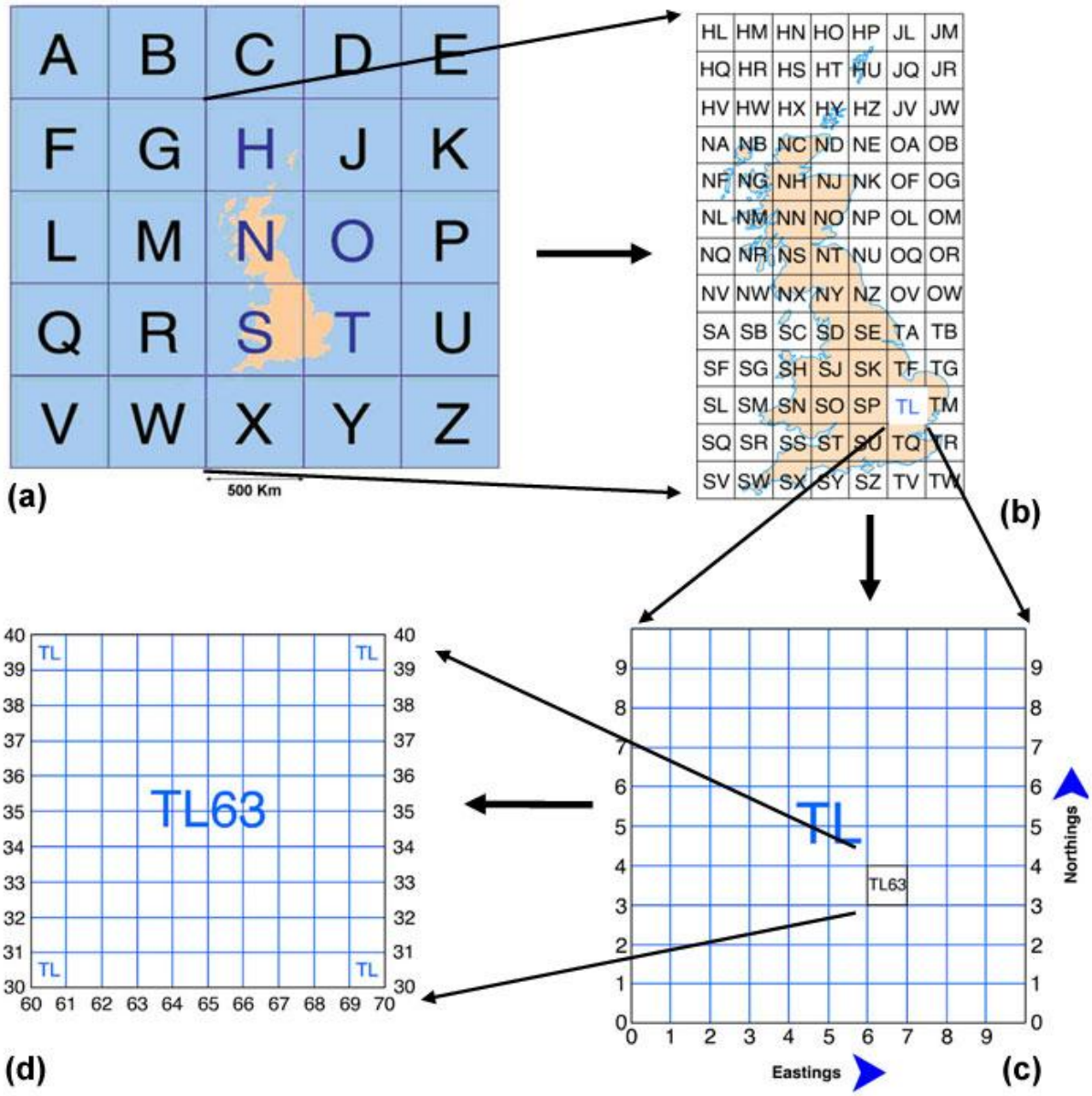


Figure 13. The mechanism of Great Britain National Grid creation: a) the 500km Grid, b) the 100km Grid, c) the 10km Grid and d) the 1km Grid.
source: Ordnance Survey

The 1km^2 tiles of the National Grid dataset have also been used by Geograph in their data collection method. This fact allows performing a like for like comparison between the data collected from the other three sources and the Geograph's data.

Since the 1km^2 National Grid has tiles that reside outside the Great Britain boundary, a spatial selection was necessary to define the tiles that would be used in the data collection process. The selection yielded 238,920 tiles. For each one of the tiles that belong to this subset, it was examined how many photos have been submitted to each selected Web application (see Section 3.3.1.3). Thus the results of the analysis have a spatial resolution of 1km^2.

c. Population dataset.

As part of the analysis a population dataset for Great Britain was needed. The dataset was constructed using the Output Area (OA) provided from EDINA and the population datasets for England, Wales and Scotland as provided by Casweb based on the 2001 census. Initially the two datasets were joined (i.e. the population data were assigned to the OA) and then a software (Surface Builder) developed by Prof. Dave Martin (Martin 2007) was used to generate a population surface with a spatial resolution of 1km^2 and the same origin as the OS National Grid. The surface was transformed into a point ESRI shapefile that was used to join the 1km^2 National Grid with the population data. This resulted into creating a 1km^2 National Grid enriched with population data for each tile.

3.3.1.2 Use of APIs

Generally, the use of the APIs in the case of the photo-sharing Web applications allows the access of the actual photographs, their location as well as the descriptive details that are associated with the photograph. More specifically, an API is an interface that allows a user or a programmer to manipulate the responses of a Web service. The exploitation of the API takes place through the use of a Uniform Resource Locator (URL). A general form of such a URL is formed from a series of parameter-value pairs as shown below:

```
http://somewebservice/api_key=key_value&param_1=param_1_value&param_2=
param_2_value&....&param_i=param_i_value
```

where:

`http://somewebservice/` : is the URL of the Web service's API

`api_key=key_value` : is a unique code that allows the access to the API service (whenever such key exists)

`param_i=param_i_value` : is the parameter values posted to the Web service that serve as a series of selection criteria (e.g. bbox = LowerLerftX, LowerLeftY, UperRightX, UperRightY)

The response to a properly formatted API request is a text document that follows a known structure (e.g. XML, KML, JSON, GeoJSON, ATOM etc.). The response contains all the data that fulfil the criteria posed at the URL's parameters section. The known format and structure of the response's document allows the creation of algorithms that can parse it and select the bits of data that are necessary in each case.

3.3.1.3 Data collection Web application

One of the basic components of the methodology was the data collection mechanism. The mechanism had to provide the means to collect the data needed at each step of the analysis. For example, at the first step of the analysis it was required to collect the number of geo-tagged photos submitted to each tile for all four sources. In practical terms this meant that there was a need to perform almost 1 million API requests and process their responses. It is easily understood that such a task could only be accomplished with the help of an automated processes and for that reason a Web-based application was developed. The architecture of the Web application is shown in Figure 14.

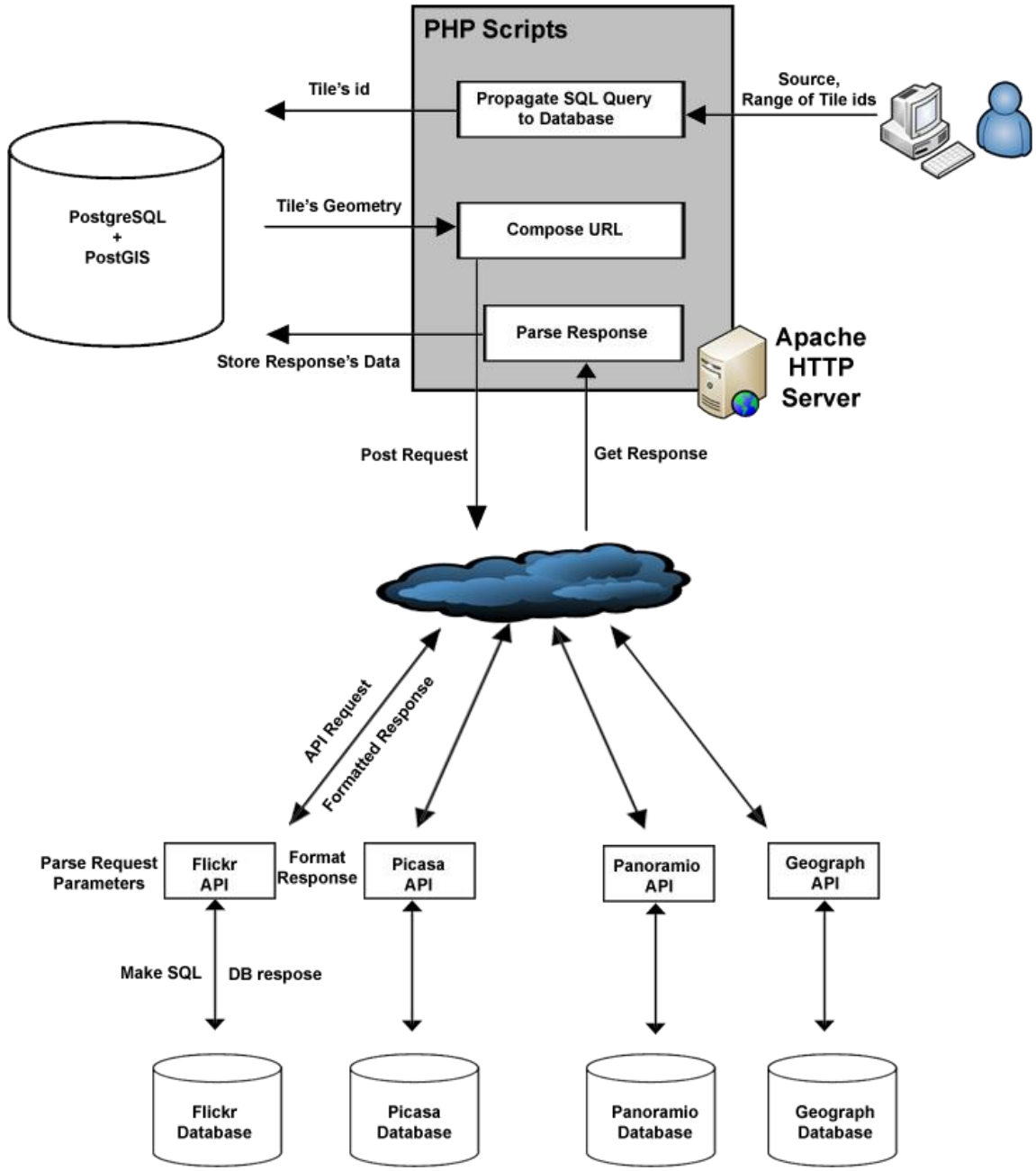


Figure 14. The data collection mechanism.

The basic components of this Web application are the spatial database, the application scripts and the HTTP Server. The software chosen to serve the needs of the spatial database was the open source PostgreSQL enabled with the spatial extension of Postgis. The necessary application scripts were written using the PHP server-side and the JavaScript client-side programming languages. The Apache Server was chosen as the HTTP Server.

The first step was to insert the subset of the 238,920 tiles into the Spatial Database as the tiles' geometry was necessary to properly form the parameters' part of each API request. Once that task was completed, it was possible to use the application's interface to start collecting the necessary data by just submitting the necessary parameters (such as a range of tiles of interest). These parameters with the help of the PHP script were transformed into valid SQL queries and then were submitted to the spatial database. The response of each SQL query, that contained the geometry and the id of each tile, was received and parsed again by the PHP script. The returning geometry value was one of the information pieces used to compose the proper, for each source, API request (i.e. a correctly formed URL). The API request could be complemented with any necessary parameters depending on the needs of the analysis (e.g. introduce date restrictions). Then, the API request was posted to the Web Service hosted by each source. As discussed, the response of each source is encoded using known, yet different formats. The next step in the process was to receive and parse at the server side the response. The parsing step allowed getting grip of all the necessary data contained in each response. Finally, the pieces of information extracted from the Web Services' responses were stored in the correct record (i.e. the correct tile) of the database.

3.3.1.4 Data collection and analysis

The gathering and the creation of the auxiliary datasets on the one hand and the development of a suitable Web application on the other, formed the basic infrastructure that enabled data collection according to the needs revealed during the different steps of the analysis.

Initially, the aim was to map the phenomenon of UGSC for the selected photo-sharing Web sources and thus to provide a better understanding of the general pattern of UGSC. As this was a relatively new, uncharted and evolving phenomenon, the first step was to recognise and document its basic principles and realise the relationship of the phenomenon with space. Therefore, there was a need to collect data regarding the geo-tagged photos submitted to each one of the 238,920 square tiles of the 4 selected Web sources. When this process was completed the dataset was a $1km^2$ grid of Great Britain for every tile of which the number of geo-tagged photos submitted was known.

The exploration of this dataset started with the examination of the basic descriptive statistics of each source (i.e. minimum and maximum values, sums, means and standard deviations). Although this is a fairly simple process, it is a necessary to see the basic characteristics of each Web application. Furthermore, this step shed light on the volume of the phenomenon while at the same time provided the first insight of the dissimilarities among the examined Web applications. Nevertheless, while the sheer number of geo-tagged photos available can provide a basic understanding about the magnitude of the phenomenon, it does not provide adequate answers about its spatial dimension.

Through the methodology followed, the effort was on investigating the relationship of space and geography in the development and the evolution of such crowdsourced efforts. Although the null hypothesis that all places are equal in terms of geo-tagged photo coverage was never expected to be corroborated, the interesting point in this analysis was to realise the degree of differentiation. The examination of the effect that space and geography have on the creation of those differences, and consequently the effect that these factors have on user participation and contribution, will allow forming valid conclusions regarding the role that such Web sources can play in a GI retrieval effort.

Interestingly, these differences can be examined from two different points of view. On the one hand, it is important to realise the differences in user participation and thus in data submission for each specific source. Examine, for example, which are the places with high, moderate and low user contribution for each source and then try to interpret why this is happening. On the other hand, it is equally interesting to realise the differences, if any, among the sources. Are all sources behaving in the same way? If not, what is the reason behind that? What is the role of space in both of these differentiations? The methodology was able to provide adequate answers to questions of that type.

The first step to investigate these aspects was made through the visualisation of the data collected by exploiting the spatial reference of the National Grid. This allowed the creation of thematic and 3-dimensional maps for each source. Visualisation is a

powerful method for data exploration and hypotheses testing especially for large and unknown datasets (MacEarchren and Kraak 2001, Thomas and Cook 2005, Dykes et al 2008). This process helped to examine the spatial dimension and added to the effort of understanding the nature of the phenomenon. More specifically, data visualisation revealed eloquently the spatial distribution of the phenomenon; it became evident where the clusters of geo-tagged photos or places that received no coverage from each source are located. This element of completeness is a point of high importance when it comes to using such sources for GI retrieval, especially for national mapping agencies. Another interesting point revealed was the volume difference in the data submitted inside the formed clusters. Moreover, by comparing the spatial distributions of the 4 sources allowed to further realise the differences of the Web applications under examination

In a second layer of analysis, the focus turned to the effect of population. As the analysis deals with a user-centric phenomenon, it is important to realise whether, and to what extent, the users' location is affecting the data collection and the data submission processes to these Web sources. In that context, the methodology followed connected at the same statistical analysis the geo-tagged photos and the auxiliary population data of Great Britain. As Dykes and Wood (2008) explain, building a surface that both relates to the intensity of a phenomenon and to the location of people, allows the statistical analysis to be carried out. This can be accomplished by calculating expectation surfaces using the chi-statistic:

$$chi = \frac{(ObservedValue - ExpectedVa\,lue)}{\sqrt{ExpectedVa\,lue}}$$

In our case, the observed values is the number of photos submitted for each tile and the expected values is the value from the population density surface for the corresponding tile. This comparison allowed to understand the correlation of UGSC phenomenon with the population density. The chi index will be negative for the tiles where the observed value (i.e. the number of geo-tagged photos) is lower than expected (according to population data) and positive when it is greater than expected.

In the next stage the effort was to further analyse the differences between spatially implicit and explicit sources. The focus in this part of the analysis was on the areas where clusters of user generated data were located and specifically to the areas where there was a relatively constant submission of geo-tagged photos. For those sites the analysis examined the data flow (i.e. the number of photos submitted) for the most popular tiles of Flickr and Geograph. A threshold of 15 photos was set to characterise a tile as 'popular', as this represents an average of one submitted photo per tile per quarter over a four year period, since the Web applications were launched. Additional to the data flow, the currency of the data available for the popular tiles was examined since when it comes to using spatial information (e.g. in updating mapping products) currency of data is paramount.

In the next layer of analysis, in an effort to examine the phenomenon in more detail, there was a need to collect data below the coarse level of $1km^2$. For that reason 15 popular areas in Great Britain were randomly chosen and new datasets of the geo-tagged photos submitted from January 2005 until April 2009 were collected for the 1/25 ($141km^2$) of the common popular tiles of both Flickr and Geograph. This time the datasets collected were more detailed as they included the actual photographs submitted (50,504 to Flickr and 11,937 to Geograph) and details associated with each photo such as title, tags, comments, date of capture and submission, user name and the recorded location. The new detailed datasets collected for the chosen areas and especially the co-ordinates of where each photo was positioned, allowed a kernel density analysis for these areas which helped to examine the spatial distribution of the phenomenon at a large scale and enabled the comparison between users' behaviour in explicit and implicit sources. In an effort to quantify this observation in more practical terms, an event analysis for each study area was carried out. This allowed the quantification of the repetition observed in the photo locations for each source.

As Kuhn (2007) describes, UGSC can be studied both as a social phenomenon that raises scientific questions (i.e. Why people participate in such initiatives? What is their motivation? Will they be committed to this effort?), and as a social phenomenon that is used in science. Although clearly this research is focused on the latter it cannot leave untouched the basic issues of the former. Along the same vein, Coleman et al. (2009)

support the contention that the analysis of the human element, which is the driving force of this phenomenon, will enable all interested parties to understand the process of content generation. In that context, a preliminary investigation of the phenomenon was necessary through the empirical examination of the users' behaviour. Therefore, the next step of the analysis focused on user behaviour with regard to user data submission, as measured by the time difference between capturing and uploading a photo. This will reveal the element of currency in the nature of the submitted geo-tagged photos and provide an insight into how users' participation is evolving through time. Furthermore, the time period that the users remain active in an area (i.e. time difference between first and last photo submission for each user) was also examined. This will be helpful in differentiating the tourists or visitors of an area from the locals. It is well accepted in the literature (see for example: Haklay and Tobon 2003; Dunn 2007; Budhathoki et. al 2008; Elwood 2008c) that local knowledge is a very important factor in retrieving geographic information and in documenting and mapping both spatial and non-spatial elements of a place (see also Section 2.2.2.5).

Finally, as the differentiation between spatial implicit and spatial explicit sources was becoming more and more evident the questions raised concerned the ability of implicit sources to represent a sufficient database to aid the update of spatial datasets at a national level. To deepen our understanding on the subject a new dataset regarding the number of geo-tagged photos submitted in Great Britain was collected for Flickr, six months after the first dataset. This step of the analysis was undertaken to monitor the evolution and examine the productivity of the phenomenon over a period of time. At the same time, by comparing the two Flickr datasets, it was possible to examine the type and extent of changes in the content provided by this type of Web sources as well as the effect that these changes can have in terms of GI retrieval. Additionally, this time, beyond the sheer number of geo-tagged photos submitted to each tile, a new element was also recorded: the user id of each geo-tagged photo. This resulted into creating a dataset containing all Flickr users that had submitted a geo-tagged photo in Great Britain, the number of photos that each user had contributed and in which tiles that contribution was distributed. This dataset gave the opportunity to examine whether the Pareto Principle, that dictates that the 80% of the effects is generated by the 20% of the causes, is corroborated in the field of geo-tagged images. In our case the effect is the

submission of a geo-tagged image at the social networking Web application of Flickr and the causes are the users that submitted geo-tagged photos.

3.4 OpenStreetMap[5]

OpenStreetMap (OSM) is a Web 2.0 initiative that allows users to create and freely use maps. Pursuing this aim, OSM has been developed and functions on the basis of a crowdsourced mechanism. Notwithstanding the crowdsourced nature of OSM it must be noted that a great portion of the data currently available in OSM has been donated by institutional organisations. There is a long list of sources that the OSM project has used to gather spatial data. In any case, OSM manifests its principle to strictly collect data from out of copyright sources (OSM 2010a).

On the other hand, OSM provides the means to contributors from all over the world to participate in map production by submitting spatial data using the OSM infrastructure. This infrastructure includes OSM editors that use the satellite images available at Yahoo! Maps as a backdrop, and thus enable users to digitise any spatial entity they can see. Through this process users are not allowed to record street names, but the level of compliance with this rule cannot be monitored. Alternatively the users can upload data (both geometry and attributes) that have been captured using GPS devices. Furthermore, the OSM community is regularly organising gatherings of the OSM contributors, known as Mapping Parties, in a more co-ordinated effort to map certain areas. Since the Mapping Parties take place physically at the mapping area, participants are able to attribute data in addition to contributing the spatial entities' geometry. The data from both sources (digitisation and GPS) is regularly gathered and after the necessary steps of map composition and styling the outcome is rasterised into map tiles.

[5] Parts of this section have been adapted from:

Antoniou, V., Haklay, M., Morley, J., 2010a. A step towards the improvement of spatial data quality of Web 2.0 geo- applications: the case of OpenStreetMap. *Proceedings of the GIS Research UK 18th Annual Conference*. London: UCL, pp.197-202.

While the geometry capture is an easily understandable process, the data attribution is somewhat different. More specifically, there is a widespread manifestation in the OSM wiki pages that OSM community does not want to impose any rules on its participants regarding the attribution of spatial entities. On the contrary, through the wiki pages it is claimed that participants can freely use any lawful method and practice to create spatial content and are free to assign any kind and type of attributes (using tags) to real world features (OSM 2009):

> *"OpenStreetMap does not have any content restrictions on tags that can be assigned to Nodes, Ways or Areas. You can use any tags you like"*

In practice though, OSM users have created numerous wiki pages that are full of instructions regarding procedures to describe geographical objects (OSM 2009):

> *"However, there is benefit in agreeing on a recommended set of features and corresponding tags in order to create, interpret and display a common basemap"*

These instructions are not presented as hard and fast rules but rather as lessons from other contributors' experiences or as best practice proposals. Nonetheless, this wiki-made user guide has evolved into a quite complicated and some times hard to follow technical document. It is interesting to note that the road map to create or change such a rule is totally democratic. In brief, users can start a proposal procedure whenever they feel that a mapping feature should be added or changed. This procedure includes a discussion and a voting step which determines whether the proposal will be rejected or accepted and consequently implemented. The active and approved map features are documented with proper instructions and both written and visual examples. This is a continuous process; entities from the map features list can be replaced with new ones and the old entities become deprecated.

An interesting point in the whole process of data capturing and attribution is that the users are not only able to create new data but they can as well modify or update existing data that have been created by other users. This theoretically and practically, perpetually iterating collaborative process enables thousands of users to create small patches of a

world map. Yet when these patches are put together the final outcome is quite impressive. OSM repository nowadays contains millions of spatial entities that belong to a variety of thematic layers such as road network segments, points of interest, administrative boundaries or land use polygons that form a world map. The level of map completeness though, differentiates greatly from place to place around the globe. There are parts of the world that have very little or no coverage, and there are parts, such as London for example, where the map created by the OSM users is highly detailed. As a result, a wide range of applications and new mapping products (e.g. opencyclemap) has started to spring from the core OSM spatial datasets.

As stated earlier, this collaborative work is provided freely to the OSM users (irrespective of whether they have contributed to the data collection or not) through different ways of data dissemination. The easiest way for users is to navigate to an area at a certain zoom level and download what they see either as a raster map, a pdf file or as vector-encoded data. This method though, covers only the need to get hold of a map for viewing purposes. A more sophisticated way of data downloading, mainly designed for developers, is to use the OSM API, which at the time of writing was at the 0.6 version. As was the case with the photo-sharing Web sources, it is possible to pass through the OSM API a range of parameters that serve as query criteria. The OSM API response is an XML-encoded document that contains the OSM row data that fulfil the query criteria. Once again, the need to parse the XML document arises in order to get hold of the necessary bits of data. Finally, private companies (such as Cloudmade and Geofabrik) provide the ability for anyone to download OSM data as ESRI shapefiles.

The success of OSM has drawn the attention of scholars and researchers that started to examine issues like the potentials, the credibility, the sustainability, the quality and the fitness for purpose of such data at a theoretical level (see for example Goodchild 2007a, Sui 2008, Flanagin and Metzer 2008). Little was the empirical research conducted (see for example Haklay 2010) that provided insights in the quantification of several characteristics of the phenomenon. In an effort to extend that quantification effort the research was oriented towards an empirical examination of different aspects of the quality of the OSM data. On the one hand, the quantification of the data quality will provide the common language for any interested party to understand the nature of the

OSM data. On the other, by examining the OSM data quality, it will be made clear if and at what extent such data can be used in the mapping procedures of an institutional organisation such as a mapping agency. In contrast with the photo-sharing Web applications, OSM provides content that is considerably more familiar to any mapping agency (i.e. spatial entities' geometry and attributes) and therefore the aim of the research in addition to the examination of the UGSC potentials, is also to provide conclusive results on the quality elements of such data.

3.4.1 Methodology for OSM datasets

The initial experiment was conducted so to realise the evolution of the phenomenon over a period of time by using England as the geographic area of scope. This step was necessary to examine at first hand the nature of the phenomenon through the data contributed. It would also provide insights on the evolution of the phenomenon in terms of the volume, the type and the attribution of the data captured. Based on the results of this reconnaissance step, the focus was turned towards the evaluation of the dataset's positional accuracy. The direct examination of the OSM's positional accuracy gets to the heart of the data quality issue and sets the tone for the overall quality of the dataset. As this is a vector based pool of spatial data, geometric accuracy is paramount. The next step was to correlate the positional accuracy results with the outcome of the completeness evaluation provided by Haklay (2010) to examine any similar behaviour of these two quality elements. The final step was the quality evaluation of the data attributes. The collaboration-based method described earlier, coupled with the findings of the initial step regarding the evolution in the data capturing, provided an interesting area of research regarding the consistency and the attribution process and raised questions whether the data collected can stand against a series of quality tests.

As described, OSM is a constantly changing and thus dynamic dataset as users around the world can delete, modify or add data at any time. Nevertheless, a benchmark procedure (see Section 2.4.2.3) was used for the quality evaluation of both the positional accuracy and the data attribution.

3.4.1.1 Data examination

In the first step of the analysis the focus was on the evolution of the OSM data repository for England, and therefore there was a need to collect, over a period of time, OSM data sub-sets. As it has been explained, there are different ways for downloading OSM data. In the first step of the analysis the download service provided by Cloudmade was used. Data was downloaded at three different times: the first datasets was downloaded in January 2009, the second in April 2009 and the third in July 2009 (i.e. with a three months interval). The Highways and the POIs were the two broad thematic layers that were chosen for monitoring. It must be noted that in the OSM terminology the 'Highways' layer covers the entire road network (including pedestrian and cycling routes) and includes the geometry of the entity and the road type attribute. Similarly, the POIs include the geometry of the spatial entities, their type (e.g. banks, ATMs, places of worship, gas stations etc.) and their names.

As this was the first interaction with the OSM data, it was deemed necessary to realise the overall nature of the data and the evolution of the phenomenon. In this effort a like for like comparison was made between the first and second, and between the second and third highway datasets. Each comparison revealed the changes in the numbers of the spatial entities recorded for each road category as well as the percentage of entities that remained geometrically unchanged. At the same time the number of spatial entities with and without names, the basic attribute element of a road network entity was also monitored.

A similar approach of comparisons between the three datasets was followed for the POIs dataset. The comparison here revealed the changes in the numbers of entities of each point category. Additionally, the numbers of entities that were unchanged, deleted or moved over the 6 months period was examined. In order to perform that type of analysis the fact that all POIs had both a type and a name recorded, was exploited. More specifically, using simple overlap spatial queries it was possible to retrieve the information about the *unchanged* points. Then, for the rest of the dataset's points a proximity query was made to find a point of the same type and name that fell under a distance threshold. It is clear that these points refer to the same spatial entity but have

been *only geometrically changed*. Next, for the points that also fell under that distance threshold but had the same type but not the same name a manual examination was made to find out if each spatial entity had subjected to *both geometric and name change* over that period. After retrieving all these point categories, the remaining points either have been geometrically moved more than the specific threshold or they have been *deleted*.

The decision on the value of the distance threshold (i.e. the query's search radius) was made after manually examining different sample areas. It was realised that the proper naming of a point had the following format: "Type:Name" (e.g. Pub:Byron's Arms). So, for the points that followed that proper format, and their name remained unchanged between the two datasets under comparison, it was reasonable to extend the search radius to 800m (the preliminary examination showed that such distances occurred only to facilities that covered a large space such as super markets, and the POI had changed position between the entrance, the main building or the parking lot). On the other hand, for the points where their names did not follow that pattern the search radius was specified to 10m, approximately the approximate positional accuracy of a GPS receiver.

Finally, by further examining the types and the names of the POIs that remained geometrically unchanged between two periods it was possible to record the number of points that had a type or a name change or not have been changed at all (neither their geometry nor their attributes).

This initial step was needed to understand the evolution of a highly promising phenomenon. More fundamentally though, this analysis gave a first insight on the nature of a global-wide crowdsourced process regarding the creation of spatial data and the linkage between that process and the outcome. As both the process of creation and the content itself are newly introduced to the Geomatics community, it is not adequate to focus only in the final outcome while being indifferent regarding the mechanisms that generate the data. The stimulus for turning the attention towards the production mechanism was given from that first analysis' results. The understanding and the evaluation of the overall data generation process can reveal potential weak points, it will set the context at which the data quality results are valid, and it will enable to recognise any flows or systematic errors that the process itself introduces to the data produced.

Finally, once that level is reached the improvement of the process and therefore the improvement of the data output can be considerably facilitated.

3.4.1.2 OSM's road network positional accuracy

As discussed in Section 2.4.2.3, a direct external data quality evaluation method can be used for the evaluation of a vector dataset's positional accuracy. In order to implement that methodology here, an external reference dataset of greater accuracy must be available apart from the dataset to be tested. For this particular experiment the road network of England was chosen as the external reference dataset. The rational behind the choice to examine the positional accuracy of the road network is because the road network covers the main body of the OSM datasets, and because there is external reference data available that can be used for that examination. As external reference dataset was used the road network layer included in the OS Meridian 2 dataset.

a. OSM Data

Instead of using the OSM API that would mean to parse the XML formatted data fragment of the dataset for England and then insert it into a spatial database, it was much more convenient and time efficient to directly use the OSM data provided from Geofabrik without affecting the analysis' outcome. This dataset is produced by the original user contributed OSM data; it is provided as an ESRI shapefile format and has the same positional accuracy as the OSM data. The Geofabrik option was preferred over the Cloudmade download service because its shapefiles include also the original unique OSM ID of every spatial entity. Moreover, the Geofabrik datasets are updated more frequently and thus provided more current data for the needs of the analysis.

The dataset used for the evaluation was downloaded on the 14th of September 2009 and, as described earlier, contained the geometry of all types of roads including types such as bridleways, paths and footways. A preliminary visual examination of the data against the Meridian dataset showed a considerable number of misclassifications in the types of roads. For that reason, it was decided to include the entire Geofabrik dataset during the evaluation process so not to exclude valid road intersections, as the inclusion of layers such as paths and footways did not affect the algorithm's outcome.

b. OS Meridian 2

The positional accuracy of the OSM data was examined against the roads layer of the OS Meridian 2 product. The Meridian 2 dataset is derived from both large-scale and small-scale digital databases and contains a variety of layers such as road network, railways, administrative boundaries etc. Regarding the road network, the data is derived from the roads centrelines of the Ordnance Survey Roads Database and the scale of the data can be 1:1,250, 1:2,500 or 1:10,000 depending on the area (the accuracy increases in the urban areas and lessens in the rural ones). The road network consists of road line segments and nodes that represent the intersections and the ends of the segments. A crucial point here is that in order to construct the Meridian 2 the Ordnance Survey Roads Database has a 20 metre generalisation filter applied to the centrelines of the road. It is important to note here that, this generalisation process does not affect the positional accuracy of node points. The resolution of the data supplied is one metre (OS 2009).

c. Positional accuracy evaluation

Driven by the fact that the accuracy of the Meridian 2 nodes have not been geometrically affected by the generalisation process, in contrast with the road segments themselves, it was decided to use the geometric position of nodes to evaluate the positional accuracy of the OSM dataset. As supported by Goodchild and Hunter (1997), this is a valid methodology for examining the positional accuracy of a dataset against another. The key difficulty, thought, in the whole process is to accurately match the corresponding nodes of the two datasets. If this step is made successfully then it is a fairly straightforward process to measure their in between distance and consequently find the positional accuracy of the test layer's node based on the position of the node in the reference layer.

Before completing the analysis, it was necessary to prepare the datasets for this type of analysis. The basic step was to build a node topology for the Geofabrik data. The ESRI shapefile was converted into an ArcInfo Coverage and a node topology was built using the ArcInfo Workstation. In that way, the OSM data was transformed to a similar form with the one followed by Meridian 2 and thus the comparison between the two datasets

was considerably facilitated. Both datasets were loaded into a PostgreSQL, and Postgis enabled, spatial database. A final step before the implementation of the positional accuracy algorithm was the harmonisation of the road's name data of both datasets. This was necessary since street names would be used in the node matching. The harmonisation consisted of the removal or the replacement of trivial parts of a street name such as articles and abbreviations that would reduce the algorithm's efficiency.

The evaluation process was only practical through an automated way and therefore an algorithm was developed to perform this task. The algorithm detects the correct matching of the nodes between the two datasets and calculates their euclidean distance and stores it in a spatial database.

More specifically, the algorithm, which is repeated for each Meridian 2 node, starts by searching for the Geofabrik node that is closest to a given node from the Meridian 2 dataset. This is done by searching for the minimum distance between the Meridian 2 node and the Geofabrik nodes that fall under a threshold of 40m. Then the algorithm verifies that the selected Geofabrik node is not closer to another Meridian 2 node. If there is no other Meridian 2 node closer to the selected Geofabrik node then this pair of nodes is recorded as a possible positive match. If the distance between the two nodes is equal or less than 15m (which is approximately the positional accuracy of a hand held GPS device and also taking into consideration the uncertainty of the Meridian 2 accuracy according to the scale of origin) then the node matching was directly accepted as a true matching. Otherwise (i.e. the distance between the two nodes was from 15 up to 40m), a second level of verification was introduced. Here the verification takes place through the examination of the road names that intersect each of the pair nodes. The algorithm finds out and records the number of road segments that start (or end) from the specific node of each dataset and how many of those segments have a name. Then the harmonised names are compared and the number of matched names is recorded. The whole process creates a dataset that in every record holds the following data (iv up to viii only when the distance is more than 15m):

i. The Meridian 2 node ID.
ii. The matching Geofabrik node ID.
iii. The distance between the two matching nodes.

iv. The number of road segments that start (or end) from the Meridian 2 node.
v. The number of those Meridian 2 road segments that have a name.
vi. The number of road segments that start (or end) from the Geofabrik node.
vii. The number of those Geofabrik road segments that have a name.
viii. The number of matching road segment names.

Initial tests were conducted to examine the validity of the results yielded by the algorithm particularly examining the 15m threshold for the direct acceptance of the nodes matching and the upper threshold of 40m. After compiling this dataset, a manual exclusion of possible miss-matches, based on the name matching results, was implemented using SQL queries in the database. Only the recorded positive matches contributed to the calculation of this experiment's results. Figure 15, shows schematically how the algorithm works.

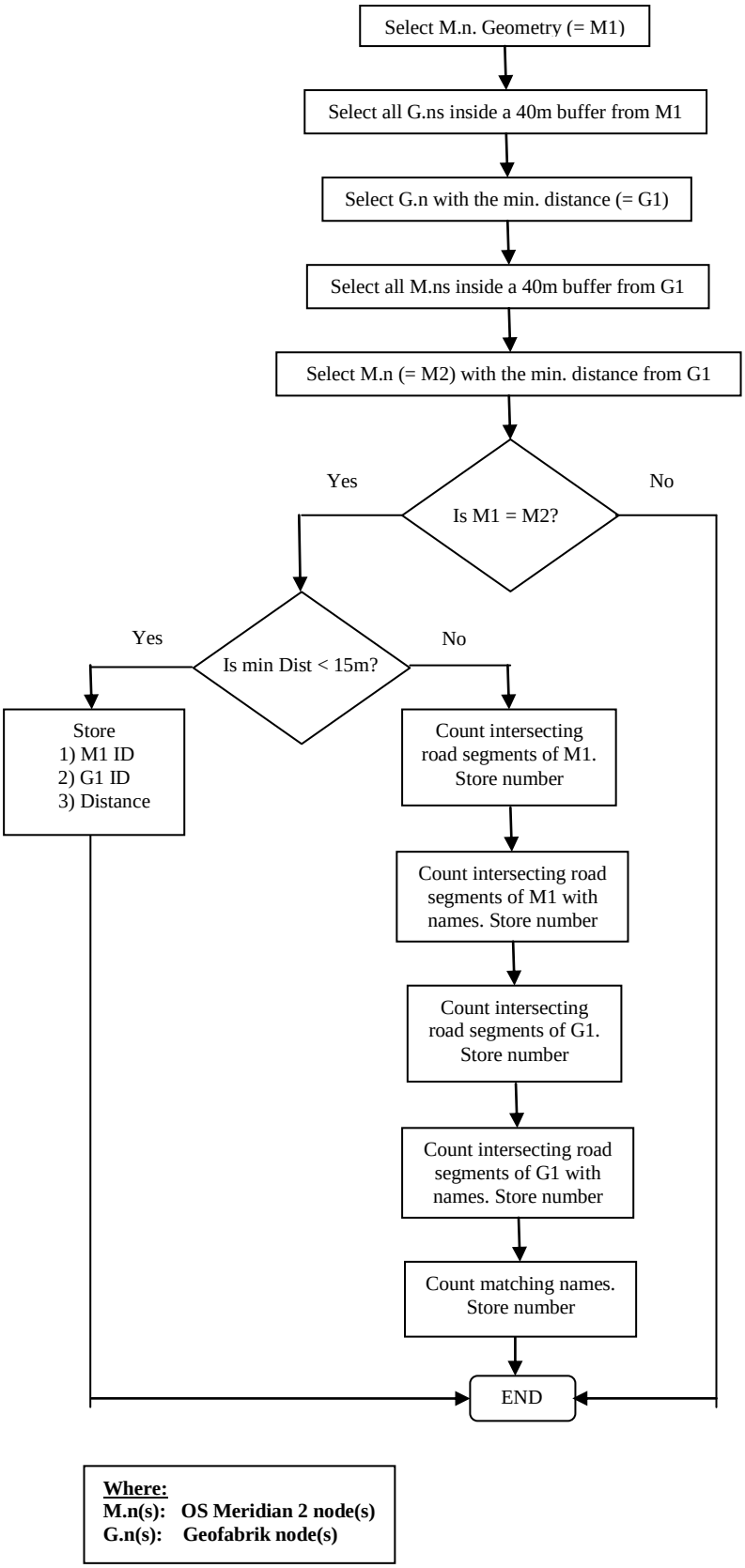


Figure 15. The positional accuracy algorithm applied to each Meridian 2 node

3.4.1.3 Positional accuracy and completeness correlation

The outcome of the positional accuracy evaluation process was also summarised into tiles, using the 1km^2 National Grid, and the average positional accuracy was computed creating a new dataset. This step was necessary to perform a comparison experiment with the data produced by Haklay (2010) in regards with the completeness of OSM data. It must be noted that the completeness experiment was conducted also using the OS Meridian 2 data as the reference dataset and its results was also summarised for the National Grid 1km2 tiles. In this context, the comparison of the two results would be able to provide meaningful outcomes. This final step regarding the positional accuracy examination included the spatial correlation between the completeness and positional accuracy in an effort to realise whether there is a connection or similarity in the trends between these two quality elements. Definitely the results from the OSM positional accuracy examination are very important in an effort to determine the degree of usability of UGSC. Another component that will assist this effort is the analysis of attribution process as described in the next Section. But the overall role and impact of the OSM dataset as a source for GI retrieval purposes at a national level takes on significance when these results are further examined under the prism of existing social patterns. Since Haklay (2010) has demonstrated the correlation between completeness and the social index of multiple deprivations, it would be particularly interesting to examine whether that correlation expands also to the positional accuracy quality element.

3.4.1.4 OSM attributes (tags) quality evaluation

As described earlier, there is a totally democratic process in place that determines the way of the real world objects attribution. This openness in the process perhaps is one of the key factors for the popularity of OSM. However, in the first leg of the OSM data research (Section 3.2.1.1) it was realised that this freedom created a lot of inconsistencies regarding both the categorisation of the spatial entities but most importantly in the attribution of the entities. This was also affirmed by the OSM community itself as it was realized that there was a need for some form of quality assurance mechanism that would enable users to correct inaccuracies. Indeed, today

there are a variety of efforts that try to achieve that. Examples of such efforts can be found in a separate section of the OSM wiki pages: http://wiki.openstreetmap.org/wiki/Quality_Assurance. These early attempts for identifying and correcting errors in the OSM dataset, while they present interesting paradigms of a self-correcting mechanism for a crowdsourced Web 2.0 application, are still incomplete and patchy. The 'Keep Right' application, which was created by the OSM community and monitors the violation of some OSM rules, provides a prime example. The application evaluates the OSM data conformance against an arbitrary set of pre-defined rules and presents to the users the positions of possible mistakes. Yet, there is no consistent way of data quality quantification or a firm methodology for data quality monitoring and reporting.

This study instead of focusing directly into the corrections that the OSM end-product needs, takes a step backwards and examines the data generating process. The aim is twofold. On the one hand, it will examine how the attribution of the spatial entities is implemented and realise if there are any erroneous parts in this crowdsourced methodology. On the other, it will provide an OSM product specification. As the aim of the research is to evaluate the quality of the OSM attributes, there is a need for a specification that would be used as reference for the evaluation. The interesting point here is that OSM does not provide a data specification other than the wiki pages. Thus, in order to proceed with the evaluation there was a need to translate the guidelines included in the wiki pages into a set of rules and then use these rules to evaluate the actual data produced. These two steps will provide the knowledge needed so as to correctly interpret the results of the OSM attribute quality evaluation. At the same time, it will highlight methodologies and practices where special attention should be paid in order not to introduce errors into UGSC. Finally, experience will be gained for improving the OSM data generation process.

In an effort to understand the process of data attribution, an extensive study of the wiki pages that describe the methodology of the data attribution took place. The starting point for this study was the http://wiki.openstreetmap.org/wiki/Map_Features OSM wiki page that serves as the container of the OSM data description. The web page further links to pages that describe a real world object that users are encouraged to capture. These pages

are actually the implementation of the OSM community common decisions and as such they contain the rules that OSM contributors are asked to follow. This set of rules functions as a user guide for the creation of OSM spatial data and it can be considered as the specification of the OSM product.

It must be noted here that the ISO 9000:2005 provides different definitions for guidelines (i.e. documents stating recommendations or suggestions) and specifications (i.e. documents stating requirements) (ISO 2005d). Given the ISO approach on these concepts, it is understandable that the OSM wiki pages are much closer to the former concept as the content of the pages is mainly instructive. Yet, these wiki pages are the closest evidence of the OSM data specification available. After all, there is a direct link between the two documents: guidelines are written in such a way that if followed closely the outcome will fulfill the requirements and thus it will meet the specifications. Hence, the effort was to extract a set of rules from the wiki pages that would function as an OSM specification.

After studying and analysing the rules introduced in the wiki pages, the outcome was a formalised version of the knowledge contributed by the users (this work did not cover the entire OSM data range due to time limits). To practically implement that formalisation the development of an XML schema took place (see also Section 6.2). The wiki pages study process was used to translate the user defined OSM rules into an XML schema that would formally describe the OSM data specification. The completion of this part of the research gave a different perspective to the OSM quality evaluation. Now it was possible to accurately measure the level of the dataset conformance against the product specifications. But first the data collection step had to be completed.

By using the original OSM IDs contained in the shapefiles downloaded by Geofabrik it was possible to access and store the tag values from the OSM servers. An automated procedure was developed (similar to the one described in Section 3.3.1.3) that used the OSM API to get the system defined attributes and the user assigned tags for each one of the spatial entities that belonged in the OSM Highways and POIs datasets of Great Britain. Each OSM XML response was parsed and the necessary data were stored into a PostgreSQL database.

After the successful data collection two final experiments were made. The first focused on the analysis of the tags. Issues like the number of tags per each thematic layer or per spatial entity were examined in an effort to realise the effect that the open crowdsourced process of attribution has on the overall dataset. Lastly, the availability of the attributes/tags dataset collected in conjunction with the OSM product specification constructed according to the wiki pages, enabled the implementation of a series of ISO-based tests to evaluate the quality of the OSM data against its own specifications. These ISO test were conducted for a number of OSM Highway and POIs categories. The tests provided a clear quantification of the OSM quality and a commonly accepted language in which it was feasible to convey the results of the OSM quality elements examination.

In order to evaluate the quality of the OSM attributes, the methodology described in the ISO 19114:2005 Specification was adopted (ISO 2005c). More specifically, the method dictates the development of a series of quality evaluation tests. Each evaluation test is designed to examine a specific data quality component of a data quality element (Table 2).

Data quality component	**Component domain**
Data quality scope	All items classified as Highways
Data quality element	Enumerated domain 1 – completeness 2 – logical consistency 3 – positional accuracy 4 – temporal accuracy 5 – thematic accuracy
Data quality sub-element	Enumerated domain (Dependent upon data quality element)
Data quality measure	
Data quality measure description	Free text
Data quality measure identification code (ID)	Enumerated domain
Data quality evaluation method	
Data quality evaluation method type	Enumerated domain 1 – internal (direct) 2 – external (direct) 3 – indirect
Data quality evaluation method	Free text or citation (depends

		description	on data quality evaluation method type)
	Data quality result		
		Data quality value type	Enumerated domain 1 – Boolean variable 2 – number 3 – ratio 4 – percentage 5 – sample 6 – table 7 – binary image 8 – matrix 9 – citation 10 – free text 11 – other
		Data quality value	Record (Depends on data quality value type)
		Data quality value unit	(Depends on data quality value)
	Data quality date		ISO 8601:1988
	Conformance quality level		value or set of values

Table 2. A typical ISO 19114:2005 evaluation test.
(Source: ISO 2005c)

* * *

This step of the OSM attributes quality evaluation, along with the completeness and the positional accuracy results, largely complete the puzzle of the overall OSM data quality. The systematic method adopted for the approach of the quality issue allows both producers and users of the OSM data to validate how accurately the data meets the product specifications (even if these specifications are loosely defined) and thus make safe judgments regarding the suitability of the data for specific tasks and applications.

3.5 Summary

Although the entire process of the analysis is conducted from the point of view of the mapping agencies (either national or private sector), it cannot be supported that all relevant issues have been thoroughly considered. Perhaps the most important point that has not been part of the scope of this analysis is the IPRs of the content available from

the selected sources. As discussed earlier (Section 2.2.1), IPRs might not be a major obstacle in the engagement of NMAs with UGSC.

The methodology described here provides a variety of experiments that examine the UGSC phenomenon from different angles. As it is expected of mapping agencies (private or national) to exploit any trends and technological advances, both in the IT and the Geomatics domains, the results generated from this study will provide a solid knowledge base for the future steps on the subject of UGSC. The experiments were chosen and conducted in such a way that their individual outcomes will be the components of an overall resultant that will make mapping agencies realise the nature and the potentials of the phenomenon and allow them to judge UGSC on its merits.

The implementation of the methodology described here is expected to affect the conceptualisation of UGSC phenomenon. As described, even from the early steps of the analysis the important role of space, and consequently the importance of the spatial orientation of a Web 2.0 geo-application were clearly revealed. At the same time, the analysis will provide answers to important issues such as distribution, positional accuracy and attribution validity of the data provided. Moreover, as a research tied up solely in the technical details of the user generated content process would fail to realise the important role of the human factor, the analysis expands sufficiently into the study of participants' behaviour, either directly as in the case of photo-sharing Web sources or indirectly by jointly analysing the outcomes of this and of existing research. Overall, the analysis will provide solid answers to the research questions posed at the beginning of this Chapter.

Chapter 4

Results of the geo-tagged photos analysis

4. Results of the Geo-tagged Photos Analysis[6]

4.1 General

In Chapter 3 the methodology's road map for the evaluation of the two UGSC's main types was described. This Chapter will present the findings and results of the experiments conducted with photo-sharing, Web 2.0 applications. As discussed in Section 3.2, for this part of the research the applications selected to be examined are Flickr, Picasa, Panoramio and Geograph.

Flickr belongs to the social networking family of the Web 2.0 applications. The main concept behind Flickr is that it uses photos as the medium to enable its users' interconnection. Flickr's users, after creating a personal account, can upload photos, manage their sharing policy and either make the submitted content freely available to all or provide access to specific users or disabling the access altogether. Flickr claims that pursues two main goals (Flickr 2010):

> *"We want to help people make their photos available to the people who matter to them.*
> *...*
> *We want to enable new ways of organizing photos and video."*

Despite the fact that the concept behind Flickr and the two explicitly articulated goals seem to be irrelevant with any form of spatial data and even more so with mapping agencies, in fact Flickr has created a Web platform that enables the submission of spatially related content. This is realised through the ability to contribute geo-tagged photos (see also Section 3.3) to the application. Currently, Flickr hosts more than 4

[6] Most parts of this Chapter have been adapted from:

Antoniou, V., Haklay, M., Morley, J. 2010b. Web 2.0 Geotagged Photos: Assessing the Spatial Dimension of the Phenomenon. *Geomatica (Special issue on VGI),* 64(1), pp.99-110.

billion photos (Yahoo! 2009) where more than 100m of them are geo-tagged (Yahoo! 2010). Not much different is the concept behind Google's Picasa Web Albums. Google, like Yahoo!, has set up a social network around photos. Both companies have leveraged their position in Web mapping applications (i.e. Yahoo! Maps and Google Maps) and introduced to their users the ability to place their photos on maps. For Panoramio though, things are a little bit different. The Web application started aiming to provide the means to the users to post their photos from places they have been to or like and thus the concept of space was present from the beginning. On top of that, a social network was built that enabled users to comment photos and socialise with other users. Finally, Geograph was firmly oriented towards the description of space through photographs as it was systematically evangelising the full coverage of Great Britain with geo-tagged photos.

4.2 Chapter's overview

The analysis will start with the presentation of the descriptive statistics of the datasets so to understand the fundamental numbers that describe the phenomenon in each case, followed by an analysis of the photo submission per $1km^2$ tile for each source. Then, the spatial dimension of the phenomenon will be examined by analysing the spatial distribution of the data. This process will add the importance of the space element in the realisation of the phenomenon. Next, the users' contribution in geo-tagged photos will be examined against population data through the construction of expectation surfaces. The analysis of the user-generated datasets under the prism of the users' location will reveal how the former is affected by the latter. The first phase of the analysis will be completed with the comparison of the findings of the four sources.

The result of this comparison will be helpful in understanding the nature of the photo-sharing, Web 2.0 applications along with their fundamental differences and commonalities. This will enable the development of a typology that divides the photo-sharing Web applications in two broad classes taking into account their nature and the content generation process. Throughout this first phase of analysis, particular attention will be paid to how space affects the overall behaviour of such sources and whether the

results justify their use as universal sources of GI retrieval for the purposes of mapping agencies. The knowledge gathered from this first phase will be used to gain further insights in the following steps.

More specifically, in the next level a more focused analysis will be carried out by comparing a spatially implicit and a spatially explicit source (see also Sections 3.3.1.4 and 7.4 for more on this typology). The start will be made with the comparison between the volumes of geo-tagged photos submitted to the most popular areas in each one of these different types of applications over a period of 18 months. This will reveal whether users' contribution differentiates between implicit and explicit sources. Furthermore, the spatial distribution comparison will move from the national level to a smaller scale (i.e. examine how the phenomenon is realised inside the $1km^2$ tiles) in order to better understand the behaviour of these two types of UGSC sources. This will be achieved by examining the spatial distribution of geo-tagged photos (for both implicit and explicit Web applications) in 15 test sites in the UK. Next, the analysis will turn to the examination of the users' behaviour in an effort to realise whether there is a difference in the user-base between the spatially implicit and explicit sources. Finally, as the demarcation line between implicit and explicit sources is clearly revealed by the analysis, the focus will turn on the ability of the spatially implicit sources to provide a sufficient pool of UGSC so to provide the necessary volume of data to support, in a national level, the aims of mapping agencies.

4.3 Web sources comparison

4.3.1 Descriptive statistics

The first step of the analysis is the calculation and the examination of the descriptive statistics of each application using the National Grid's $1km^2$ tiles as a measurement unit. Although this is a fairly simple and straightforward process, it gives a clear indication of the magnitude and the characteristics of the phenomenon. So for example, the fact that 1.65m geo-tagged photos have been submitted to Flickr for Great Britain provides a solid understanding of the data volume and it can trigger the interest of mapping

agencies for possible ways to exploit this data repository. Additionally, although it was expected that there was going to be tiles with many geo-tagged photos submitted to them, it is particularly interesting to note that there is one tile that has received more than 38,000 photos. Questions are raised regarding the popularity of this specific area, and consequently the need for literally immediate updating of such high-demand areas. Table 3 provides the combined summary of the descriptive statistics for all four sources.

	Num. of Tiles	Min	Max	Sum	Mean	Std. Deviation
Geograph	238920	0	1317	1078150	4.51	12.95
Flickr	238920	0	38506	1654277	6.92	196.56
Panoramio	238920	0	10191	410236	1.72	34.97
Picasa Web Albums	238920	0	31947	1255515	5.25	136.46

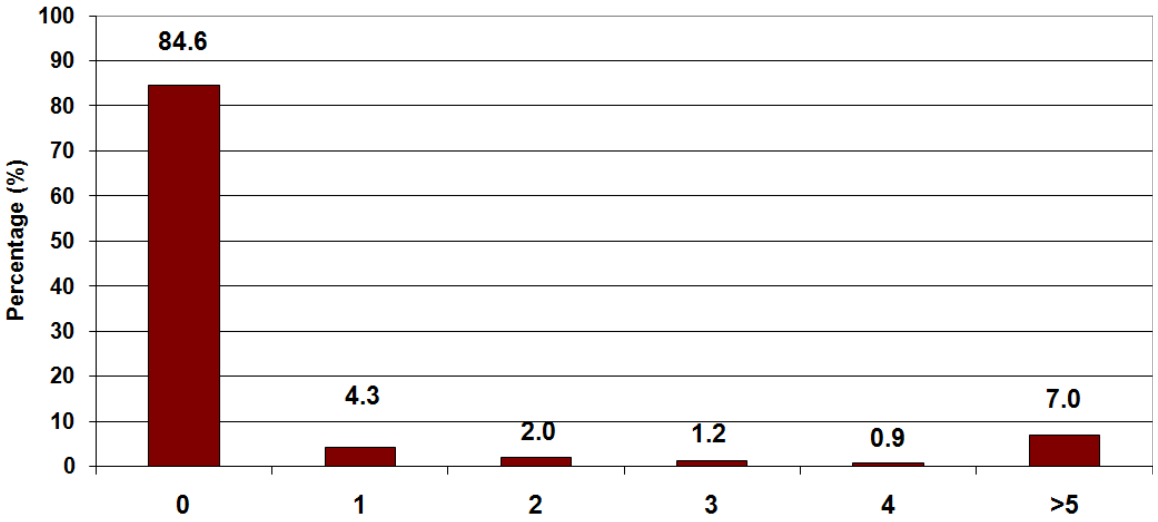
Table 3. Descriptive Statistics of the photo-sharing Web applications examined.

The source with the most geo-tagged photos for Great Britain is Flickr with 1,654,277 photos, followed by Picasa Web with 1,255,515, Geograph with 1,078,150 and last Panoramio with 410,236. Yet, despite the huge volume of geo-tagged photos recorded for each source, the average number of photos per tile is quite small for all four sources with Flickr holding the maximum of the means (6.92 photos per tile) and Panoramio holding the minimum with 1.72 photos per tile (it must be noted that Panoramio's API failed to report back the tiles that had one photo submitted to them and thus the results are open to interpretation).

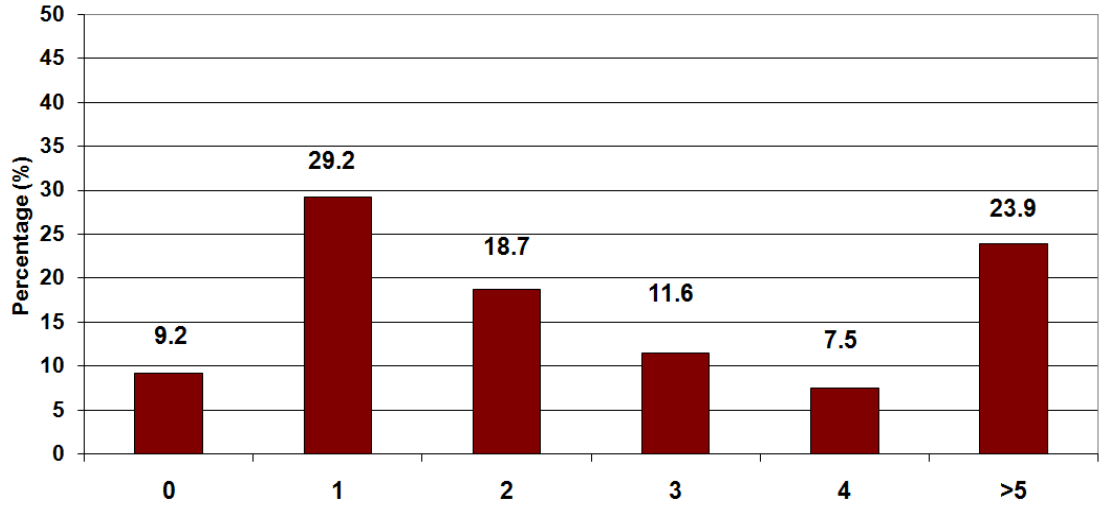
On the other hand, in contrast with the small average values is the maximum numbers of photos per tile which also gives an indication of the different level of popularity among the sources. This is evident when comparing the Geograph's maximum value (1,317 photos) to the maximum values of Flickr (38,506) and Picasa Web (31,947). Panoramio's maximum value (10,191) is located in between these two groups, resembling more to Geograph than to the other two sources. A similar observation can be seen when examining the Standard Deviation (Std. Dev.) of each datasets. Geograph has the smallest Std. Dev. (12.95) whereas Flickr and Picasa Web have 196.56 and 136.46 respectively. Once again Panoramio with Std. Dev. equal to 34.97 is closer to Geograph than to the group of Flickr and Picasa Web.

4.3.2 Geo-tagged photos per tile

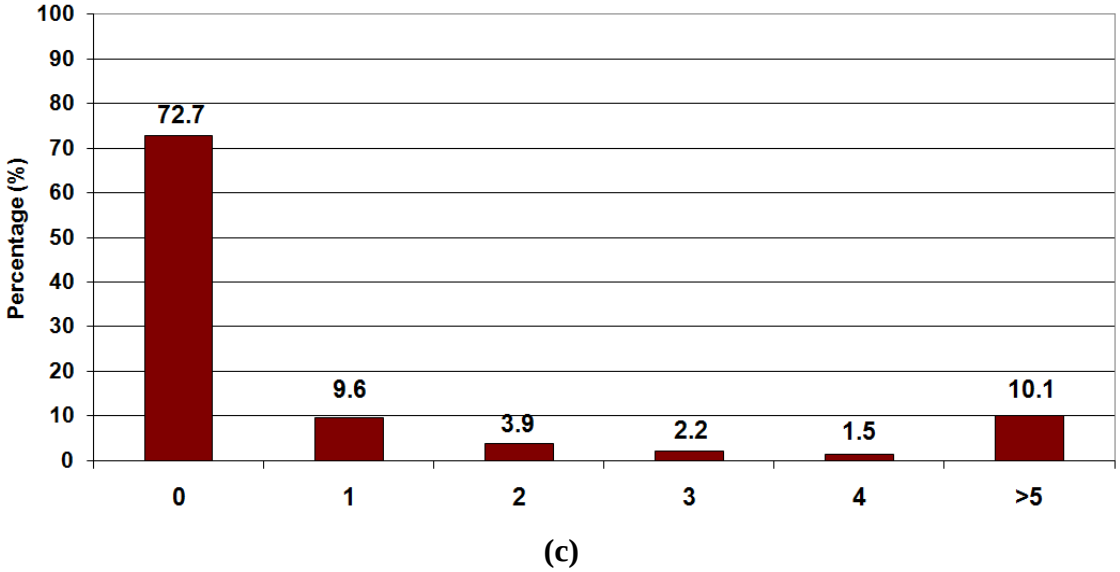
Even though the average number presented earlier is a mathematically correct number, in a power-law distribution it does not give a completely accurate picture of the underlying reality. For that matter, it is constructive to present the frequencies of the number of photos submitted per tile (Figure 16).



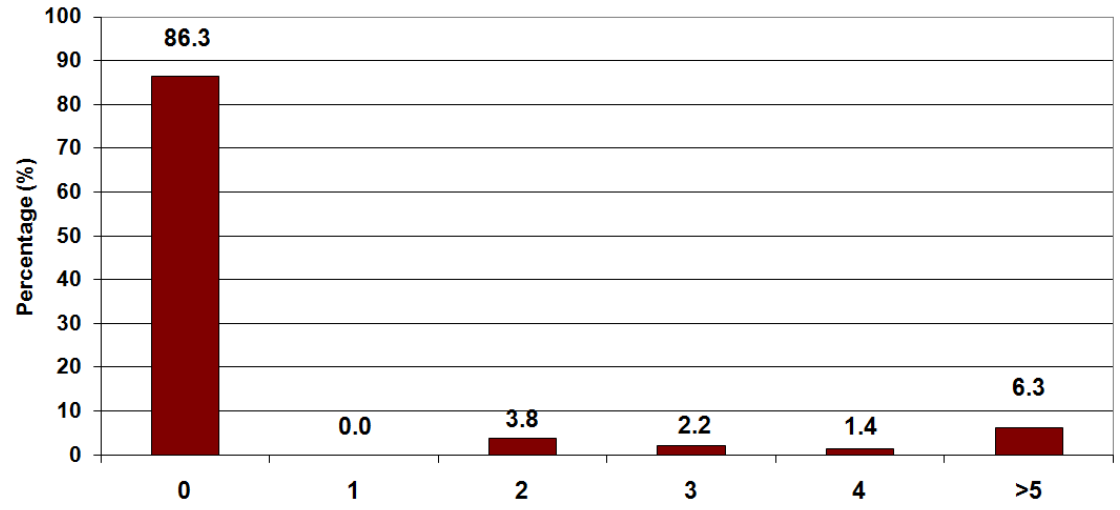
(a)



(b)



(c)



(d)

Figure 16. Frequencies of photos per tile for (a) Flickr, (b) Geograph, (c) Picasa and (d) Panoramio

It is evident that, for all the sources except Geograph, for the majority of the tiles there are no geo-tagged photos that have been submitted since the Web 2.0 applications were launched. More specifically, for Picasa Web there are no photos for the 72.7% of the tiles, for Flickr the percentage climbs to 84.6% and for Panoramio 86.3%. In contrast, for Geograph only the 9.2% of the 1km^2 tiles in Great Britain is not covered with at least one geo-tagged photo.

Another interesting point revealed from the presentation of this data is that for three of the Web 2.0 applications examined, just a very small part of the area in scope (i.e. Great

Britain) has been repeatedly covered by five or more geo-tagged photos. For Flickr a mere 7% of the space manages to surpass this threshold where as for Picasa the figure raises slightly to the 10.1%. For Panoramio the number is even smaller (6.3%). A simple observation of the four graphs presented can reveal that the Panoramio's behaviour is considerably more similar to the Flickr and Picasa's results compared to the Geograph's as originaly thought. Indeed, Geograph stands out of the group of Web sources examined as almost 24% of Great Britain has been covered by five or more geo-tagged sources in the course of Geograph's activity.

The importance of this fact is further highlighted when the overall photo contribution of each source is taken into account. This can be seen when for each source the number of tiles that belong to each category (i.e. 1, 2, 3, 4 or 5 or more photos per tile) is normalised by the total number of geo-tagged photos submitted to each source (Figure 17). It can be seen that taking into account the overall users' contribution, the Geograph is the most productive source in terms of covering space repeatedly with geo-tagged photos. On the other end of the spectrum and despite the fact that Flickr has almost 32% more geo-tagged photos than Picasa, it is falling behind when it comes to efficiently covering the space with photos.

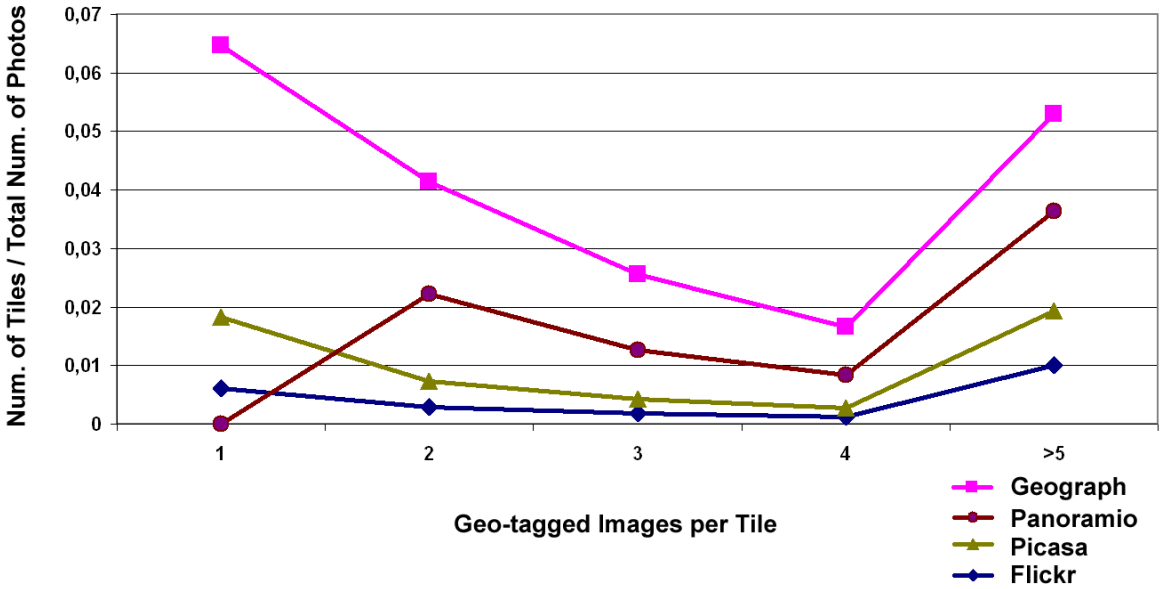


Figure 17. Number of Tiles covered by geo-tagged photos normalised by the total number of photos submitted to each source.

4.3.3 Spatial distribution

After the completion of the data gathering process, an effort to analyse and understand the phenomenon of UGSC and how this is realised through the photo-sharing Web applications started. The first step was the presentation of the descriptive statistics for each dataset. While this gave a basic understanding mainly of the magnitude of the phenomenon it did not reveal the whole picture. The second step towards that direction was the examination of the number of geo-tagged photos submitted to the tiles of the $1km^2$ National Grid. This step enabled the observation of differences and similarities among the four sources but still the whole picture remained elusive. As this is a spatially related phenomenon, this gap can be filled only when the impact of geography is considered. For that matter, the $1km^2$ National Grid of Great Britain was associated (joined) with the data collected from the Web sources and the results (i.e. the spatial distributions of the datasets) were visualised. Figures 18, 19, 20 and 21 show the spatial distribution from each source's photos.

The spatial distribution of Flickr's photos is shown in Figure 18. The first clear observation is associated with the figures discussed in Section 4.3 and has to do with the number and the location of the tiles that have no geo-tagged photos submitted to them. There are large areas of Great Britain that are fairly empty on the one hand and there is the formation of clusters (small and larger ones) on the other. At first sight it is obvious that the larger clusters are formed in the main urban areas of Great Britain. For example, the larger cluster is located in London, followed by clusters in other main cities like Manchester, Edinburgh, Birmingham, Glasgow and Bristol. Also, there is an increased presence of geo-tagged photos in smaller cities and especially in those located in the south coast like Plymouth, Southampton, Portsmouth and Brighton. This preference over tourism or leisure areas is observed also in other parts of Great Britain like the Lake District, the National Park of Snowdonia and Stonehenge. At the other end of the spectrum are the rural and the less popular parts of Great Britain. For example, much of Wales and Scotland, the east coast and the northern areas of England have not been covered by any geo-tagged photos. Yet, an interesting observation in these particular areas is the formation of the outline of the major highways by the photo's traces. An

indicative example is the A82 highway that connects the city of Inverness with Fort William and travels along side the famous Loch Ness in Scotland.

Similar patterns with the one described for Flickr are formed from the spatial distributions of Picasa (Figure 20) and Panoramio (Figure 19). In the latter there are fewer and smaller clusters than the ones appearing in Flickr. This is largely explained by the fact that Panoramio has approximately the ¼ of Flickr's geo-tagged photos. In contrast Picasa appears to provide a slightly better coverage of the area in scope, an observation that was also discussed in the previous Sections. Yet, overall, both patterns are not much different compared to Flickr's. Their main clusters are spotted at the urban and popular places while the places with no photos submitted are located in the rural and less popular areas.

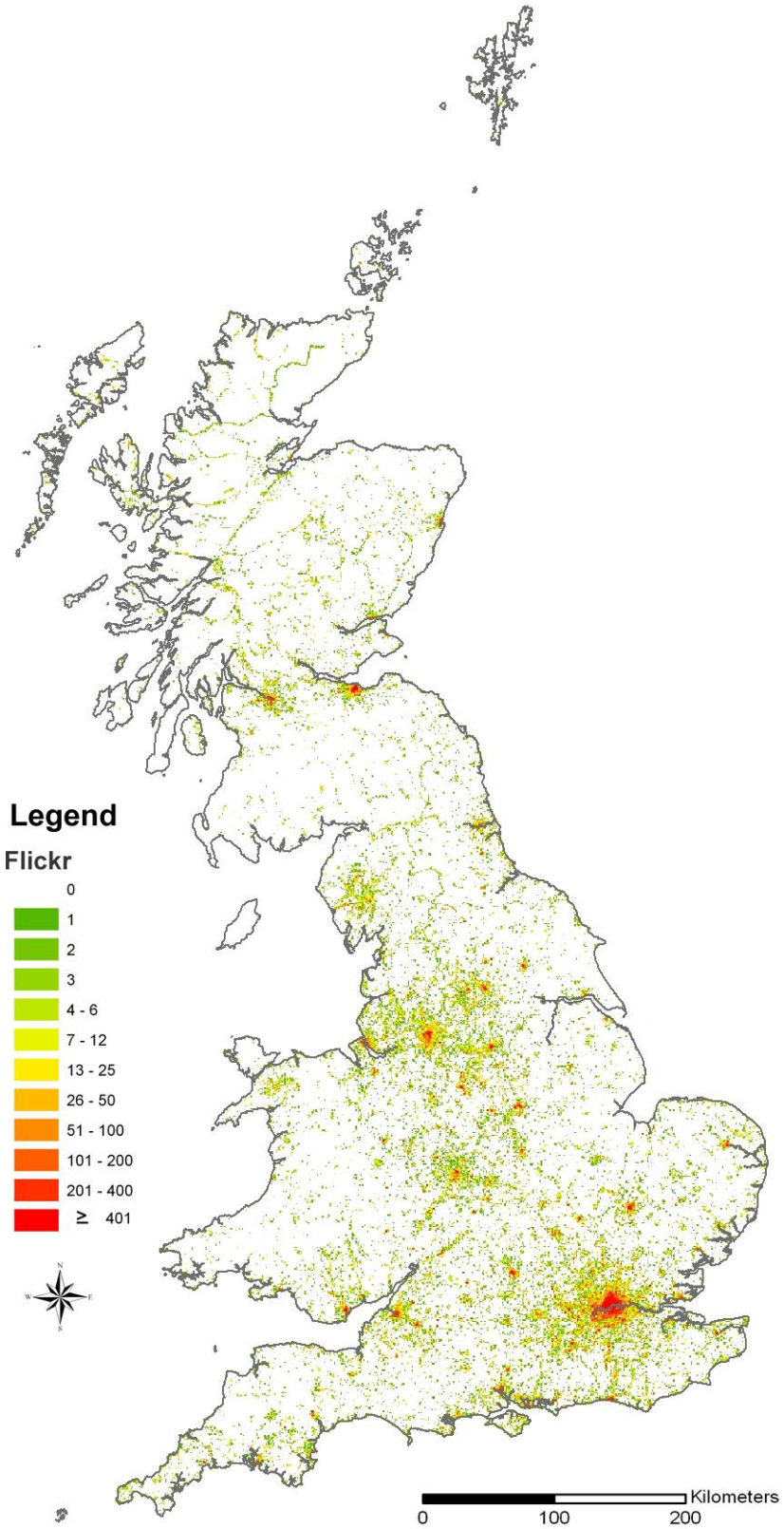


Figure 18. Spatial distribution of Flickr's geo-tagged photos

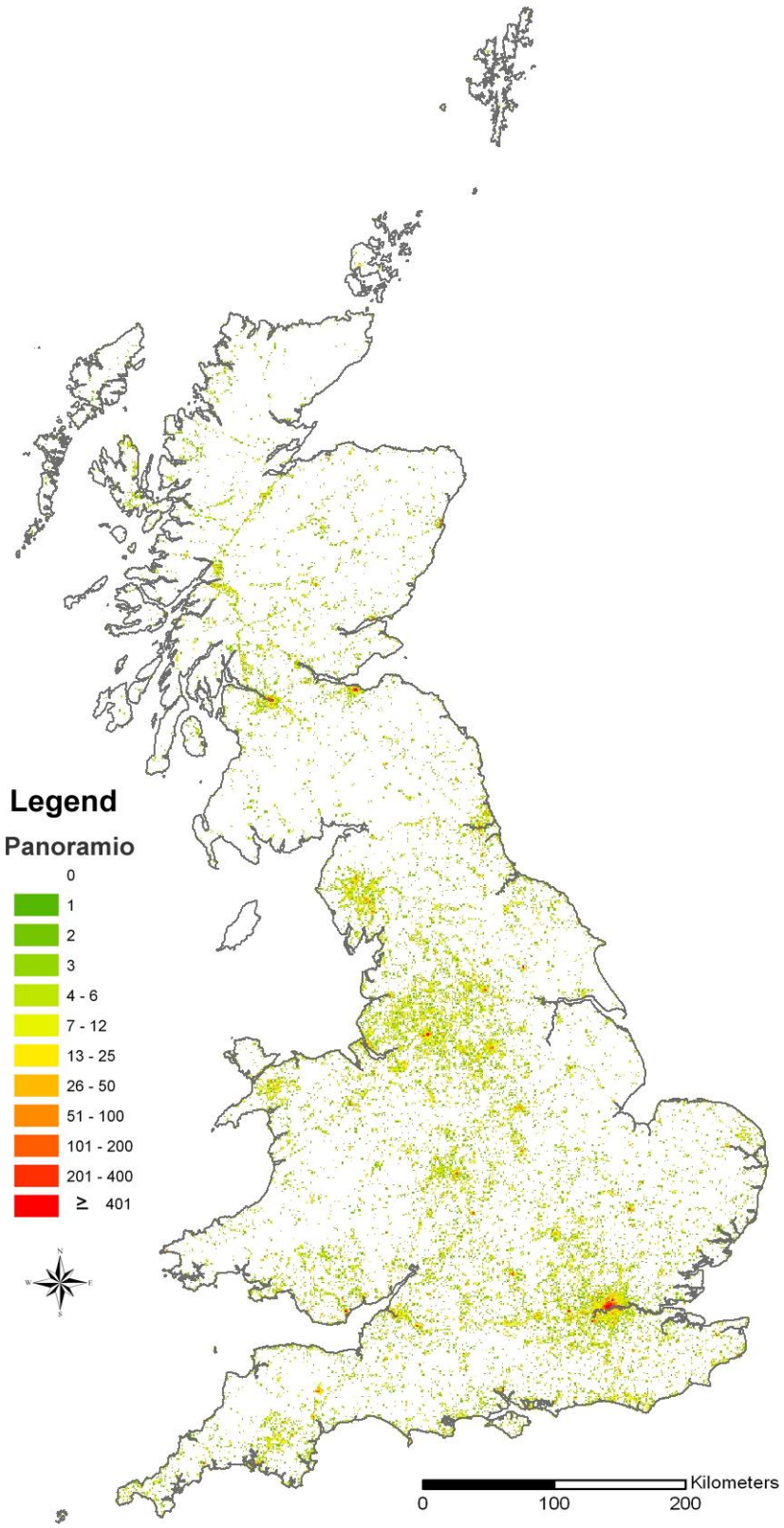


Figure 19. Spatial distribution of Panoramio's geo-tagged photos

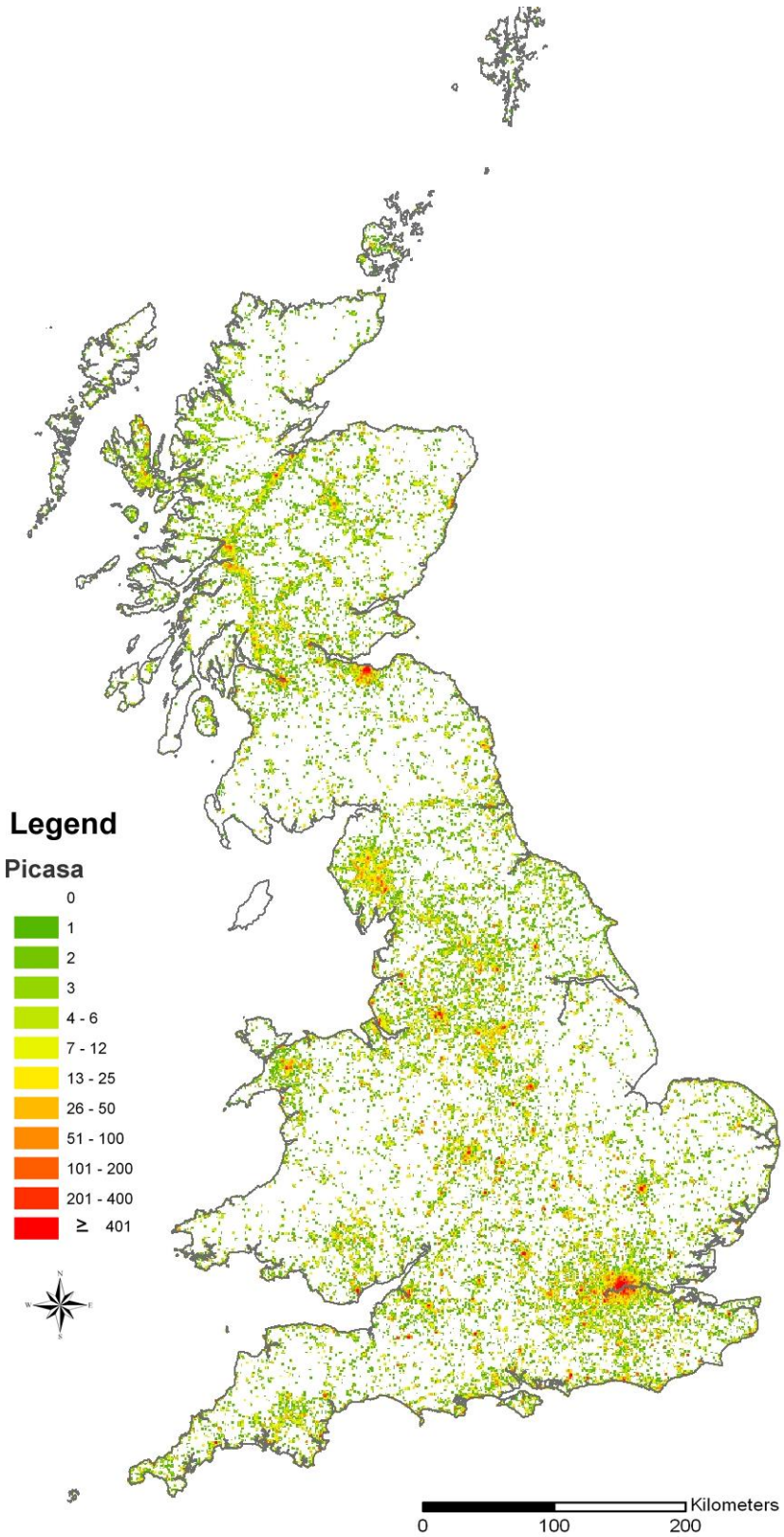


Figure 20. Spatial distribution of Picasa's geo-tagged photos

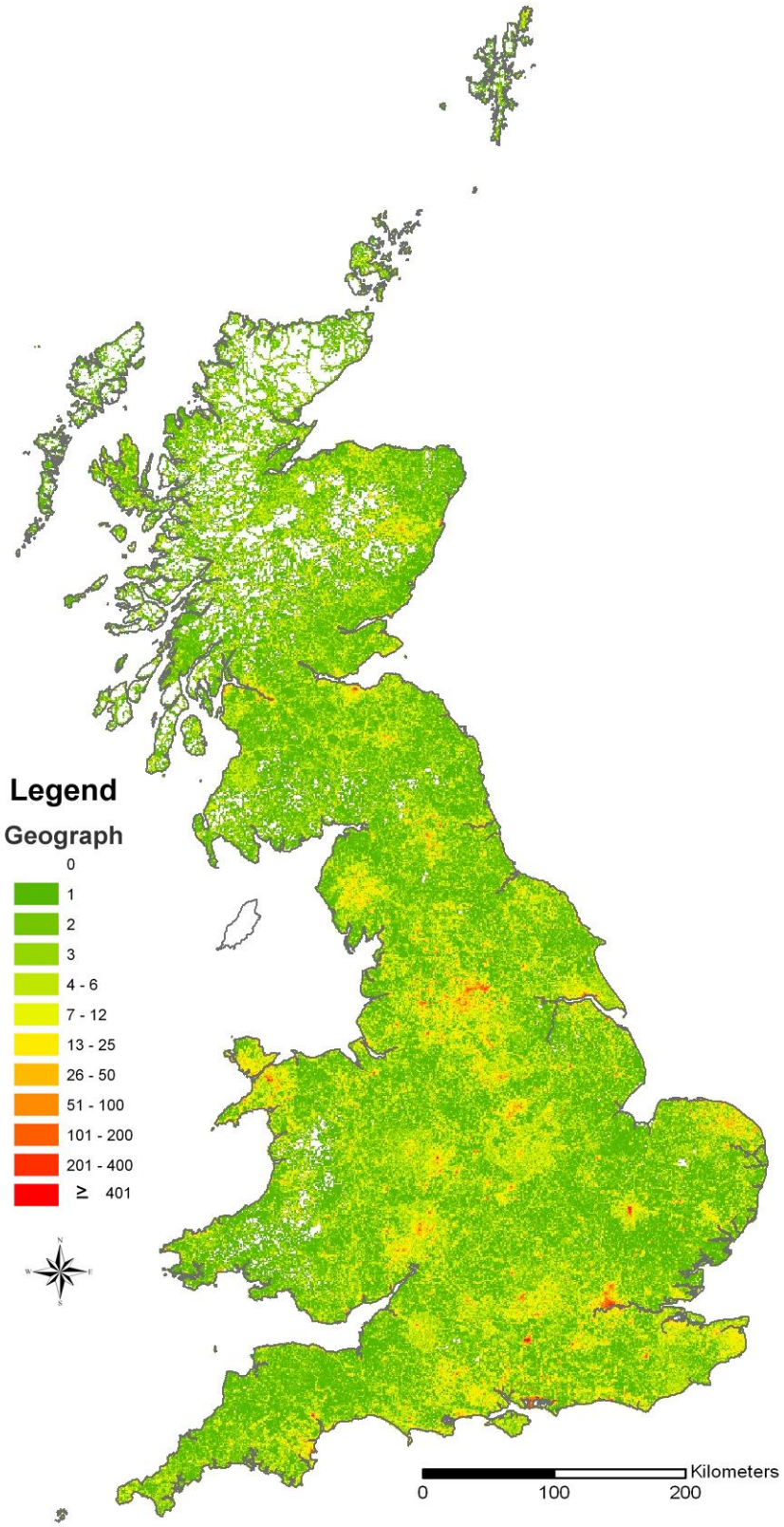


Figure 21. Spatial distribution of Geograph's geo-tagged photos

However, a totally different pattern, appears when examining the spatial distribution of Geogrpah's photos (Figure 21). The geo-tagged photos submitted by the users cover most of Great Britain with very few blank spots mostly in the sparsely populated, barren areas of the Highlands in Scotland. On top of that, there is a large number of clusters located in the urban and tourism areas. These clusters are more stretched and not that intense (see the maximum values in Table 3 and Figure 23) as the ones recorded in the previous Web applications thus covering considerably larger areas. For example, for the Greater London area a total of 35,275 photos have been submitted which approximately corresponds to 22 photos per km^2. This number is almost 5 times higher from the overall average of 4.5 photos per km^2. In contrast, the 338,198 photos that have been submitted to Picasa Web for the Greater London area correspond to an average of 211 photos per km^2, and this is 40 times more than the overall average of only 5.25 photos per km^2. In fact, Geograph's photos cover about 60% with 1 to 3 photos, another 30% is covered with 4 to 50 photos and just 1% (which corresponds to approximately 2,400 tiles) is covered with more than 50 photos. The rest 9% of the space remains uncovered (Figure 22).

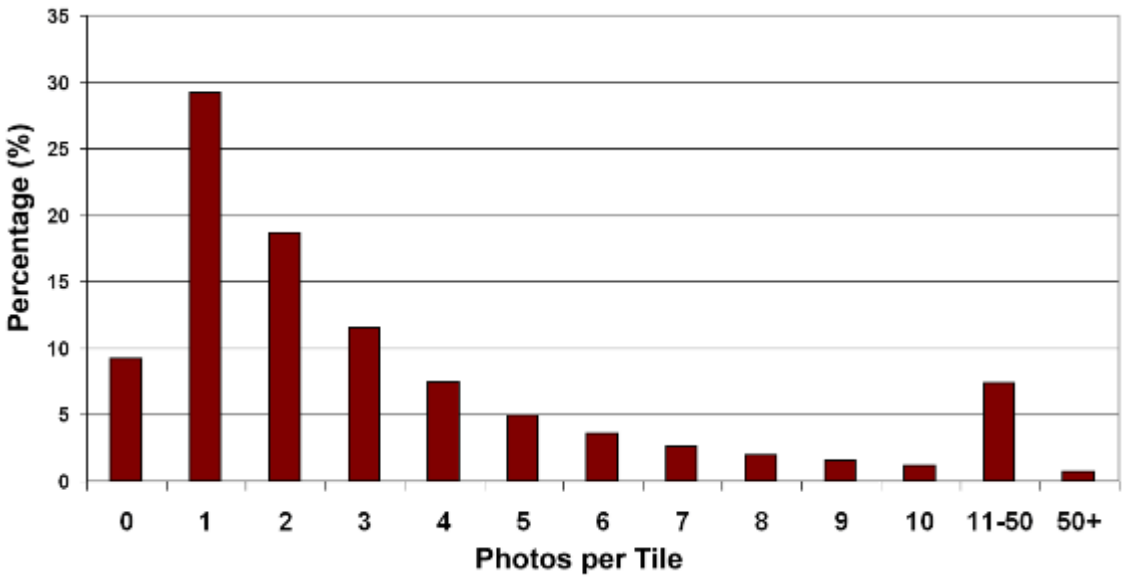


Figure 22. Frequencies of geo-tagged photos per Tile for Geograph

Finally, using 3D visualisation (Figure 23) we can understand the magnitude of the phenomenon on the clustered areas.

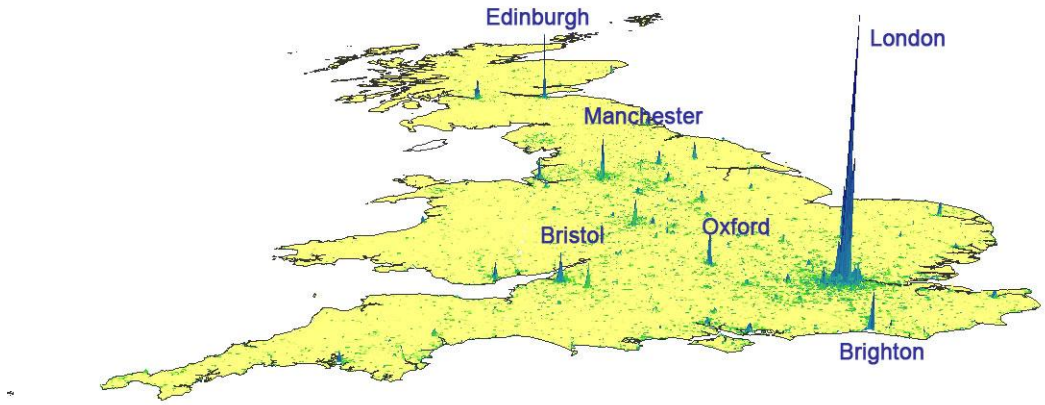


Figure 23. 3D visualisation of the geo-tagged photos collected from Fickr

A final observation at this point is that if the analysis of the number of photos submitted to each tile is confined only in the areas where there are available photos (i.e. exclude the empty tiles) the comparison of the frequencies of the photos per tile reveal a similar pattern for both explicit and implicit applications (Figure 24). Both types of sources cover the areas in scope (national-level for the explicit and the popular and tourism for the implicit ones) basically with one photo per tile and they provide a more substantial coverage for a smaller part of the areas. Here the implicit sources are performing better than the explicit one as the percentage of tiles that are covered with 10 or more photos is considerably larger. Therefore, it can be suggested that the role of explicit and implicit Web applications as sources of UGSC can be complementary in a national level: the implicit ones will provide huge volumes of data for the urban and tourism areas and the explicit sources for the rest of the area. This hypothesis will be further tested in Section 4.4.

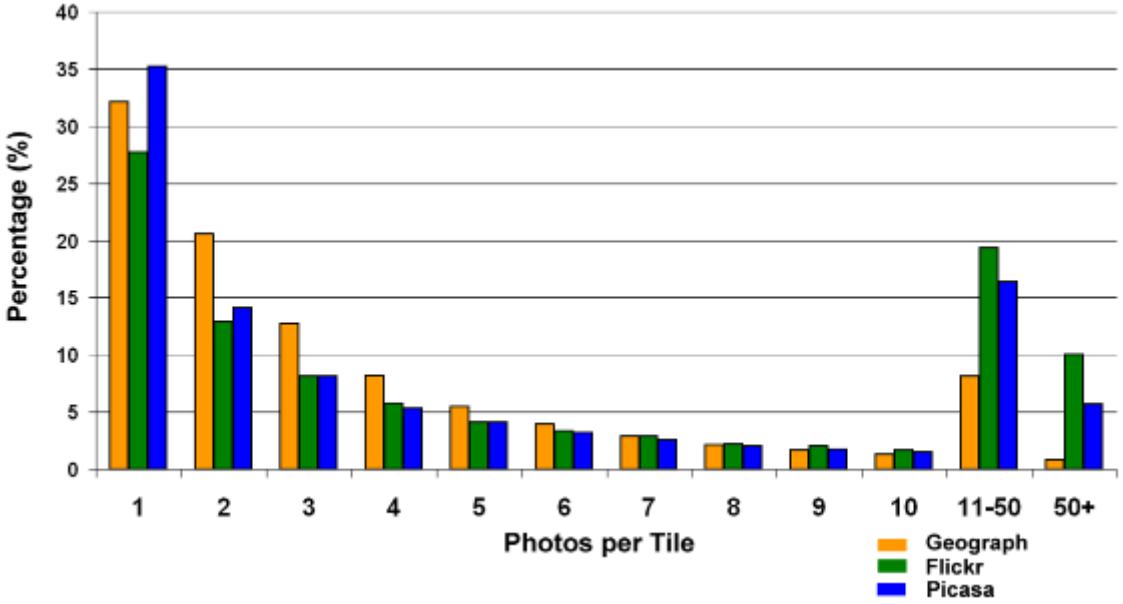
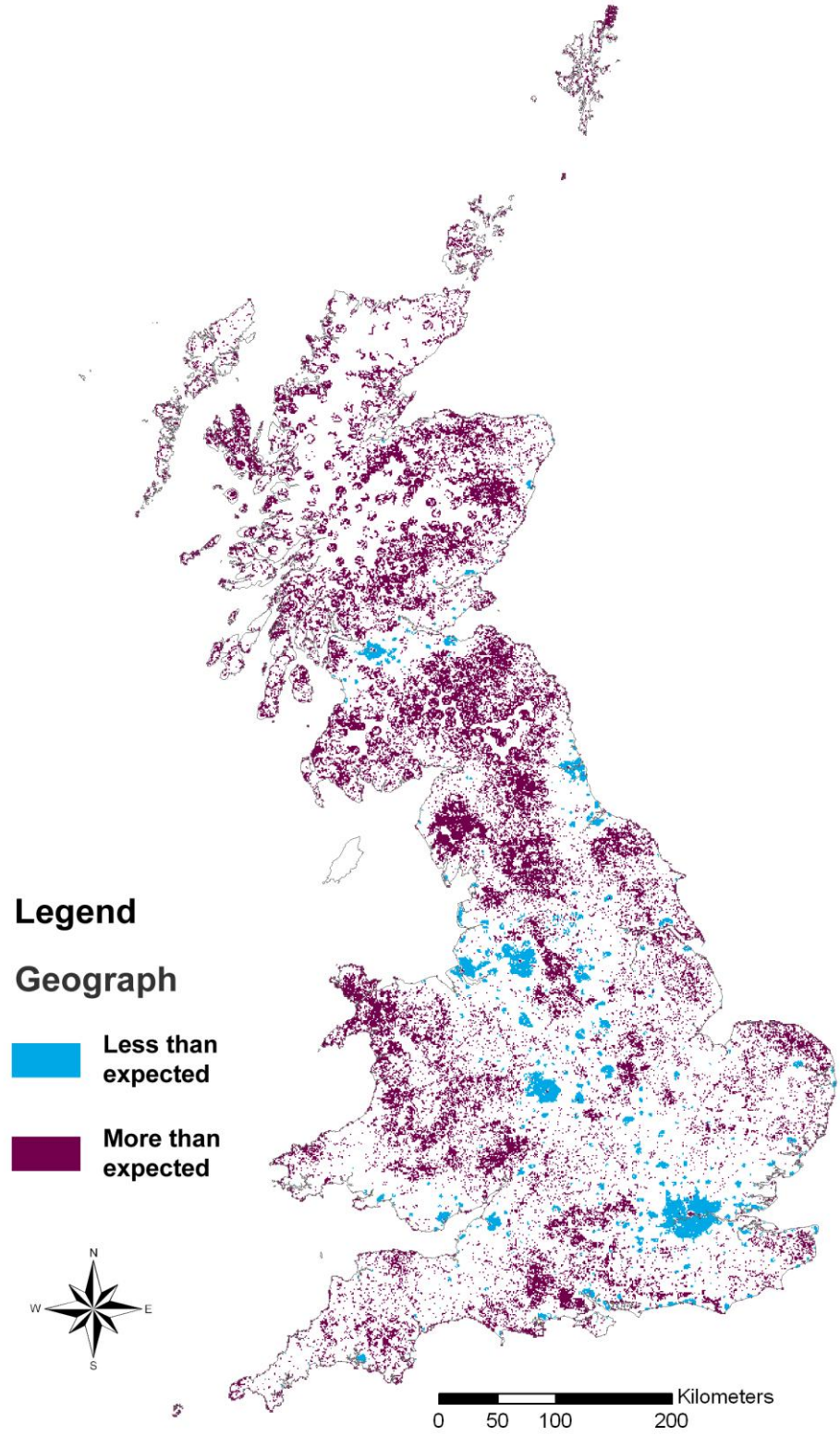


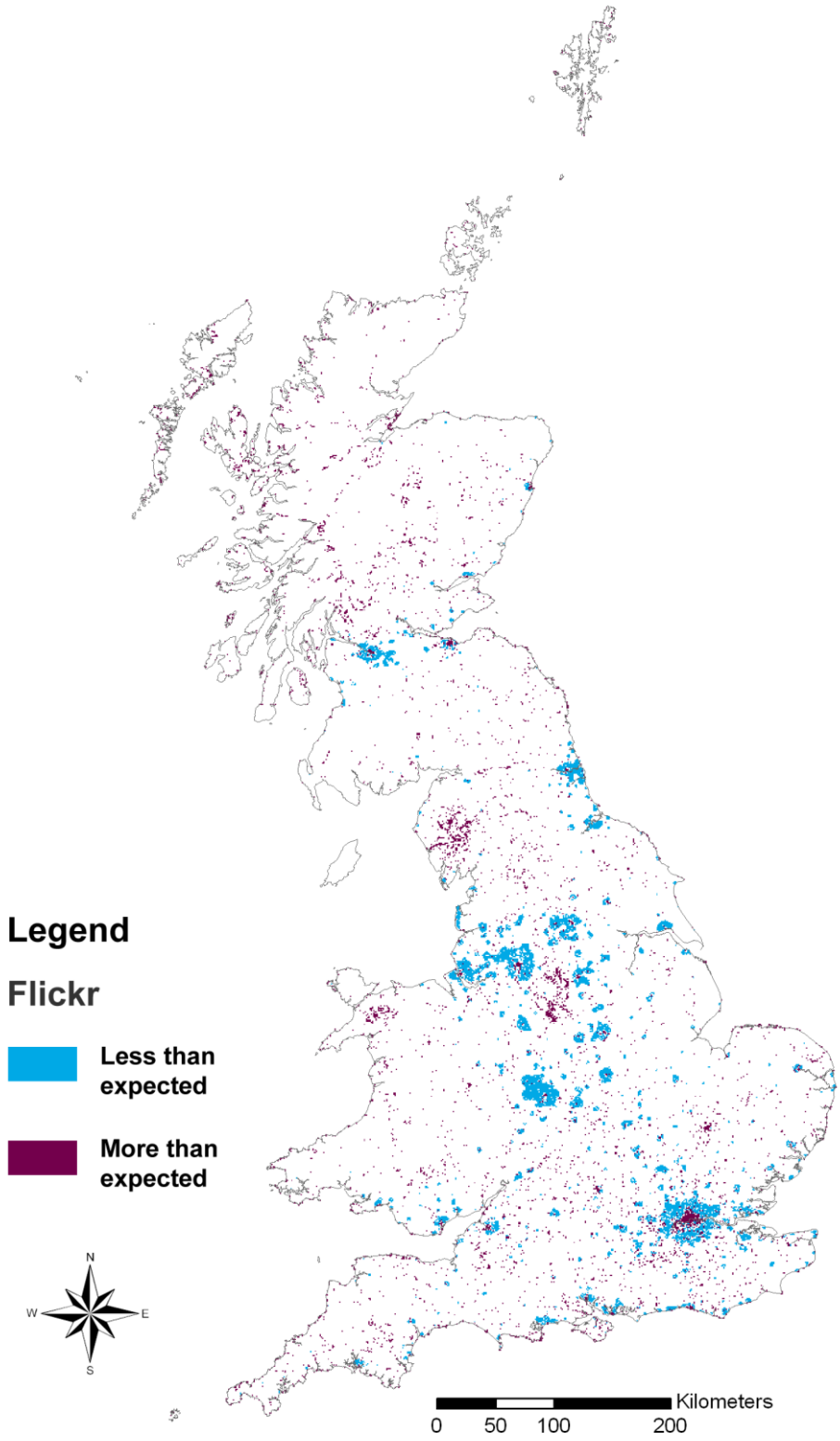
Figure 24. Comparison of the frequencies of the number of photos per tile for Geograph, Flickr and Picasa without taking into account the areas with 0 photos

4.3.4 Expectation surfaces

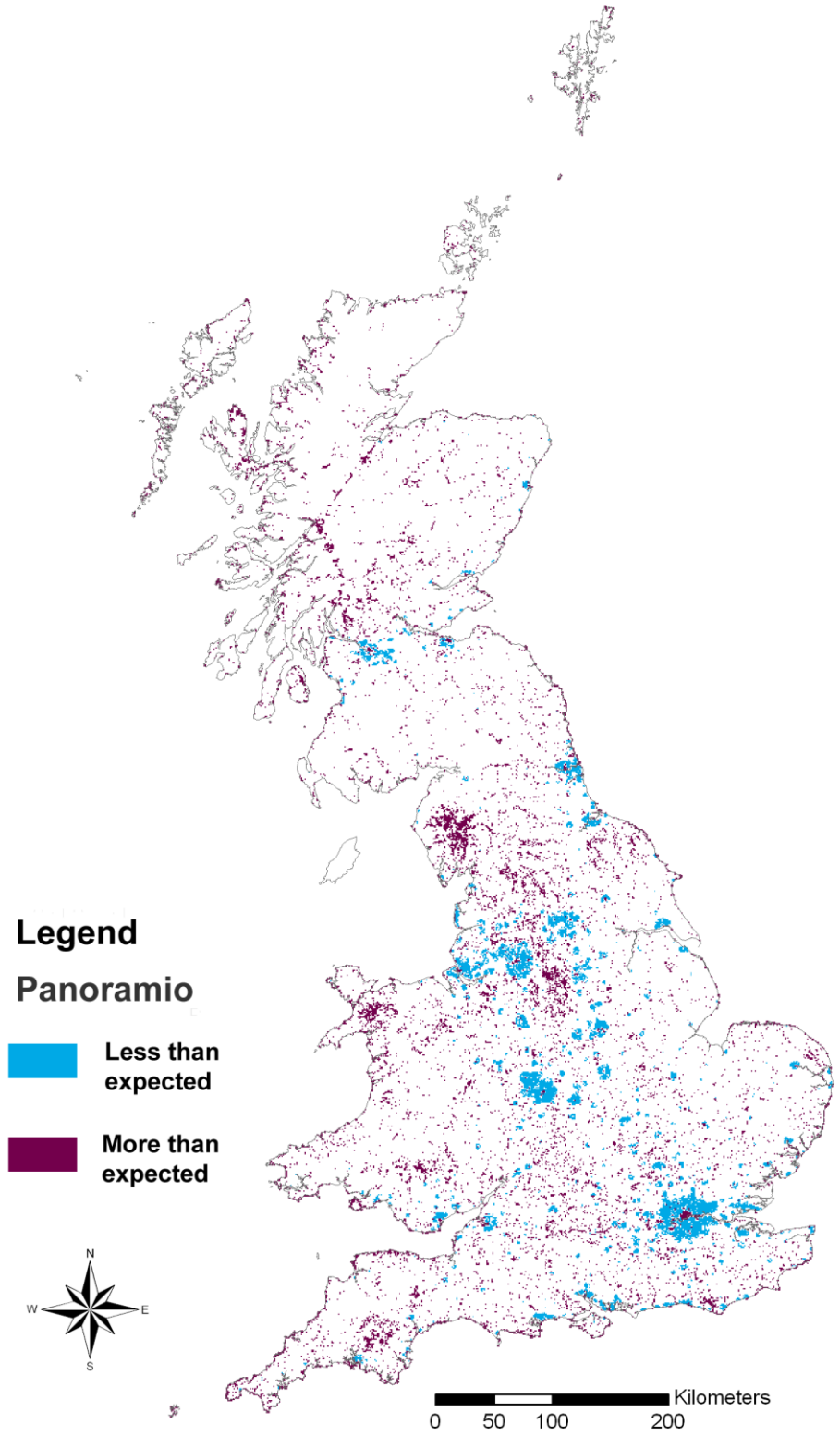
It has been mentioned repeatedly in the course of this Thesis that UGSC apart from a major evolution in the Geomatics domain is also an important social and user-centric phenomenon. In that context, it is important to examine its nature by taking into account factors like the population of the study area. More specifically, as explained in Chapter 3 about the methodology followed, building a surface that both relates to the intensity of a phenomenon and to the location of people facilitates the understanding of the phenomenon. In this case this was accomplished through the calculation of the expectation surfaces using the chi-statistic (Dykes and Wood 2008). For that matter, the number of geo-tagged photos submitted to each source and the 2001 population data have been used. The chi-index will be negative for the tiles where the observed value (i.e. the number of geo-tagged photos submitted) is lower than expected (according to population data) and positive when it is greater than expected. Figure 25 (a to d) shows the expectation surfaces for the four different sources; the purple (dark) shade indicates areas where there are more photos than expected compared to the underlying population in contrast with the cyan (light) areas.


Legend
Geograph
Less than expected
More than expected
N
W
E
S
Kilometers
0
50
100
200

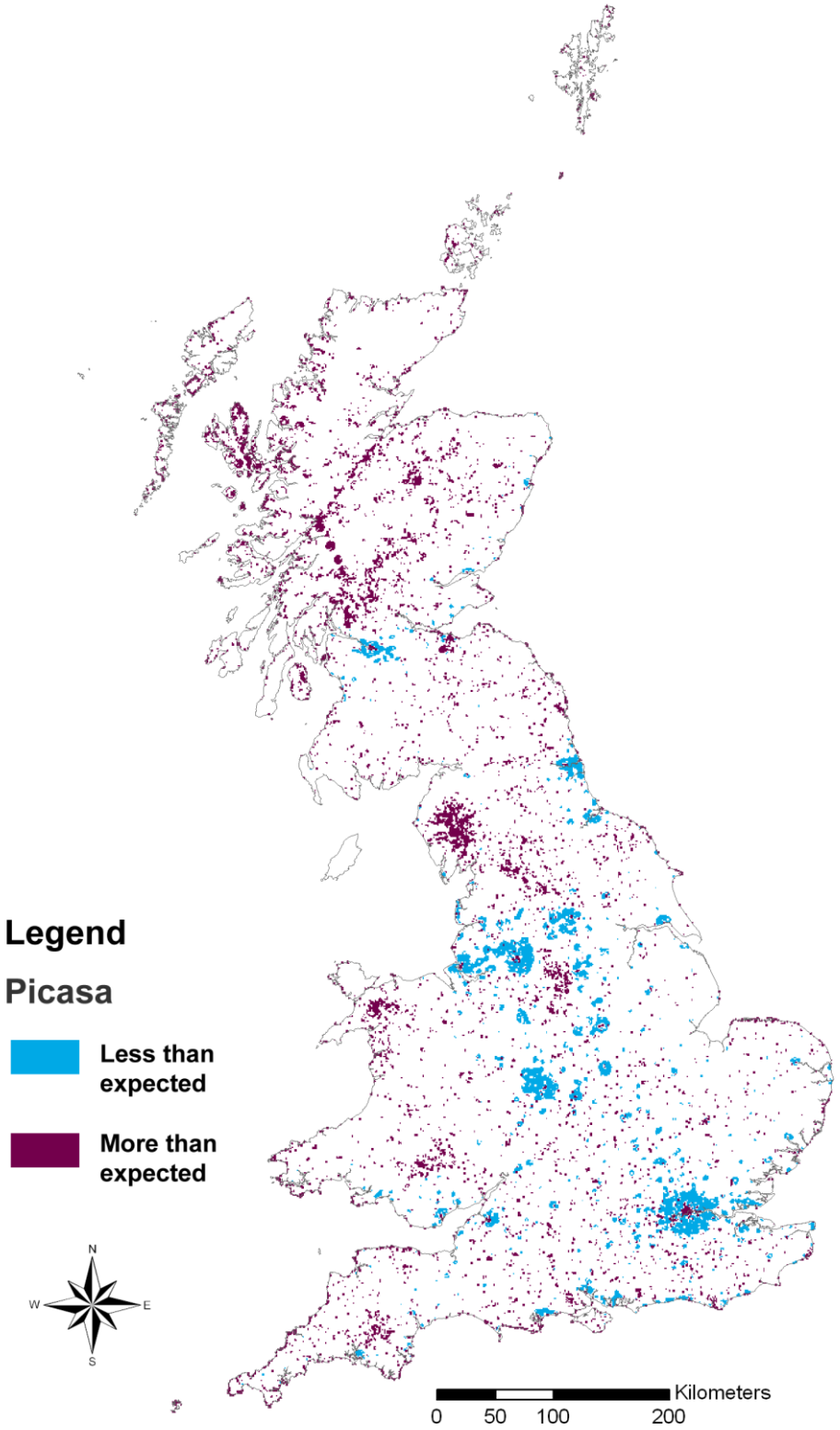
(a)


Legend
Flickr
Less than expected
More than expected
N
W
E
S
Kilometers
0
50
100
200

(b)


Legend
Panoramio
Less than expected
More than expected
N
W
E
S
Kilometers
0
50
100
200

(c)



(d)

Figure 25. Expectation surfaces of geo-tagged photos for (a) Geograph (b) Flickr (c) Panaromio and (d) Picasa Web Albums, based on the population data

In the majority of the study area the number of Geograph's photos is higher than expected compared to the underlying population. In spatial terms, this characteristic can prove valuable for a crowdsourced application because it demonstrates that the population is not the only predictor of data collection and a linear assumption that where there are people, data will be collected is invalid. In contrast, the rest of the sources cover better than expected only the city centres and the tourism areas where huge volumes of data have been submitted but only few people have permanent residence.

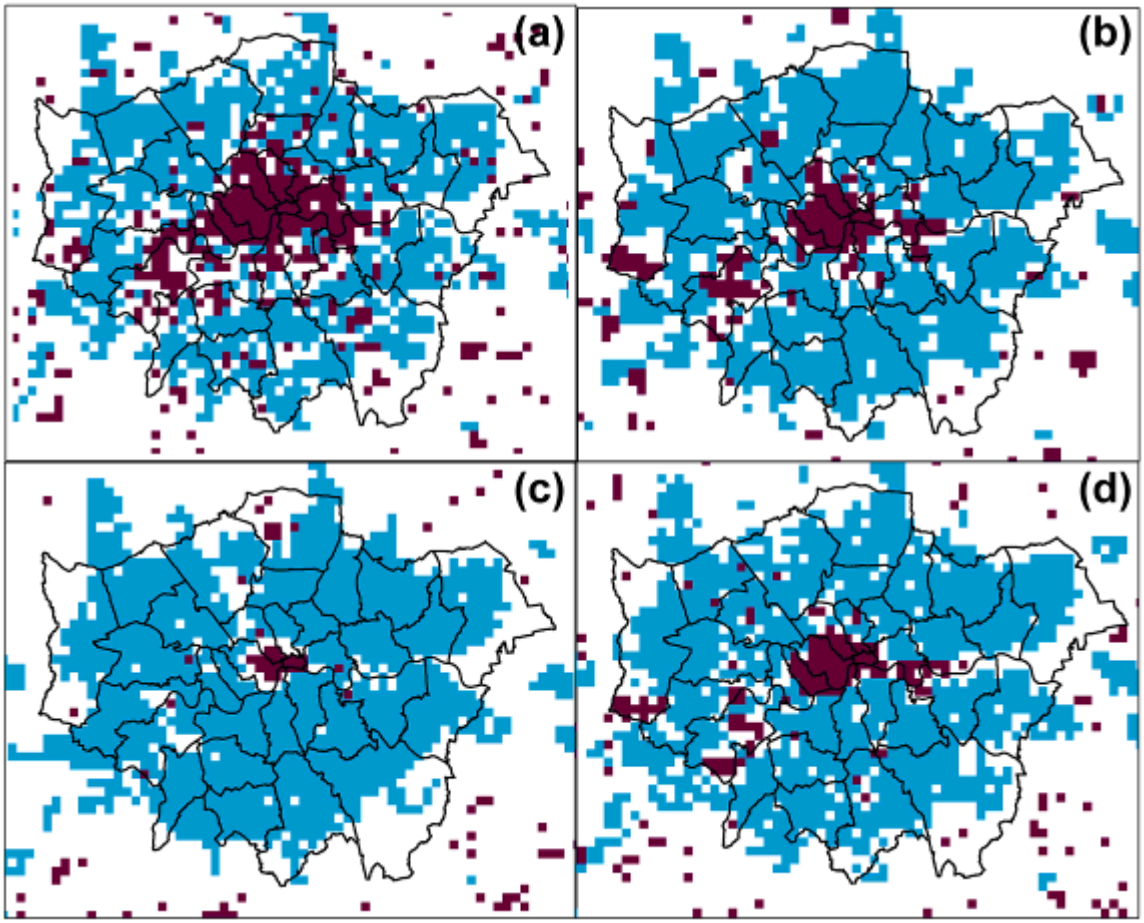


Figure 26. The chi expectation surfaces for the area of Greater London: (a) Flickr, (b) Picasa (c) Geograph (d) Panoramio

Figure 26 shows a magnified view of the chi expectation surface for the Greater London area. As can be seen, even in the urban areas where users' contribution is high for all four Web applications, the better-than-expected performance is confined to a core at the centre of the city, ranging from a very small area, for Geograph, up to a broader one for Flickr. In all four cases though, the outer suburbs of London are underrepresented. In

spatial terms, the characteristic observed this time can prove challenging for mapping updating purposes. It is reasonable to assume that the majority of the man-made changes on the ground are taking place in areas where there is increased human presence. In that context, the limited coverage of the suburbs indicates that there will not be sufficient volume of data to cover these changes.

*** * ***

At this point the first clear conclusion about the nature of the photo-sharing, Web 2.0 applications can be drawn. Although the distributions differ between the sources, it can be suggested that there are two different patterns of distribution. The first one, to which only Geograph belongs, covers almost all the area in scope as its users are not limiting their contributions only to popular areas. Apparently, the aim of Geograph conveyed to its users to collect geo-tagged photos for every square kilometre of Great Britain and Ireland has proven to be a strong motivation. In contrast, the second pattern of distribution (to which Picasa Web, Flickr and Panoramio) covers less percentage of the study area and their pattern is considerably more clustered in urban and tourism areas.

However, both types of applications are based on the fundamental principles that powered the evolution of Web 2.0 and they form a special part of the Web 2.0 world. From a GI retrieval point of view though, it has been shown that such Web sources can be categorised into *spatially implicit* and *spatially explicit* ones and thus the initial hypothesis is corroborated. Spatially explicit applications like Geograph, urge their contributors to interact directly with spatial features (i.e. to capture spatial entities in their photos) while at the same time encourage the photos, and thus the content, to be spatially distributed. In contrast, Flickr and Picasa Web are more socially-oriented, and thus are aimed to allow people to share their photo albums. The support of geotagged photos is one of the many interesting features that these applications have but spatial information is neither the core issue nor the main motivation for their users, and thus they are spatially implicit Web applications. This is in contrast with what takes place for spatially explicit sources where the users are explicitly expected to use geography and location as a motivational and organisational factor.

The evaluation of the latest results made clear that this diversity in the spatial patterns of Web sources is due to the difference in the nature of the Web applications. For example, Panoramio's spatial distribution is closer to the distribution of the spatially implicit sources, in contrast with the initial hypothesis. Its API inconsistencies might play a role but when taking into account the photo frequencies (Figure 16d) it becomes evident that Panoramio behaves like a social-oriented and thus like a spatially implicit application. Thus the fact that Panoramio does not explicitly encourage a consistent and complete spatial coverage of space but rather urges its users to submit photos for places they like or visit makes the application to behave like a spatially implicit one.

4.4 Comparison between spatially explicit (Geograph) and spatially implicit (Flickr) sources

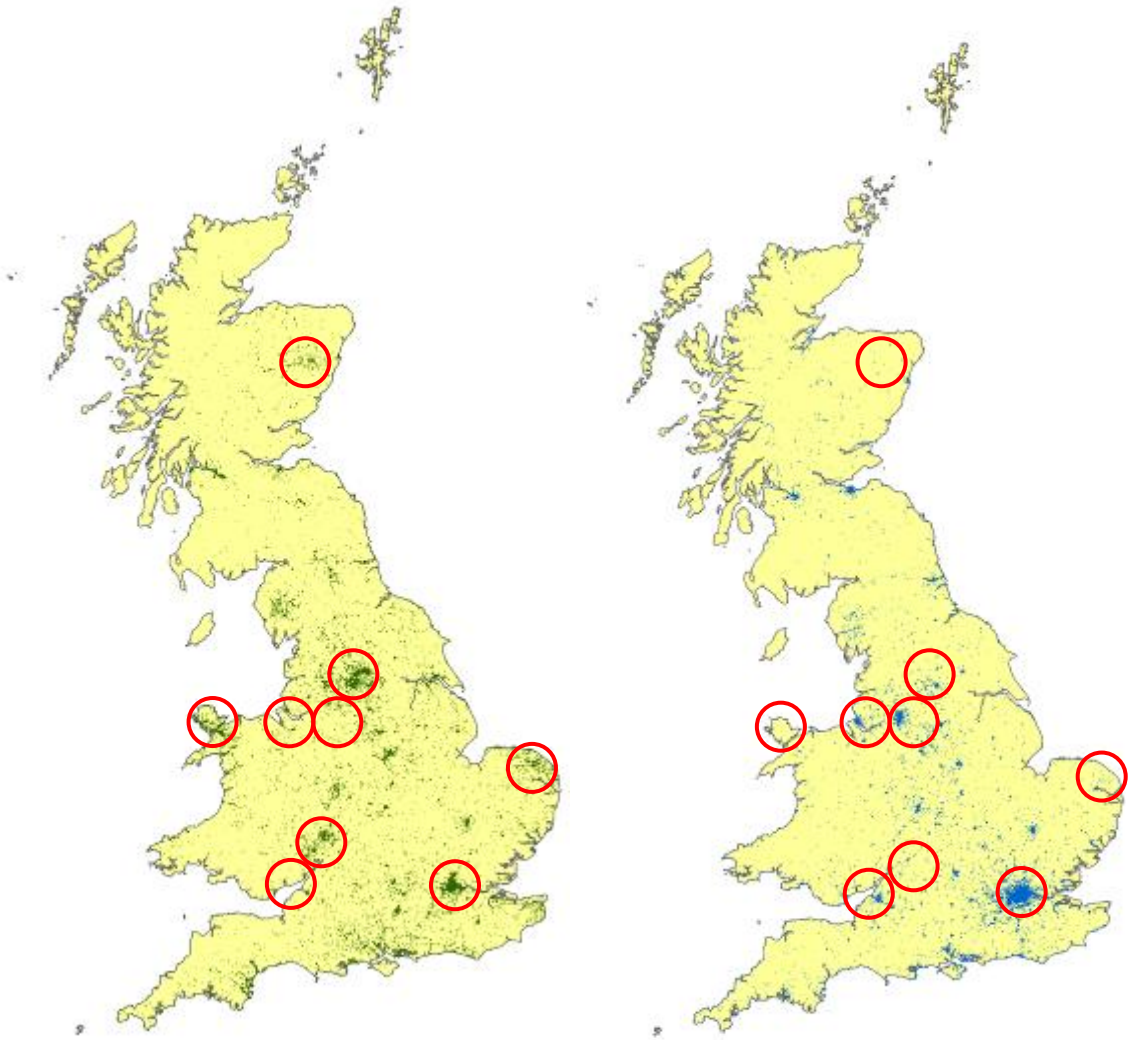
The analysis of UGSC through the examination of photo-sharing Web applications, shows that the spatial explicit applications have the potential to serve as nation-wide sources of spatial content. Yet, the overall data flow of such applications remains an issue as the majority of the area in scope is covered with very few photos. On the other hand, what remains largely unknown for the spatial implicit applications is their ability to serve as sources of UGSC for the more productive to them areas (i.e. the urban and tourism ones). While the formation of a concrete answer to both these question needs further research, a first attempt to explore this issue and lay the ground for the next steps is made here. The methodology followed is based on the monitoring of the tiles that mostly receive the attention of the users (i.e. the popular areas) in both cases. For this analysis, a comparison between Geograph and Flickr, which belong to the spatially explicit and spatially implicit Web applications respectively, will take place.

4.4.1 Data flow in the popular tiles

In an effort to examine the data flow (i.e. the number of photos submitted) to the most popular tiles of Flickr and Geograph, a threshold of 15 photos was set. This number corresponds to an average of one submitted geo-tagged photo per quarter, per tile since

the launch of the Web applications (i.e. the March 2005 when Yahoo! acquired Flickr and Geograph was launched) and for the next 45 months. Thus, for both datasets a tile was characterised as popular if there were 15 or more geo-tagged photos submitted to it. In this category there are 8,889 tiles for Flickr and 12,081 for Geograph, covering just 3.72% and 5.06% of Great Britain respectively. Significantly, although Flickr has 65% more photos than Geograph, its popular tiles cover 26% less area than Geograph's.

Figure 27 shows the spatial distribution of the popular tiles for Geograph (a) and Flickr (b). The important observation here is that the popular areas recorded in the spatially explicit application are not similar to the popular areas of the implicit one. The common popular tiles are 3,533 which correspond approximately to 40% of Flickr's popular tiles and to 30% of Geograph's. In Figure 27 the red circles mark few characteristic differences between the popular areas of the two applications.



(a) **(b)**

Figure 27. The most popular tiles (with 15 or more photos submitted to them) in (a) Geograph and (b) Flickr.

In a second layer, the total number of photo submission to the popular tiles per quarter over a period of 18 months was examined for both Geograph and Flickr; the results are shown in Figure 28.

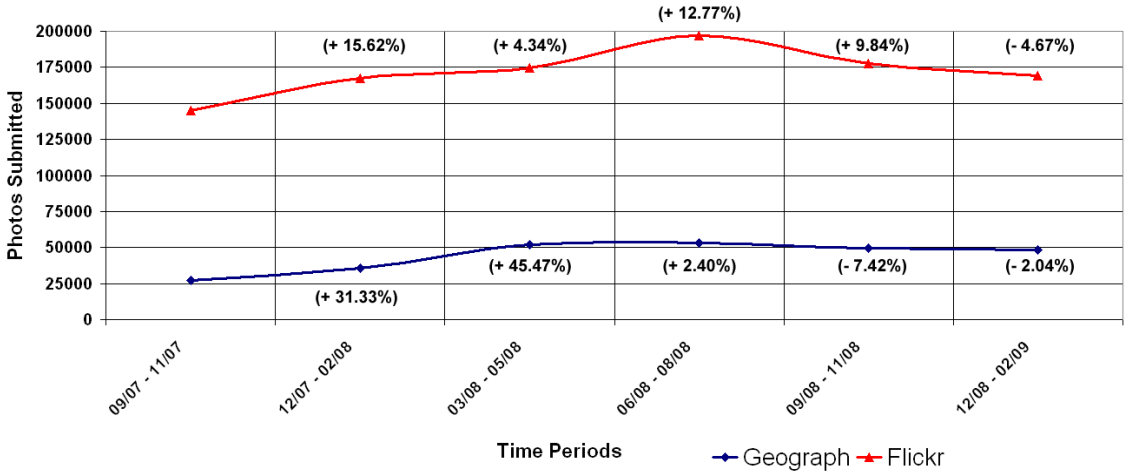


Figure 28. The submission of geotagged photos to Flickr and Geograph over a period of 18 months

The total data flow for both sources follows the same pattern; it shows a strong growth from the start of the test period (09/2007) and for 3 consecutive quarters. For the last two quarters though, there is a negative growth recorded for both sources. This can be attributed to seasonality as the increase in the number of photos takes place during the summer holidays and the decrease during the winter. This seasonal fluctuation is corroborated for the spatially explicit application when monitoring the overall photo submission to Geograph (Figure 29). By taking into account the similarity in the data flow shown in Figure 28, it can be also suggested that the same principle applies to the geo-tagged photos of the spatially implicit applications.

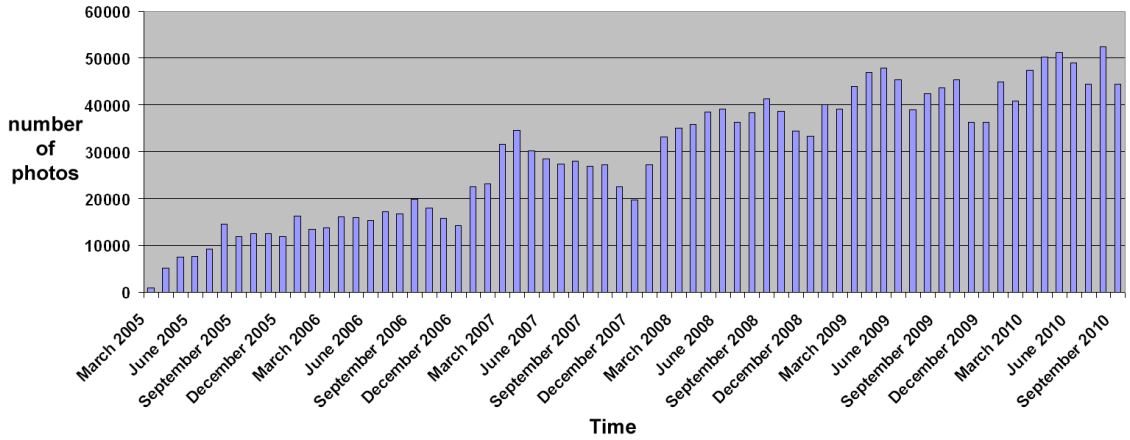


Figure 29. The monthly geo-tagged photo contribution in Geograph

Additionally to the data flow examination, the aim also was to examine the currency of the data available for the popular tiles. When it comes to the use of spatial information (for example in updating mapping products) data currency is paramount. Table 4 shows

the number of popular tiles and the overall percentage of area coverage that they represent for Great Britain, for which at least one geotagged photo has been submitted over a period of three consecutive 6-month periods.

	09/2007 - 02/2008	03/2008 - 08/2008	09/2008 – 02/2009
Geograph	7072 (2.96%)	8927 (3.74%)	7879(3.30%)
Flickr	5884 (2.46%)	6474 (2.57%)	6249 (2.71%)

Table 4. The number of popular tiles for which at least one geotagged photo has been submitted over a period of 6 months. In the parenthesis is the percentage area coverage of Great Britain

The 6-month period was not an arbitrary choice rather it was used to examine the productivity of the phenomenon in accordance with the requirements of OS, the national mapping agency of Great Britain. According to OS (OS 2010b) the aim is to represent in its database some 99.6% of the significant real-world features that are more than 6 months old. From the sub-group of the popular tiles shown in Table 4, only a mere 1.58% (3,782 tiles) of the area of Great Britain had photos submitted to Geograph for all three consecutive 6-month periods. For Flickr the number raises slightly to 1.64% (3,912 tiles). This observation leaves little room for planning on using the photo-sharing Web 2.0 applications as regular sources of spatial content able to serve the needs of a mapping agency. In contrast, the role of such applications could be complementary to the existing mainstream efforts of spatial data collection used up to now.

4.5 Large scale analysis of implicit and explicit sources

So far, the research has shed light on the behaviour of the photo-sharing Web applications at the national level by examining datasets with $1km^2$ spatial resolution. In the next step, the effort focused on the understanding of the applications' behaviour in a larger scale. For that matter, a new dataset was collected for the 15 selected areas shown in Table 5. The sites were randomly selected from the common popular tiles of both Geograph and Flickr and the total area corresponds to the 1/25 ($141km^2$) of them. The only requirement was the test sites to be at least $5Km^2$. For this new dataset there was a more detailed collection of data; for each photo the URL of the actual photo was

recorded, its co-ordinates, its title, its tags and comments, the usernames and the dates of capture and submission.

Num.	Study Area	Num. of Tiles	Num. of Flickr Photos	Num. of Geograph Photos
1	Torquay	10	930	887
2	North London	15	7993	1114
3	Chester	9	3753	1781
4	Leeds	16	3642	1888
5	Dundee	5	1156	215
6	Swindon	6	674	465
7	Oxford	14	11861	704
8	Chatham	12	2266	874
9	Glasgow	6	625	224
10	Edinburgh	9	6837	472
11	East London	9	2337	341
12	West London	9	3444	347
13	Cambridge	6	673	1047
14	Portsmouth	6	2982	1187
15	Nottingham	9	1331	391
	Sums	**141**	**50504**	**11937**
			Total Num. of Users	
			3236	**538**

Table 5. Study Areas for Large Scale Analysis of Flickr and Geograph

The detailed datasets collected for these areas and especially the co-ordinates of where each photo was taken, allowed a kernel density analysis for these areas which examines the spatial distribution of the phenomenon at a large scale and compares the behaviour between explicit and implicit sources[7]. This analysis shows that spatially explicit sources provide better coverage of the study areas, even with fewer photos. For example, Figure 30 shows the density analysis for an area of 15km^2 located in North London. The spatial resolution of the density surface is 10m and the kernel radius is 50m. Figure 30a shows the density surface created from Flickr's photo-capture points (7993 in total) and Figure 30b shows the density surface created from 1109 Geograph points. It is clear that, although Flickr has 6.2 times more photos than Geograph in this area, the spatial distribution of the photo-capture points from Flickr is concentrated in a few relatively popular spots (e.g. Parliament Hill and Dartmouth Path at Hampstead in

[7] but see also the limitations of geo-tagging discussed in Section 3.3; yet, this affects randomly both sources and thus the co-ordinates were assumed to correspond at the camera (photo-capturing) locations.

North London). In contrast, for Geograph, the distribution of the photos is more dispersed covering a considerably larger portion of the area.

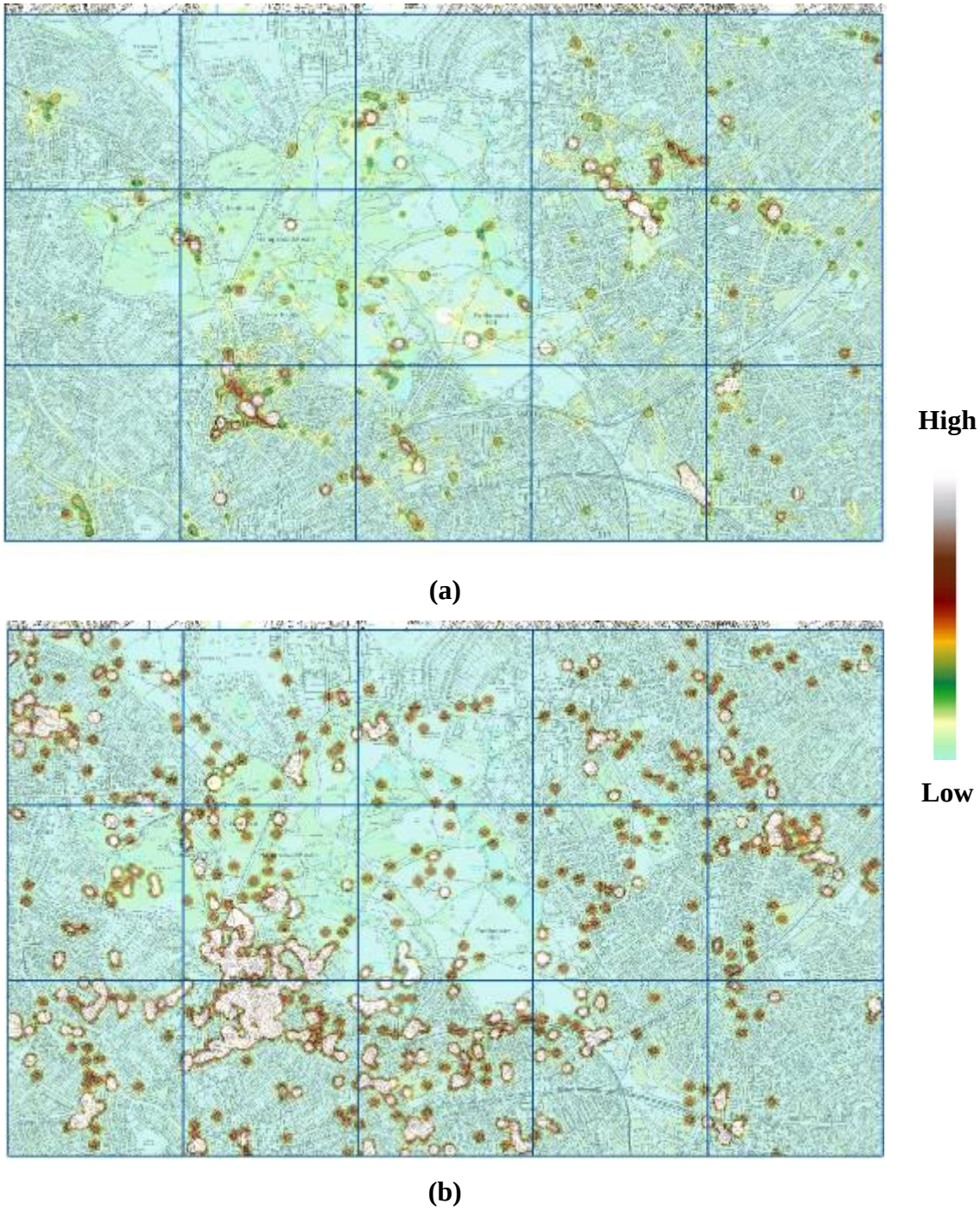


Figure 30. Density surfaces for (a) Flickr and (b) Geograph from the North London test area

In an effort to quantify this observation in more practical terms, an event analysis for each study area was carried out. This allowed the quantification of the repetition

observed in the photo-capture locations (i.e. camera locations) for each source. Figure 31 shows the percentage of different photo-capture locations for each study area.

The average percentage of unique locations for Flickr is 30.1%, in contrast with 85.6% for Geograph. This means that, on average, 100 photos in Flickr have be taken from only 30 different camera locations in contrast with 85 different locations in Geograph.

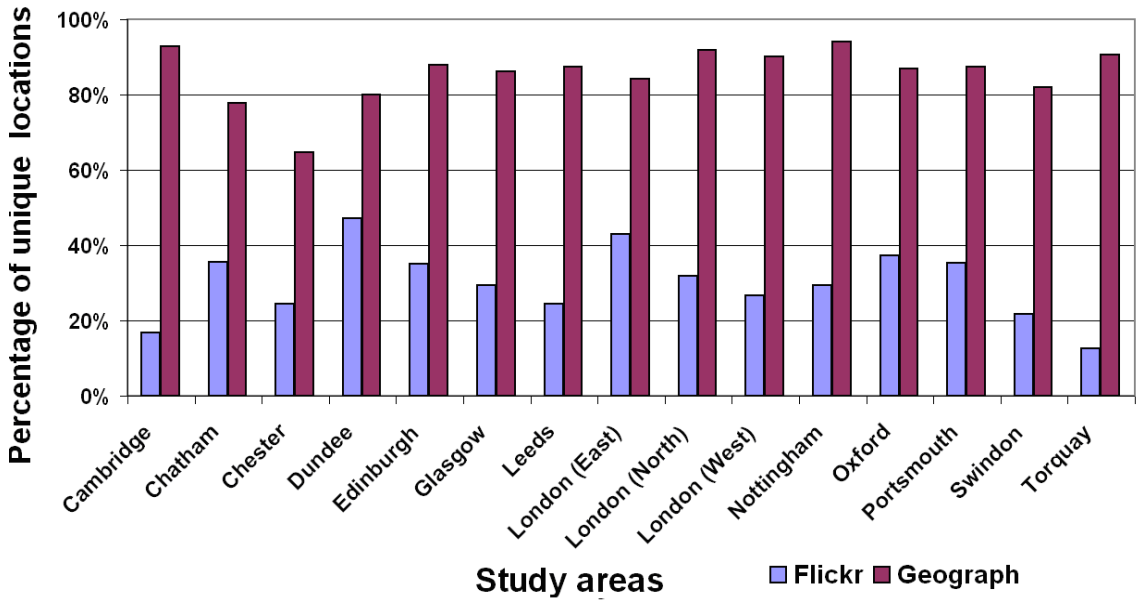


Figure 31. Percentage of unique camera location for each study area.

4.6 User behaviour analysis

In this section, the research conducted regarding the social aspects of the phenomenon in the Section 4.2.4, which examined the photo contribution versus the underlying population, is complemented by looking into the users' behaviour. The focus on this user behaviour analysis is on the users' contribution patterns, as measured by the time difference between capturing and uploading a photo, and also with regard to the time that users remain active in an area (i.e. time difference between first and last photo submission for each user). This will reveal the currency of the photos submitted to such applications and provide insights into how users' participation is evolving through time.

Starting with the analysis of the data from Table 5 (presented earlier in Section 4.5), it can be seen that Geograph users appear to be more productive than those from Flickr. There are 3,236 Flickr users that have uploaded photos for these 15 study areas in contrast to 538 Geograph users, which means that individual contribution for Flickr users is 15.6 photos per user in contrast to 22.2 photos per user for Geograph.

However, apart from this element, users' behaviour is quite similar for both sources. Figure 32 shows the percentage of photos submitted in various time spans (the negative values are due to falsely recorded timestamps by either the users or the Web sources). The results presented make it clear that in both cases the users upload their photos in the first couple of weeks after the capture date. Very few photos (8.4% for Flickr and 9.2% for Geograph) are more than a year old which indicates that such sources can be used in applications that need contemporary photos.

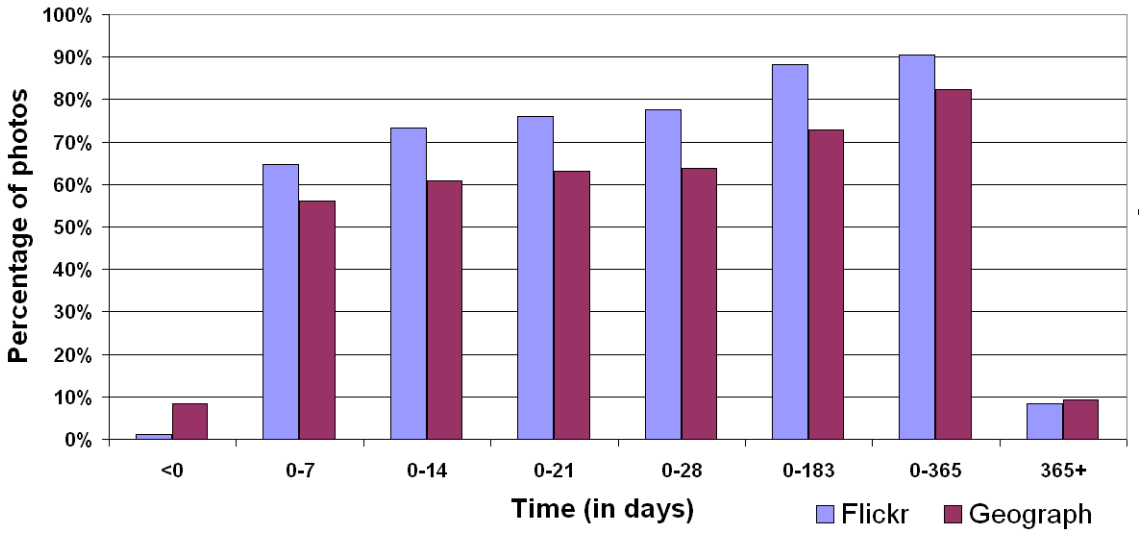


Figure 32. Time difference between capturing and submitting a photo to Flickr and Geograph

Moreover, Figure 33 shows how long users remained active in these test areas. For both Web applications, more than 50% of the users captured data over a period of less than a day and thus it can be suggested that these users were just visiting the area. On the other hand, a rough estimation can be made about the users that are permanent residents in an area (for example, by calculating the users that have stayed active for more than 14 days). This is important because local knowledge has been widely acknowledged as a very important factor in retrieving geographic information (see for example: Haklay and

Tobon 2003, Dunn 2007, Budhathoki et. al 2008, Elwood 2008a and the relative discussion in Section 2.2.2.5).

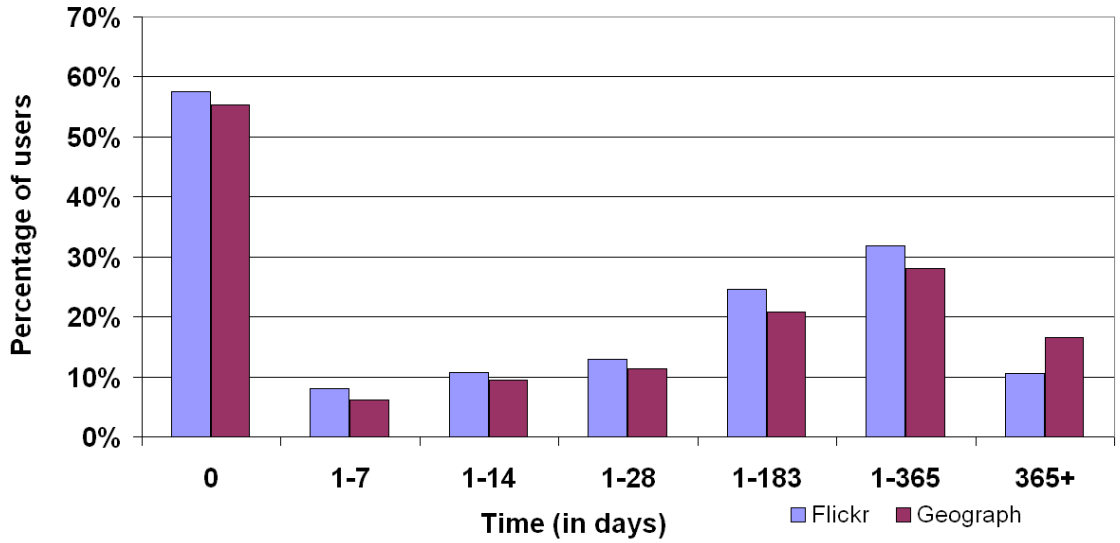


Figure 33. Periods of user activity

The final step in this comparison process between the user's behaviour of spatially explicit and implicit applications, focuses on the overall pattern of user contribution. As it will be further explained in the next Section, a new full-scale dataset (i.e. not only the popular tiles) of Great Britain was collected from Flickr six months after the first one. The total number of geo-tagged photos in the new dataset is approximately 2 millions (2,085,897) and has been created by the submissions of 32,467 users. The new Flickr dataset was compared against statistical data collected for Geograph where a similar number of geo-tagged photos (1,998,596) have been submitted by a total of 10,031 users.

The first clear observation is that Geograph managed to collect almost the same number of geo-tagged photos with 1/3 of the number of users that Flickr did. Furthermore, the joint analysis of those two parameters (i.e. photos submitted and users) for both datasets is shown in Figure 34. In fact, what this Figure shows is the accumulated volume of photos submitted versus the accumulated percentage of contributing users for both types of applications. The X-axis shows the number of photos submitted. So, for example the reader can see that 500 or more geo-tagged photos have been submitted by the 4% of Geograph's and the 2.1% of Flickr's users. In turn, these user participation percentages

have created the 87.7% of the Geograph's overall data repository and the 53.5% of Flickr's. Thus, it is interesting to note that both sources are behaving more 'extremely' than described by the 80-20 Pareto's rule, which finds many applications in social phenomena (simply described the rule dictates that the 80% of the effects come from the 20% of the causes). Importantly, in the spatially explicit application of Geograph this rule is transformed into 95-5 whereas for the spatially implicit application of Flickr the analogy is 95-10. From the point of view of a mapping agency, this asymmetry might prove particularly valuable. Tapping the productivity of a relatively small group of users (for example, the 5% of Geograph users corresponds to just 500 users for the entire Great Britain) through a specifically designed series of incentives and motivations would produce the same effect as the entire Web 2.0 application. Interestingly, this is in direct contrast with the principle of Long Tail discussed in Section 1.2.3 as the volume of data that resides in the long tail of the phenomenon does not seem to have the value expected.

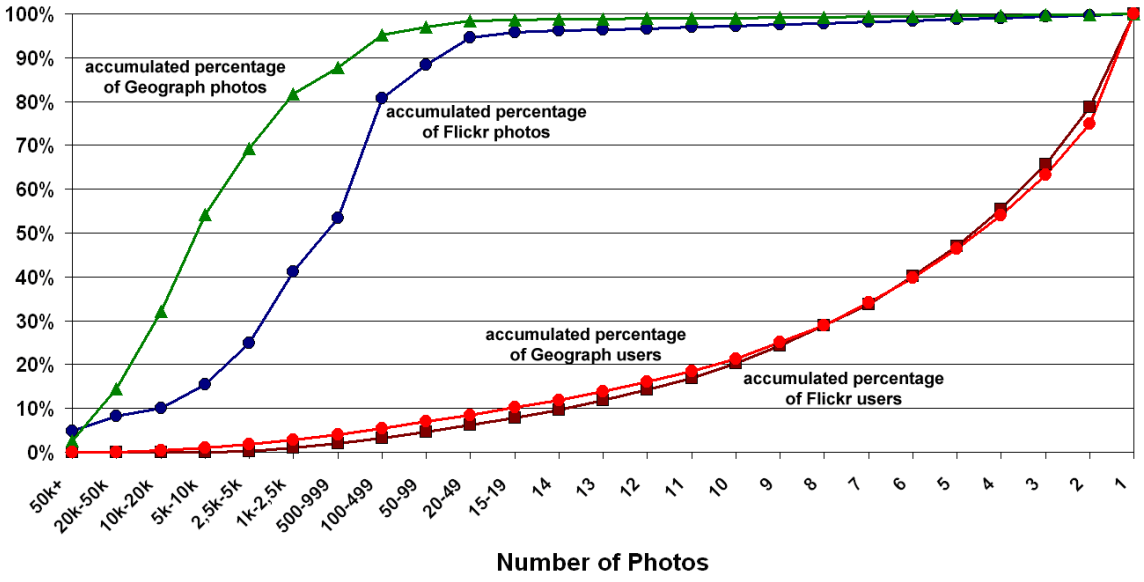


Figure 34. Accumulated percentages of photos submitted to Geograph and Flickr versus the accumulated number of contributing users.

4.7 Spatially implicit application's data flow monitoring (Flickr)[8]

As mentioned in the previous Section, a new dataset of the Flickr's geo-tagged photos for Great Britain was collected six months after the first one. This was deemed necessary as at this stage of analysis the focused turned to the examination of the evolution of the spatially implicit Web application over time. Once again, a 6-month period was selected to match the aims of the OS. The dataset from the new collection was compared to the one initially collected. Figure 35 shows the differences in the number of photos between the two datasets. It is clear that the pattern presented is similar to the one presented in Figure 18. This means that the majority of the activity took place at the same area (urban and tourism). A second observation is that there are both additions and deletions of data (although the additions are considerably more). More specifically, there were photos submitted to the 9.2% of the total area where as the deletion of photos occurred in the 1.8% of Great Britain.

Figure 36 shows the areas where new tiles have been populated with geo-tagged photos. The total number of tiles in this category is 8554 which corresponds to the 3.6% of the total research area. Importantly, from those new tiles only 1% is populated with more than 3 photos. These observations further enforce the option of using such Web applications as complimentary sources of spatial content (at least in a national level) rather giving them the role of universal sources of spatial content that can support the mainstream cartographic production.

[8] This Section is adapted from:

Antoniou, V., Morley, J., and Haklay, M., 2009c, *Do photo sharing websites represent a sufficient database to aid in national map updating or change detection?* Presented at EuroSDR workshop on Crowd sourcing for updating national databases, Wabern, Switzerland, 20-21 August.

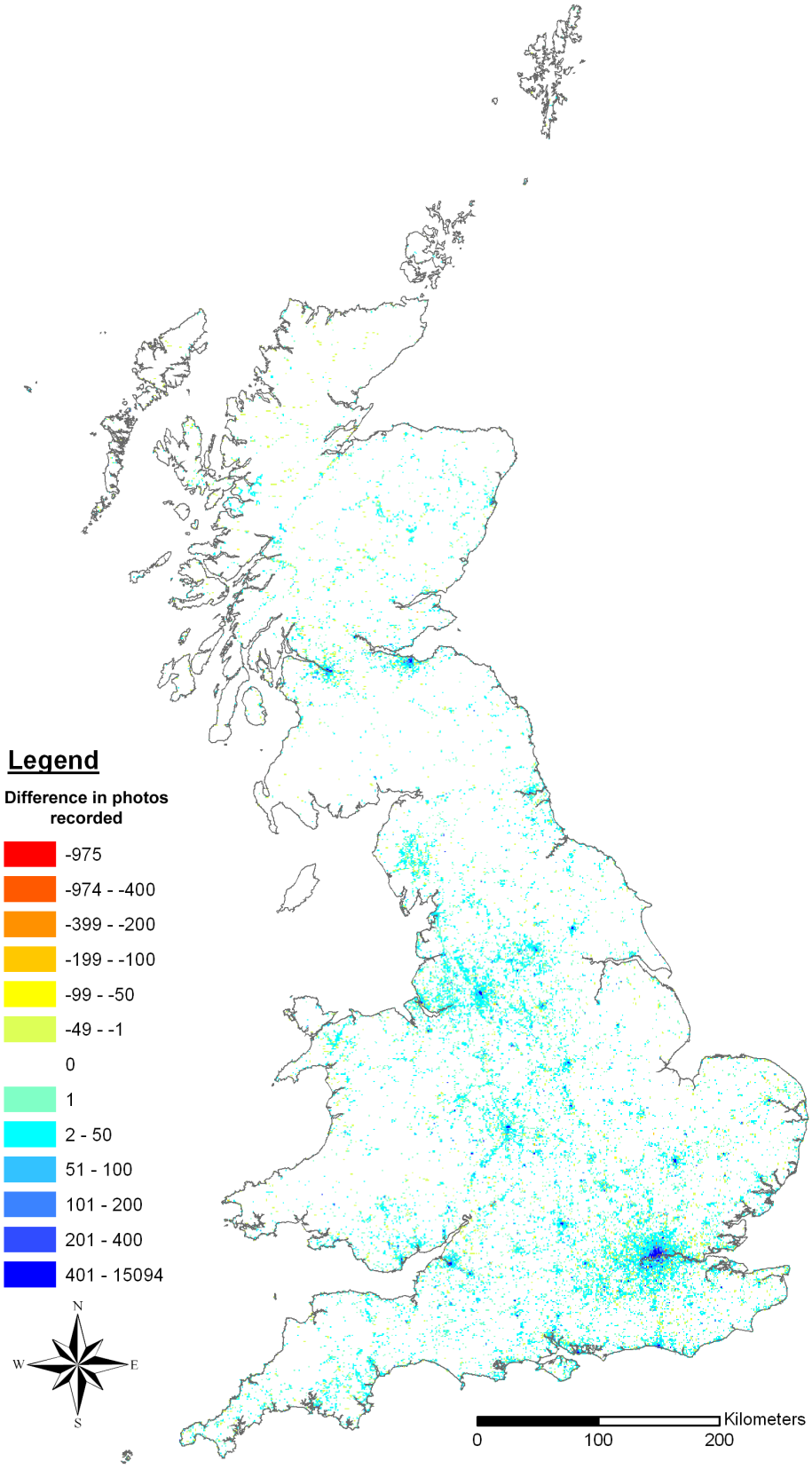


Figure 35. The changes recorded over a period of 6 months for Flickr

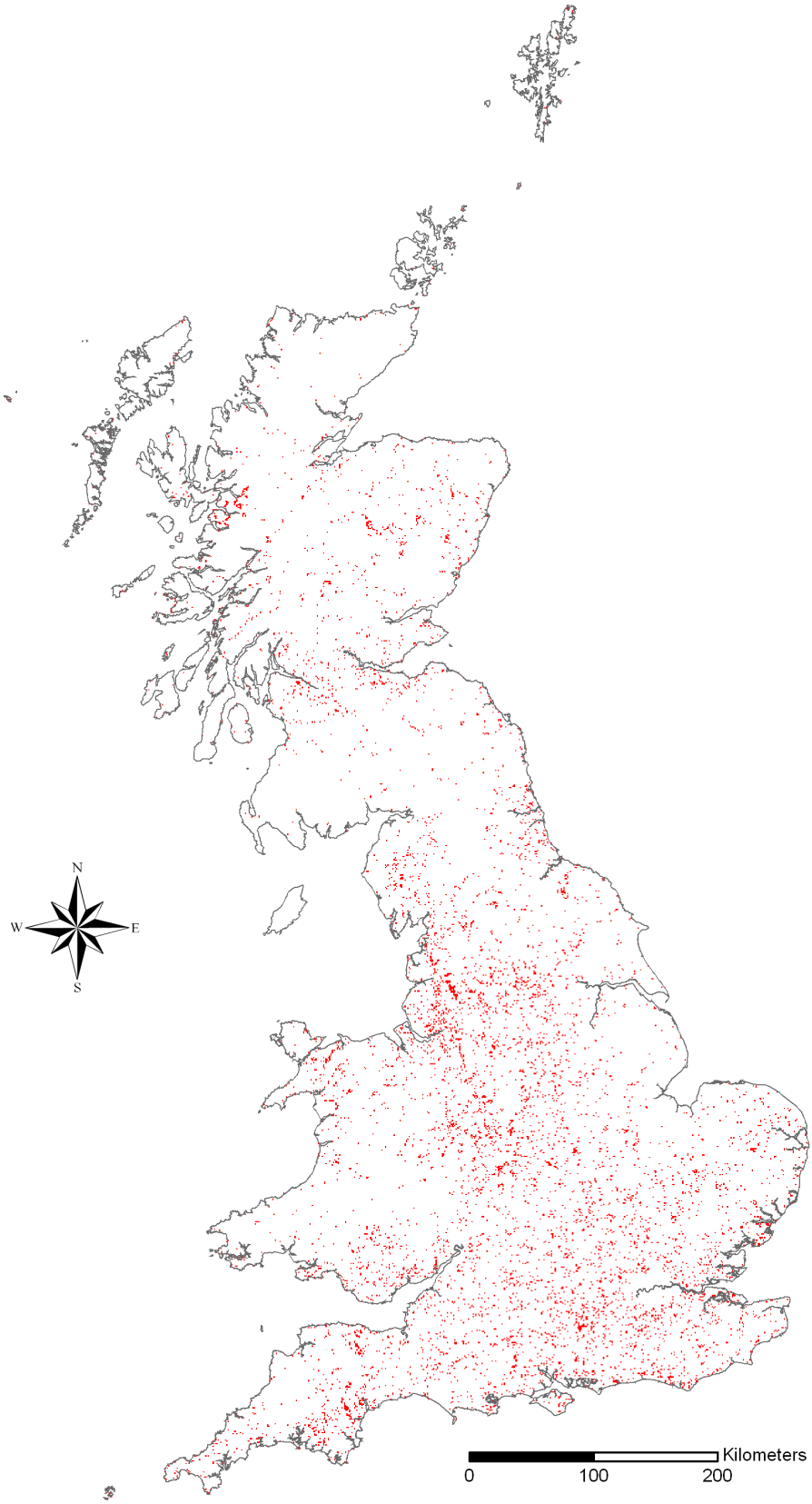
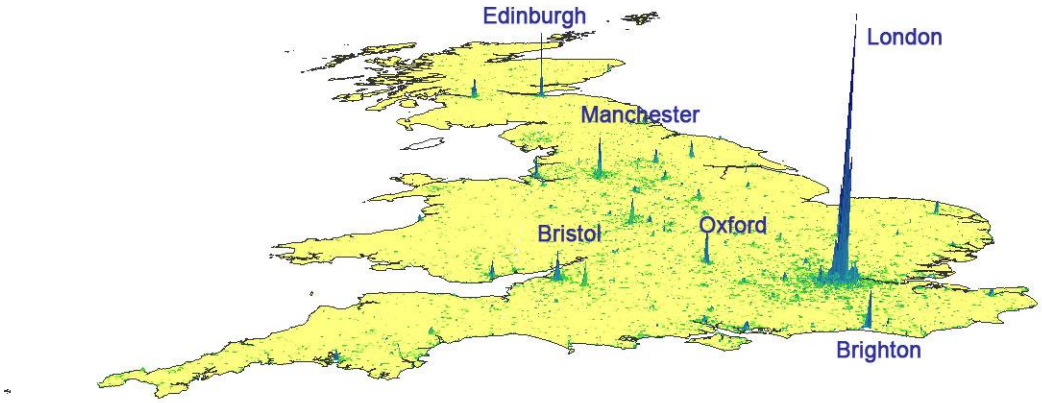
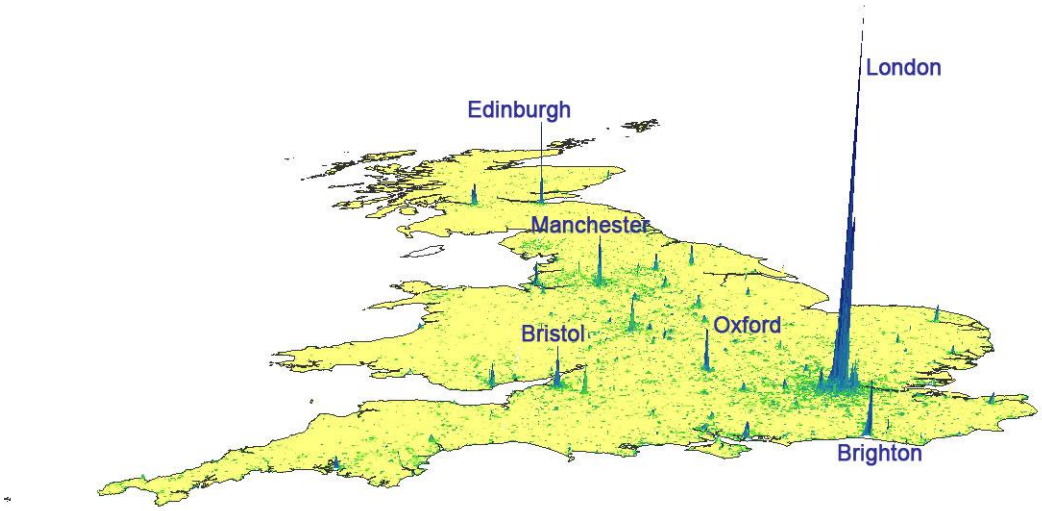


Figure 36. The areas where new geo-tagged photos have been submitted to Flickr over a period of 6 months

Finally, despite the changes that took place during the 6-month period, the overall pattern of the magnitude of the phenomenon did not change in terms of spatial distribution (Figure 37a and 37b).



(a)



(b)

Figure 37. 3D representations of Flickr's data distributions in a 6 months period

4.8 Summary

The analysis of the photo-sharing, Web 2.0 applications presented covers a considerable part of the spatial content that is generated today by lay users on the Web. The focus was on understanding the fundamental behaviour of the phenomenon and on examining its main characteristics. Additionally, given the rise and increased popularity of such social-networking Web 2.0 applications, the interest also turned to the evaluation of the phenomenon's evolution in terms of user's participation and spatial content productivity.

In this context, the analysis systematically covered many of the phenomenon's aspects in an effort to realise the potentials of the spatial data contributed. The analysis was made under the prism of a possible mapping agency's engagement with UGSC. The analysis conducted set the basis to gain helpful insights and enabled the building of valuable knowledge. Both will be further discussed in Chapter 7. In the next Chapter, the interest will shift to the second big family of UGSC on the Web: vector data.

Chapter 5

Results of the vector data analysis

5. Results of the Vector Data Analysis

5.1 General

> *"OpenStreetMap creates and provides free geographic data such as street maps to anyone who wants them. The project was started because most maps you think of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive, or unexpected ways"* (OSM 2010b).

As noted in the OSM welcoming statement, the OSM project started as a reaction to the limitations posed by mapping agencies on the online geospatial content. To overcome these limitations, the OSM project managed to set up a constantly growing community by mobilising more than 300,000 contributors in an effort to provide a free map of the world: "*Mission: To map the world and give the data away for free*" (OSM 2010c).

The fundamental concept behind OSM is not much different from the one in Wikipedia. OSM provides a Web platform that enables users to freely upload or create spatial content or modify existing spatial content submitted by other users. Spatial data upload usually takes place when the users want to contribute to OSM their GPS-recorded data where as data creation or editing includes mainly on-screen digitising. These content generation processes are accomplished with the help of various applications that have been developed by the OSM users. Figure 38 shows the usage share of the most popular OSM editors based on the analysis of the data collected for OSM.

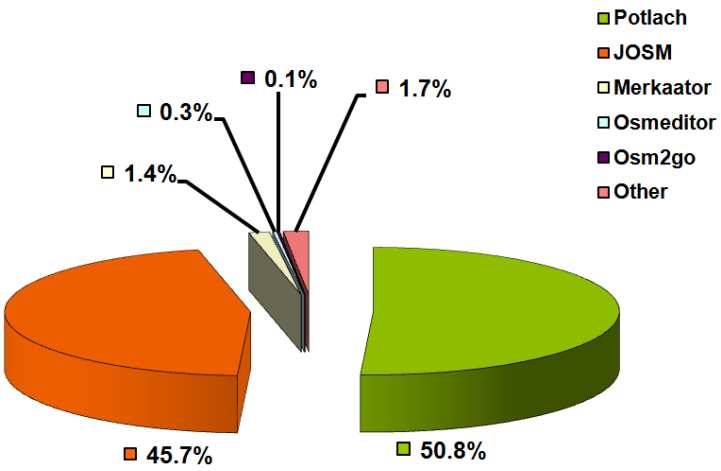


Figure 38. The OSM editors' usage share

Apart from the similarities in the fundamental concepts between OSM and Wikipedia, there are also similarities in the challenges that these two projects face. Perhaps the most important one is the concern about their quality. It is a fact that the OSM community has realised the importance of the quality. A variety of efforts have started to emerge aiming to provide solutions for improving OSM quality. For example, some of the OSM editors have already started to guide the users into entering correct tag values (i.e. values that are in compliance with the list of values that the OSM community has agreed upon). Another example is a group of independent efforts (initiated by OSM users) that focus on discovering errors and then prompting the OSM community to correct them (see also Section 3.4.1.4).

In any case, the OSM community has realised that the quality of spatial dataset is a very important issue. However, an interesting yet elusive issue is first the quantification of the OSM quality and then its communication to any interested party. Regarding the quality quantification, as seen in Section 2.4.3, there have been already some empirical studies towards this direction. These efforts range from the evaluation of a few test sites up to a national-level evaluation for selected spatial data quality element (i.e. positional accuracy and completeness respectively). Regarding the issue of communicating quality, though, no particular effort has been made. As explained, this Thesis complements the findings of the existing research by examining the positional accuracy and the attributes quality in a national level. Moreover, in an effort to address the quality

sharing challenge the Thesis implements the ISO suggested methodology for conducting the evaluation tests.

5.2 Chapter's overview

In the statement quoted at the beginning of this Chapter it seems that there is an uncritical mix of the terms '*geographic data*' and '*maps*'. However, in this Thesis the focus is only on the examination and evaluation of the geographic data provided by OSM and not on the mapping products such as maps, as the former is the actual UGSC and the latter might contain errors introduced by the cartographic process.

The research started with a general, preliminary examination of the OSM data (Section 5.3). By examining perhaps the most successful Web application in its kind, this first step helped to understand the nature of the data, the evolution of the project and, most importantly, how the UGSC phenomenon regarding vector data is realised on the Web.

After the initial analysis of the OSM datasets it became clear that there are two important issues that require further investigation. First is to examine and analyse the positional accuracy of OSM (Section 5.4). Second is to examine the attributes' quality (Sections 5.5 and 5.6).

In the first stage of the research Cloudmade shapefiles were used. However, it was realised that the data did not contain the original OSM_ID of each spatial entity, in contrast with the Geofabrik shapefiles. Thus, in the next step of the research (i.e. the positional accuracy evaluation) the Geofabrik shapefiles were used. The same datasets were used also in the final leg of the research (i.e. tag evaluation) for the gathering of the raw OSM tags for Great Britain.

5.3 Preliminary OSM analysis

The preliminary OSM analysis took place so as to realise the nature of the OSM datasets and particularly the changes/updates that took place to these datasets over time. More specifically, the England OSM Highways (i.e. the road network) and POIs datasets were downloaded from Cloudmade in January 2009, April 2009 and July 2009. The aim at this phase was to perform a like to like comparison of the datasets to understand how the phenomenon evolves. Through this process, the changes in Highways and POIs that took place over two consecutive quarters were examined. Regarding the Highways, both the geometric changes among the datasets and the completion of the roads' name attribute was monitored. Similarly, for POIs the monitoring focused on the change, deletion and alteration of the spatial entities. This was achieved by examining the geometry, the type and the name tags of each entity as described in Section 5.3.2.

5.3.1 Highways monitoring

The comparison of the three different datasets took place for 22 of the OSM Highway categories. In Table 6 the results of the comparisons are presented. The Table includes the number of features recorded per OSM category for each one of the three datasets (Dataset Jan09 has been downloaded in January 2009, Apr09 in April 2009 and Jul09 in July 2009) and the share of this category in the total OSM Highways population (i.e. *Layer Features (% of Total)* in the table header). Apr09 and Jul09 datasets are further divided into two sub-categories: the *Identical Features* and the *New Features* (compared with the previous dataset). The percentage in the parenthesis refers to the share of this sub-category inside the category. Furthermore, each sub-category is divided into two groups: the *With Name (WN)* for those Highway entities that have a name tag and *No Name (NN)* for those that have not. Once again, the percentage in the parenthesis gives the share of each group inside the sub-category.

	Dataset Jan09		Dataset Apr09				Dataset Jul09			
	Layer Features (% of Total)		Layer Features (% of Total)				Layer Features (% of Total)			
	-	-	Identical Features (% of Layer Features)		New Features (% of Layer Features)		Identical Features (% of Layer Features)		New Features (% of Layer Features)	
	WN (%)	NN (%)	WN (% IF)	NN (% IF)	WN (% NF)	NN (% NF)	WN (% IF)	NN (% IF)	WN (% NF)	NN (% NF)
Bridleway	6818 (0.9%)		8533 (1%)				10326 (1%)			
	-	-	5988 (70.2%)		2545 (29.8%)		7510 (72.7%)		2816 (27.3%)	
	899 (13.2%)	5919 (86.8%)	710 (11.9%)	5278 (88.1%)	289 (11.4%)	2256 (88.6%)	830 (11.1%)	6680 (88.9%)	351 (12.5%)	2465 (87.5%)
Cycleway	11930 (1.7%)		14602 (1.7%)				17727 (1.8%)			
	-	-	10017 (68.6%)		4585 (31.4%)		12744 (71.9%)		4983 (28.1%)	
	1524 (12.8%)	10406 (87.2%)	1249 (12.5%)	8768 (87.5%)	533 (11.6%)	4052 (88.4%)	1508 (11.8%)	11236 (88.2%)	641 (12.9%)	4342 (87.1%)
Footway	90289 (12.5%)		118848 (13.6%)				143204 (14.5%)			
	-	-	83172 (70%)		35676 (30%)		110829 (77.4%)		32375 (22.6%)	
	7531 (8.3%)	82758 (91.7%)	6594 (7.9%)	76578 (92.1%)	2382 (6.7%)	33294 (93.3%)	8371 (7.6%)	102458 (92.4%)	2602 (8%)	29773 (92%)
Primary	30903 (4.3%)		33440 (3.8%)				35768 (3.6%)			
	-	-	23011 (68.8%)		10429 (31.2%)		27794 (77.7%)		7974 (22.3%)	
	13121 (42.5%)	17782 (57.5%)	10177 (44.2%)	12834 (55.8%)	5318 (51%)	5111 (49%)	12974 (46.7%)	14820 (53.3%)	4272 (53.6%)	3702 (46.4%)
Residential	303066 (42.1%)		366718 (41.9%)				407543 (41.2%)			
	-	-	277463 (75.7%)		89255 (24.3%)		348526 (85.5%)		59017 (14.5%)	
	227508 (75.1%)	75558 (24.9%)	216004 (77.8%)	61459 (22.2%)	70706 (79.2%)	18549 (20.8%)	278638 (79.9%)	69888 (20.1%)	49484 (83.8%)	9533 (16.2%)
Secondary	23186 (3.2%)		25916 (3%)				27709 (2.8%)			
	-	-	17198 (66.4%)		8718 (33.6%)		20836 (75.2%)		6873 (24.8%)	
	11068 (47.7%)	12118 (52.3%)	8617 (50.1%)	8581 (49.9%)	4632 (53.1%)	4086 (46.9%)	10951 (52.6%)	9885 (47.4%)	3559 (51.8%)	3314 (48.2%)
Service	50967 (7.1%)		66512 (7.6%)				78751 (8%)			
	-	-	47453 (71.3%)		19059 (28.7%)		63656 (80.8%)		15095 (19.2%)	
	6533 (12.8%)	44434 (87.2%)	6075 (12.8%)	41378 (87.2%)	1919 (10.1%)	17140 (89.9%)	7638 (12%)	56018 (88%)	1489 (9.9%)	13606 (90.1%)
Tertiary	31566 (4.4%)		39403 (4.5%)				43504 (4.4%)			
	-	-	24024 (61%)		15379 (39%)		33136 (76.2%)		10368 (23.8%)	

	18014 (57.1%)	13552 (42.9%)	13918 (57.9%)	10106 (42.1%)	6089 (39.6%)	9290 (60.4%)	19991 (60.3%)	13145 (39.7%)	6387 (61.6%)	3981 (38.4%)
Track	9290 (1.3%)		13242 (1.5%)				18591 (1.9%)			
	-	-	8332 (62.9%)		4910 (37.1%)		12012 (64.6%)		6579 (35.4%)	
	1016 (10.9%)	8274 (89.1%)	849 (10.2%)	7483 (89.8%)	571 (11.6%)	4339 (88.4%)	1298 (10.8%)	10714 (89.2%)	560 (8.5%)	6019 (91.5%)
Trunk	20089 (2.8%)		22853 (2.6%)				29862 (3.0%)			
	-	-	15066 (65.9%)		7787 (34.1%)		24487 (82%)		5375 (18%)	
	5725 (28.5%)	14364 (71.5%)	4433 (29.4%)	10633 (70.6%)	2561 (32.9%)	5226 (67.1%)	7831 (41%)	16656 (87.1%)	1826 (34%)	3549 (66%)
Unclassified	113202 (15.7%)		129694 (14.8%)				138582 (14%)			
	-	-	93947 (72.4%)		35747 (27.6%)		114009 (82.3%)		24573 (17.7%)	
	47915 (42.3%)	65287 (57.7%)	40933 (43.6%)	53014 (56.4%)	14538 (40.7%)	21209 (59.3%)	50079 (43.9%)	63930 (56.1%)	10875 (44.3%)	13698 (55.7%)
Byway	679 (0.1%)		952 (0.1%)				1198 (0.1%)			
	-	-	562 (59%)		390 (41%)		814 (67.9%)		384 (32.1%)	
	548 (80.7%)	131 (19.3%)	100 (17.8%)	462 (82.2%)	99 (25.4%)	291 (74.6%)	165 (20.3%)	649 (79.7%)	81 (21.1%)	303 (78.9%)
Construction	180 (0%)		255 (0%)				316 (0%)			
	-	-	111 (43.5%)		144 (56.5%)		205 (64.9%)		111 (35.1%)	
	46 (25.6%)	134 (74.4%)	32 (28.8%)	79 (71.2%)	43 (29.9%)	101 (70.1%)	58 (28.3%)	147 (71.7%)	40 (36%)	71 (64%)
Motorway	3508 (0.5%)		3935 (0.4%)				4244 (0.4%)			
	-	-	3741 (95.1%)		194 (4.9%)		4125 (97.2%)		119 (2.8%)	
	114 (3.2%)	3394 (96.8%)	139 (3.7%)	3602 (96.3%)	21 (10.8%)	173 (89.2%)	166 (4%)	3959 (96%)	2 (1.7%)	117 (98.3%)
Motorway Link	3058 (0.4%)		3241 (0.4%)				3343 (0.3%)			
	-	-	2473 (76.3%)		768 (23.7%)		2875 (86.5%)		468 (14.1%)	
	81 (2.6%)	2977 (97.4%)	75 (3%)	2398 (97%)	30 (3.9%)	738 (96.1%)	91 (3.2%)	2766 (96.2%)	17 (3.6%)	451 (96.4%)
Path	769 (0.1%)		2580 (0.3%)				5081 (0.5%)			
	-	-	587 (22.8%)		1993 (77.2%)		2339 (46%)		2742 (54%)	
	55 (7.2%)	714 (92.8%)	42 (7.2%)	545 (92.8%)	61 (3.1%)	1932 (96.9%)	95 (4.1%)	2244 (95.9%)	95 (3.5%)	2647 (96.5%)
Pedestrian	2960 (0.4%)		3556 (0.4%)				3926 (0.4%)			
	-	-	2428 (68.3%)		1128 (31.7%)		3191 (81.3%)		735 (18.7%)	
	1645 (55.6%)	1315 (44.4%)	1355 (55.8%)	1073 (44.2%)	718 (63.7%)	410 (36.3%)	1860 (58.3%)	1331 (41.7%)	413 (56.2%)	322 (43.8%)
Primary Link	1115 (0.2%)		1250 (0.1%)				1286 (0.1%)			

	-	-	878 (68.3%)		372 (31.7%)		1155 (81.3%)		131 (18.7%)	
	80 (55.6%)	1035 (44.4%)	65 (55.8%)	813 (44.2%)	65 (63.7%)	307 (36.3%)	118 (58.3%)	1037 (41.7%)	31 (56.2%)	100 (43.8%)
Road	5710 (0.8%)		7705 (0.9%)				8179 (0.8%)			
	-	-	3695 (48%)		4010 (52%)		6257 (76.5%)		1922 (23.5%)	
	336 (5.9%)	5374 (94.1%)	205 (5.5%)	3490 (94.5%)	201 (5%)	3809 (95%)	328 (5.2%)	5929 (94.8%)	227 (11.8%)	1695 (88.2%)
Steps	3724 (0.5%)		5439 (0.6%)				6816 (0.7%)			
	-	-	3479 (64%)		1960 (36%)		5247 (77%)		1569 (23%)	
	222 (6%)	3502 (94%)	189 (5.4%)	3290 (94.6%)	78 (4%)	1882 (96%)	252 (4.8%)	4995 (95.2%)	62 (4%)	1507 (96%)
Trunk Link	4121 (0.6%)		4637 (0.5%)				4858 (0.5%)			
	-	-	3272 (70.6%)		1365 (29.4%)		4189 (86.2%)		669 (13.8%)	
	184 (4.5%)	3937 (95.5%)	156 (4.8%)	3116 (95.2%)	131 (9.6%)	1234 (90.4%)	267 (6.4%)	3922 (93.6%)	46 (6.9%)	623 (93.1%)
Unsurfaced	1451 (0.2%)		1453 (0.2%)				1413 (0.1%)			
	-	-	1257 (86.5%)		196 (13.5%)		1171 (82.9%)		242 (17.1%)	
	519 (35.8%)	932 (64.2%)	457 (36.4%)	800 (63.6%)	46 (23.5%)	150 (76.5%)	435 (37.1%)	736 (62.9%)	65 (26.9%)	177 (73.1%)
Total	**718581**		**874764**				**992209**			
	-	**-**	**628154 (71.8%)**		**246610 (28.2%)**		**807089 (81.3%)**		**185120 (18.7%)**	
	344684 (48.0%)	**373897 (52.0%)**	**312374 (49.7%)**	**315780 (50.3%)**	**111031 (45.0%)**	**135579 (55.0%)**	**403944 (50.0%)**	**403145 (50.0%)**	**83125 (44.9%)**	**101995 (55.1%)**
Change in Entities Population	-		**21.7%**				**13.4%**			

Table 6. The OSM Highways comparison results.

Starting from the general trend, it can be seen that the sheer number of OSM Highways is increasing. However, the level of increase has fallen between datasets Apr09 and Jul09. To reach sound conclusions regarding the overall trend of OSM content creation it is necessary to expand this analysis well over a 6 months period. However, it is also expected that as the OSM dataset becomes more complete the content generation will decline. A second interesting observation, closely related to the previous one, comes from the fact that the percentage of Identical Features between datasets Apr09 and Jul09 have increased by 28.5% (from 628,154 to 807,089), the New Features have declined by almost 25% (from 246,610 to 185,120). The most extreme example of this can be found in the Motorway category where a mere 2.8% was added in dataset Jul09 as the rest (i.e. 97.2%) had been already recorded in dataset Apr09.

Another observation comes from the comparison of the Identical Features between two consecutive datasets. In the majority of the cases, the Identical Features of a latter period are less than the total number of Features in the previous period. For example, in dataset Jan09 the Unclassified had 113,202 features. In dataset Apr09 the Identical Features to dataset Jan09 were just 93,947. This means that approximately 20,000 features either have been geometrically altered or assigned to another category or completely deleted. For the first case, the issue of geometric accuracy is raised; for the second the attribution process should be scrutinised; and the third is probably a matter of poor image interpretation (as in general, roads do not disappear). Interestingly, in total more than 90,000 features from the dataset Jan09 and approximately 65,000 features from dataset Apr09 belong to these cases.

Furthermore, regarding the name attribute, it can be seen that between dataset Jan09 and Apr09 the spatial entities with names recorded an increase of 22.8% in contrast to 20.7% for the entities without a name. The gap opens even more between datasets Apr09 and Jul09 as the entities with names increased by 15.0% in contrast to 11.9% for the entities without a name. This observation might be attributed to the fact that OSM users are becoming more experienced and thus more cautious when they collect road data. Such a change in the data creation attitude can result in the increase of the road entities with a name attribute.

Leaving the general trend and looking more closely to specific categories, a number of observations can be made. First, the allocation of the spatial entities to the OSM categories helped to understand the population and thus the relative significance of each category. More specifically, looking into the Jul09 dataset, the most popular category is the Residential (41.2%) followed by the Footway (14.5%) and the Unclassified (14.0%). Interestingly, compared to dataset Jan09, the Footway's share in the OSM Highways increased by 2% where as the Unclassified's share contracted by 1.7%. This provides a clear indication of the OSM contributors' preference in capturing certain road categories as it is obvious that reality (i.e. the spatial entities of the real world) did not have similar changes (Figure 39).

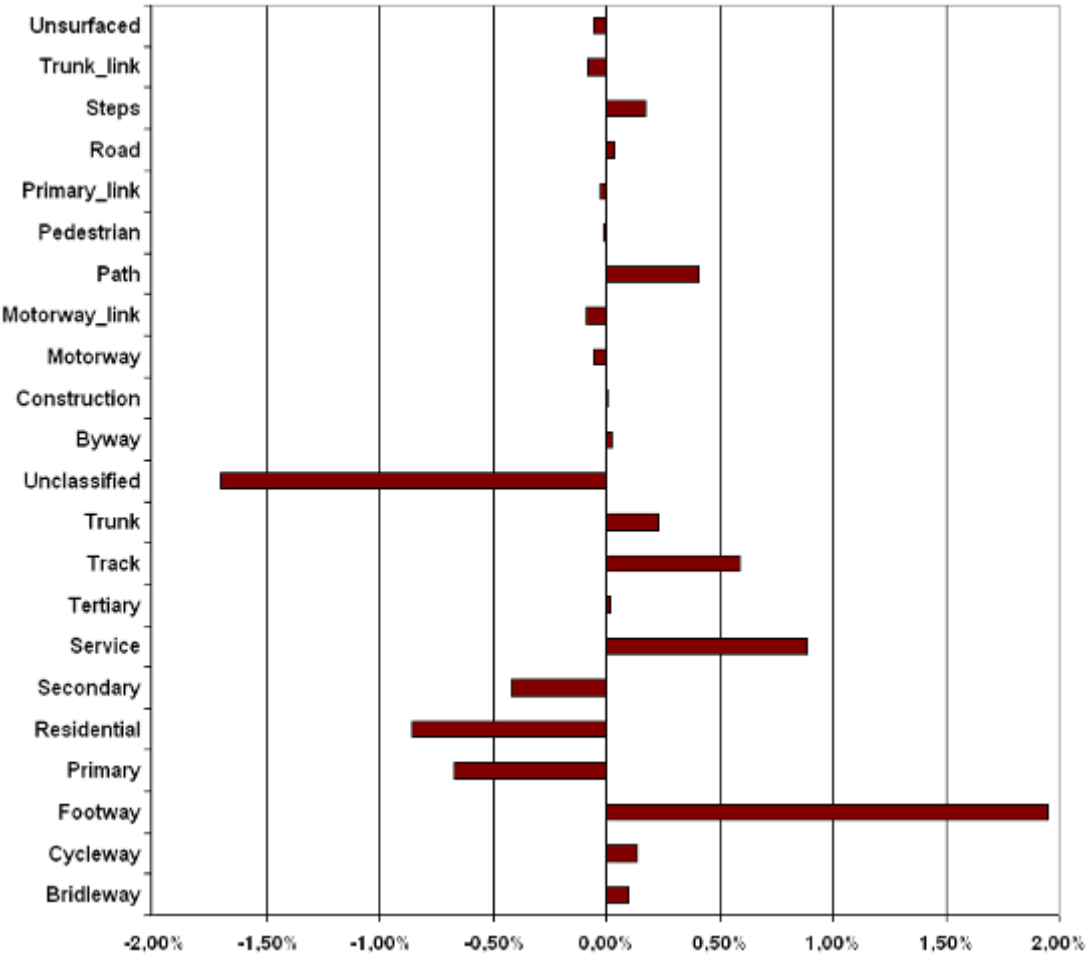


Figure 39. Relative change of the entities' share between datasets Jan09 and Jul09.

Returning back to the name attribute issue, in a category-level this time, it is expected that not all OSM Highway entities need to have a name attached. However, for certain

categories the name is a vital attribute. For example, while entities in Footways and Paths is uncommon to have a name (and hence this explains the low percentage of entities with names: 8.0% and 3.5% respectively), for Residential roads the name attribute is essential. Interestingly, for the latter category the percentage of spatial entities with names rose from 75.1% in dataset Jan09 to 83.8% for the newly introduced OSM entities in dataset Jul09. This further enhances the argument that OSM contributors started to become more careful during the content generation process.

This initial analysis of the OSM Highways data gave the opportunity to realise how the phenomenon evolves and to examine the particularities that appear during UGSC creation. One of the important observations of this initial step is the evolution in the attribution of the entities. The name, which is one of the most important attributes for certain road categories, increasingly gains the attention of OSM contributors. Still, the percentage of entities without a name is high. Finally, questions regarding the consistency of the geometric accuracy of the OSM datasets are raised.

5.3.2 POIs monitoring

A similar approach to the one discussed earlier was followed for the POIs dataset. Initially, the examination focused on the allocation of the spatial entities to different categories for all three datasets (Table 7). Noteworthy is a social aspect of the phenomenon as the primary focus of the OSM contributors seems to be the mapping of governmental and public services' entities (53.2% of POIs); information that is considered public and thus it should have been free and easily accessible. Moreover, the majority of the other categories are related to outdoors activities giving an indication of the OSM contributors' interests and possibly of their social background. In contrast, it is worth noting the particularly small share of the Shopping category with just 93 records (0.1%)

CATEGORY	Dataset Jan09			Dataset Aprl09			Dataset Jul09		
	Num. of Features	Share	Change	Num. of Features	Share	Change	Num. of Features	Share	Change
Automotive	18408	18.2%	-	23177	17.4%	25.9%	26838	17.0%	15.8%
Eating & Drinking	16210	16.0%	-	20815	15.6%	28.4%	24234	15.4%	16.4%
Government and Public Services	53248	52.7%	-	71330	53.5%	34.0%	83908	53.2%	17.6%
Health care	1161	1.1%	-	1486	1.1%	28.0%	1728	1.1%	16.3%
Leisure	2598	2.6%	-	3328	2.5%	28.1%	4022	2.6%	20.9%
Lodging	2422	2.4%	-	3290	2.5%	35.8%	4230	2.7%	28.6%
Night Life and Business	1315	1.3%	-	2019	1.5%	53.5%	2560	1.6%	26.8%
Shopping	66	0.1%	-	84	0.1%	27.3%	93	0.1%	10.7%
Sports	1590	1.6%	-	2032	1.5%	27.8%	2396	1.5%	17.9%
Tourism	3990	4.0%	-	5833	4.4%	46.2%	7629	4.8%	30.8%
Total	**101008**	**100.0%**		**133394**	**100.0%**		**157638**	**100.0%**	

Table 7. POIs categories

In the next level, regarding the POIs, the focus was on the evolution of the phenomenon. This was examined by comparing the datasets by pairs (i.e. dataset Jan09 against Apr09 and Apr09 against Jul09). In this comparison both the geometry and the attribution of the spatial entities was evaluated. As in the case of Highways, the analysis showed that each newer dataset does not include the entire data of the older one, rather a part of it (Figure 40).

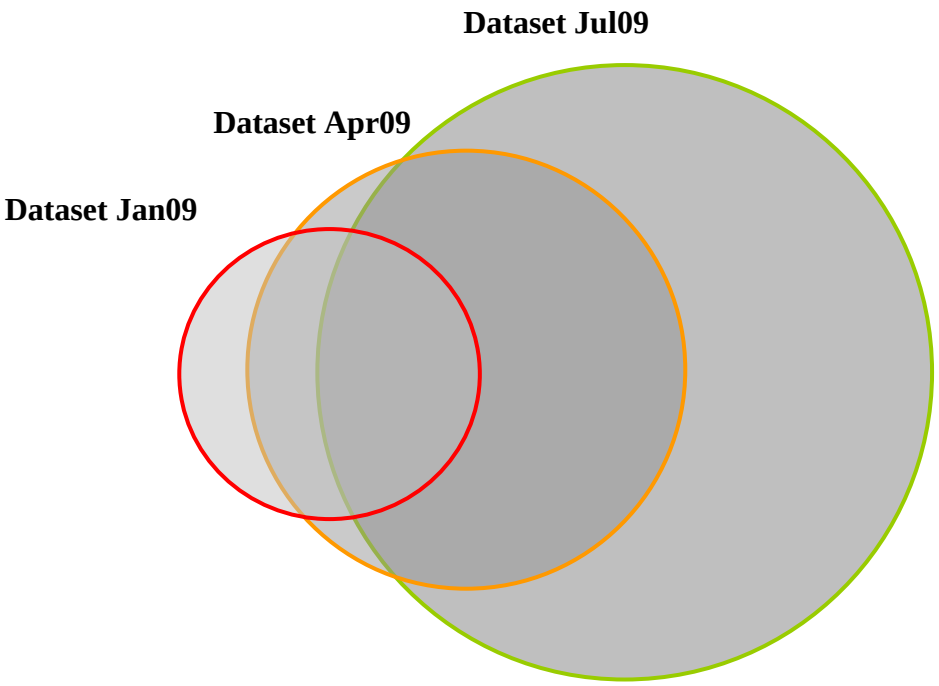


Figure 40. The evolution of the POIs datasets

5.3.2.1 Comparison between datasets Jan09 and Apr09

By comparing the dataset Jan09 against Apr09 it was realised that 6,876 (6.8%) POIs that belong to dataset Jan09 had been geometrically changed and thus did not appear at the same position in dataset Apr09 (Figure 41). On the other hand, 94,132 (93.2%) POIs of the Jan09 dataset were geometrically unchanged. Therefore, two separate issues needed to be examined. The first one was to examine why the geometric changes appeared and the second was to examine whether there were any attribute changes in the geometrically unchanged POIs.

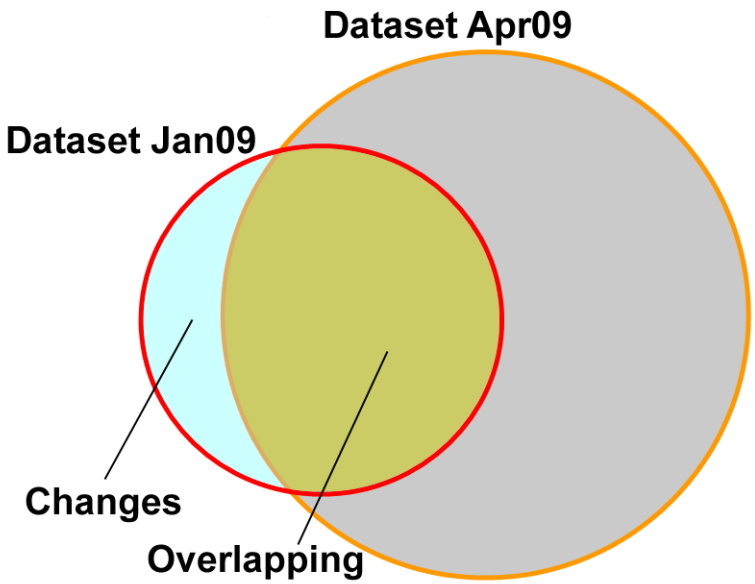


Figure 41. Spatially comparing dataset Jan09 POIs against dataset Apr09

Following the methodology described in 3.4.1.1 (i.e. by examining the POIs type and name, whenever that was available), for the geometrically changed POIs, it was possible to calculate how many POIs have been only geometrically changed, have both a geometric and an attribute change or have been deleted. The analysis showed that from the 6,876 POIs more than 50% have been only geometrically changed (i.e. no attribute change); approximately 8% have a change in their attributes and almost 40% have been deleted.

Regarding the geometrically unchanged POIs between Jan09 and Apr09, 906 (i.e. approximately 1%) have a change in their attributes: 161 POIs were assigned to a different category and 745 have a name change.

5.3.2.2 Comparison between datasets Apr09 and Jul09

A similar approach was followed when comparing Apr09 and Jul09 datasets. Here the analysis showed that 4,629 (3.5%, approximately half compared to the previous analysis) POIs that belong to dataset Apr09 had been geometrically changed and thus did not appear at the same position in dataset Jul09. Thus, 128,765 (96.5%) POIs of the Apr09 dataset were geometrically unchanged. Regarding the former group of POIs the analysis showed that from the 4,629 POIs about 36% have been only geometrically changed;

approximately 4% had also a change in their attributes and more than 60% have been deleted.

Regarding the geometrically unchanged POIs between Apr09 and Jul09, 897 (i.e. approximately 0.7%) have a change in their attributes: 230 POIs were assigned to a different category and 667 have a name change.

* * *

The preliminary OSM analysis gave a basic understanding of the project's evolution. Also, it helped to gain insight regarding the challenges and the potential pitfalls of the UGSC process. It was made clear that both the positional accuracy and the attribution process should be further examined as both issues affect greatly the overall quality of a dataset. In that context, the next steps of the analysis focused on the evaluation of the positional accuracy of the OSM Highways for England (Section 5.4) and the attribution quality of the OSM Highways and POIs for Great Britain (Section 5.5)

5.4 Positional accuracy analysis[9]

It has been discussed in Literature Review (Section 2.4.3) that there have been already some efforts for the evaluation of the OSM positional accuracy. However, these efforts have been contained to few small test sites in urban areas in London. Here, the effort was to move the evaluation of the positional accuracy quality element of the OSM road network to a national-wide level. As described in Section 3.4.1.2 the OSM positional accuracy was examined against the OS Meridian 2 dataset by using the road intersection nodes. Although Meridian 2 is constructed by applying a 20 metre generalisation filter to the centrelines of the OS Roads Database, this generalisation process does not affect the positional accuracy of node points and thus their accuracy is the best available (OS

[9] Parts of this Chapter have been adapted from the author's contribution to the following paper:

Haklay, M., Basiouka, S., Antoniou, V., Ather, A., 2010. How Many Volunteers Does It Take To Map An Area Well? The validity of Linus' law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4), pp. 315-322.

2009). Goodchild and Hunter (1997) recognise the use of the road intersections as a valid methodology to examine the positional accuracy of road network data against a reference dataset. They do raise a word of caution though regarding the need to positively match the corresponding nodes. Therefore a specialised algorithm was developed for the identification of the correct nodes between the two datasets, and the positional error was calculated for each node and as an average for each tile of the 1km^2 National Grid for England.

The positional accuracy evaluation analysis started with the creation of similar datasets for comparison. More specifically, as the OS Meridian 2 is a dataset that includes both the road segments as lines and the intersections as nodes, it was necessary to bring the OSM data (i.e. Geofabrik shapefiles) to a similar state. This step was necessary as the methodology followed was based on the intersections' comparison. Thus, a node topology was necessary to be created for the OSM data. The BUILD command of the ArcInfo Workstation was used for that purpose. The node topology building generated a separate node coverage (each node corresponds to an intersection or segment end) for OSM. This first step made the two datasets to have a similar structure. Figure 42 shows the road segments and nodes of OSM (blue) and OS Meridian 2 (red).

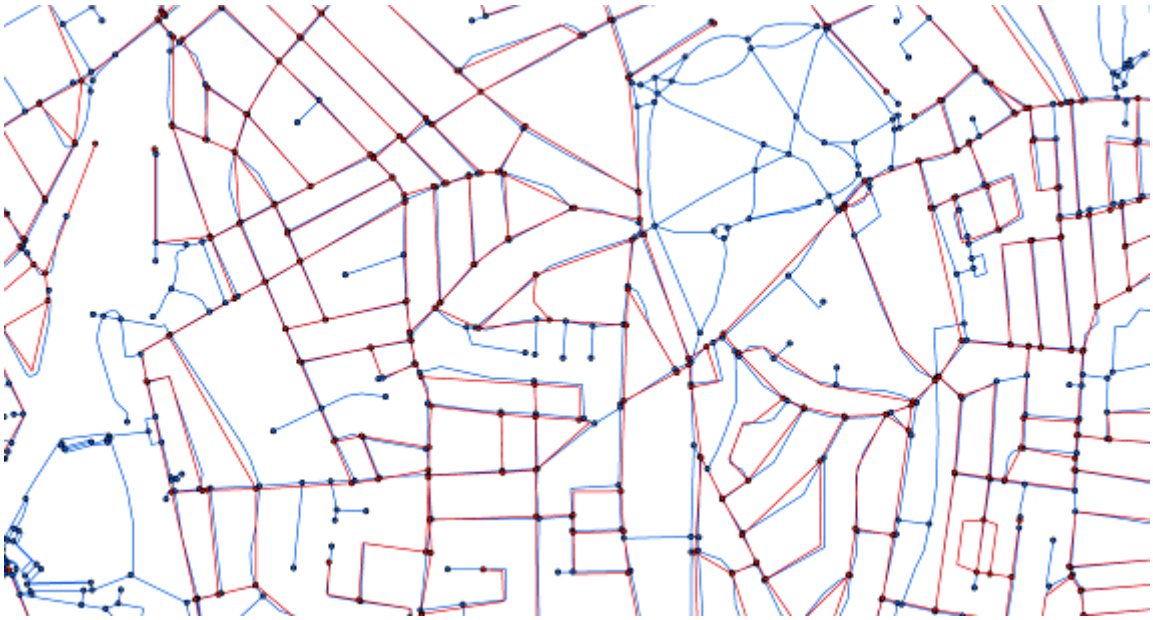


Figure 42. The road segments and intersection/end nodes of OSM (blue) and OS Meridian 2 (red)

A second step of data preparation before the positional accuracy evaluation was the road's name harmonisation. As explained in Section 3.4.1.2 the evaluation algorithm uses the name matching between the two datasets to record a positive match between two nodes. Therefore, it was necessary to harmonise the road names so to facilitate the name matching, otherwise disturbed by trivial and unimportant grammatical elements. For example, words like "Rd." and "St." where removed by both datasets. Finally, the names of both datasets were capitalised. There were two criteria for rejecting an otherwise positive match. First, the case where each node of a matching pair had one intersecting road (i.e. it is a road end) but the road names were different. Second, the case where each node of a matching pair had three or more intersecting roads, but the common road names were less than three. These criteria were implemented in the database level through SQL queries.

5.4.1 OSM and OS Meridian 2 data

After the datasets' restructure and harmonisation, the OSM dataset included 1,813,433 line segments and 1,406,616 nodes. On the other hand, the OS Meridian 2 dataset comprised 913,573 line segments and 639,554 nodes. All four datasets were inserted into a PostgreSQL (PostGIS enabled) spatial database for further analysis and spatial indexes were created.

5.4.4 Algorithm's evaluation

The final step before the algorithm's implementation was its evaluation. The algorithm was tested both for its efficiency (due to the large number of nodes needed to be examined and the multiple calculations - spatial or not - that needed to be performed) and the results yielded. Figure 43 shows the OS Meridian 2 (blue) and OSM (green) line segments and nodes from an evaluation test. The red vectors show the suggested node matching between the two datasets. The Figure also shows clearly the generalisation applied to the OS Meridian 2 line segments and thus the errors that a methodology based on these lines would introduce.

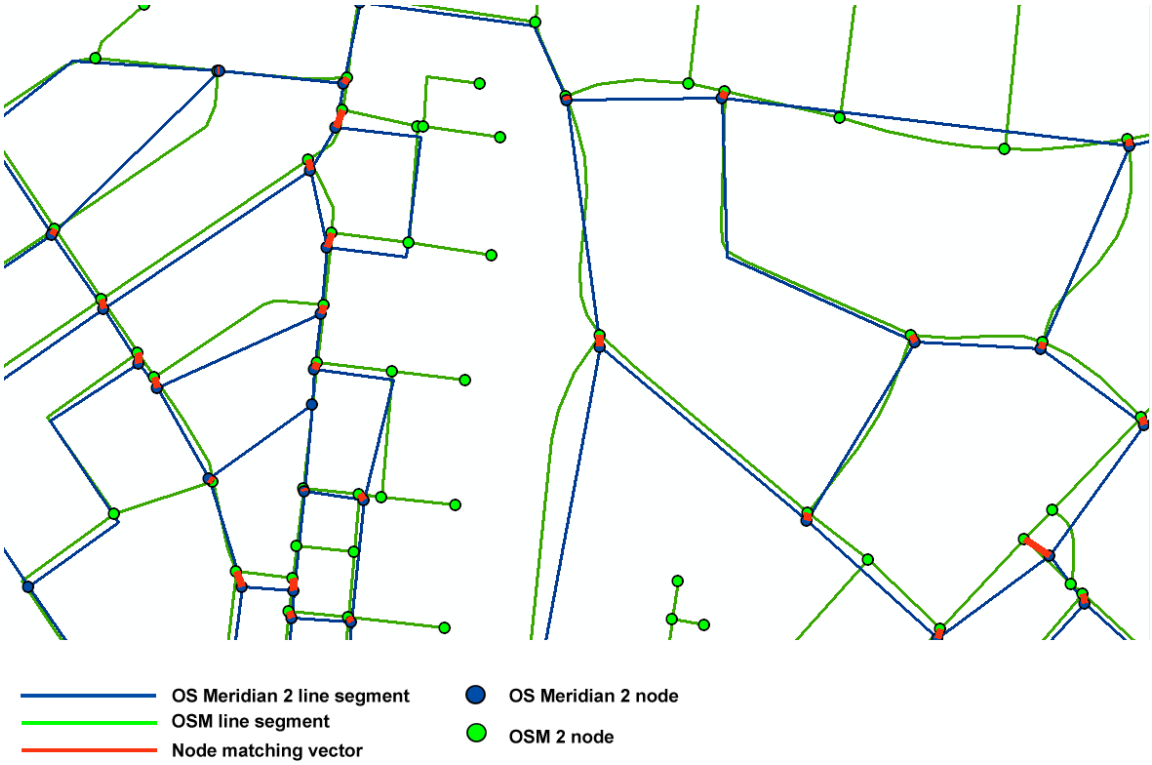


Figure 43. Positional accuracy algorithm's evaluation tests .

5.4.5 Positional accuracy results

The implementation of the algorithm presented in Section 3.4.1.2 created a dataset which included for each OS Meridian 2 node the following data:

i. The matching Geofabrik node ID.
ii. The distance between the two matching nodes.
iii. The number of road segments that start (or end) from the Meridian 2 node.
iv. The number of those Meridian 2 road segments that have a name.
v. The number of road segments that start (or end) from the Geofabrik node.
vi. The number of those Geofabrik road segments that have a name.
vii. The number of matching road segment names.

As expected, not every OS Meridian 2 node has a matching OSM node. In fact, a positive match has been recorded for the 383,810 (almost 60%) of the OS Meridian 2 nodes. However, the absence of matching OSM nodes does not necessarily mean the absence of OSM data. For example, Figure 44 shows the OSM (red) and the OS

Meridian 2 (blue) data for the National Grid SD 9107 tile. It can be seen that although there are no matches for the 39 OS Meridian 2 nodes, still OSM data exist for the road network. Nevertheless, the percentage of the positive matches gives an indication of the completeness of the OSM dataset. This factor could be further quantified with the results of Haklay (2010) regarding the OSM completeness analysis.

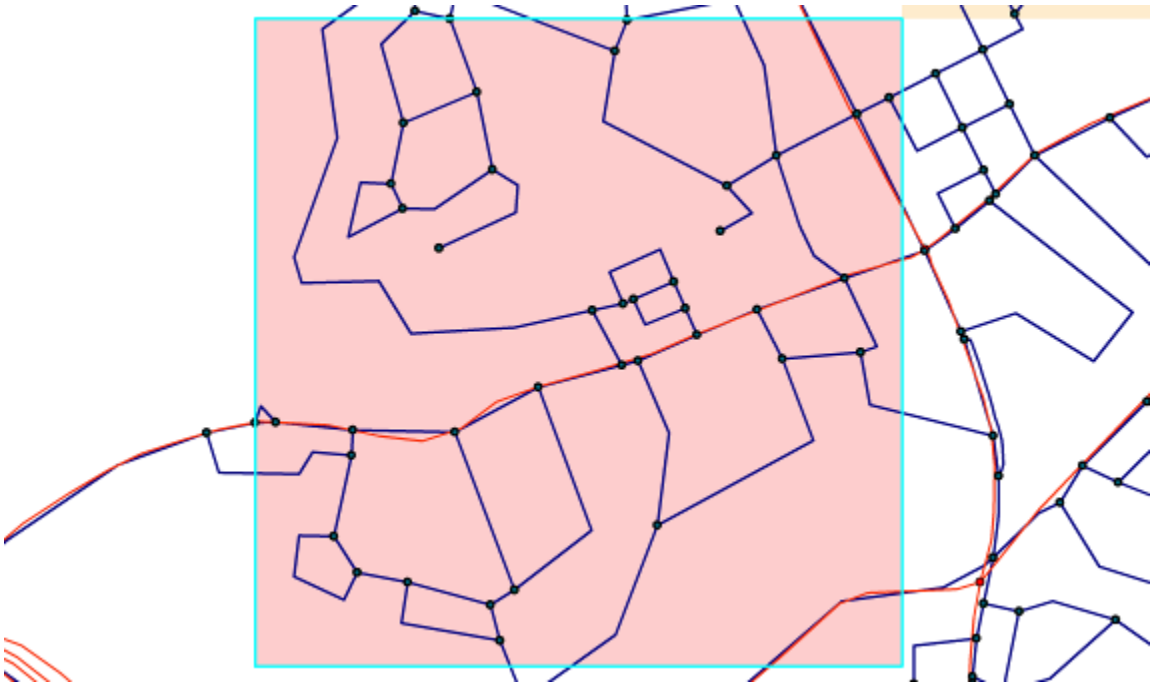


Figure 44. No recorded matching between OSM and OS Meridian 2 nodes.

In that context, only the OS Meridian 2 nodes and the OSM respective matching points were used for the rest of the analysis. Based on this sub-dataset, the average positional error of the OSM intersections is 7.90m with a St. Dev. equal to 7.02m. These results are rather impressive as they show that the average positional error is well inside the expected error from a simple hand-held GPS device. From another point of view, the findings could be read as if the average error is independent of the skill of the GPS user (i.e. professional or lay user).

Although the figure of the average positional error is an important index, it does not paint a clear picture of how this positional accuracy is distributed spatially. In other words, the positional accuracy findings need to be further analysed taking also into account the factor of space. To achieve this, the positional accuracy calculated for the

positively matched nodes was summarised using the tessellation of the 1km^2 National Grid. By doing so it was made possible to spatially examine the OSM road network positional accuracy on the one hand and to further compare and analyse the dataset with the findings of other experiments that also used the same grid (see for example the results from Haklay 2010).

Positive matches between OSM and OS Meridian 2 nodes were recorded for 55,192 1km^2 tiles. Figure 45 shows the results of the above described process. It is evident that the most accurate tiles are located in major urban areas like London, Liverpool, Manchester or Birmingham in contrast with the tiles in the rural areas that have generally larger positional errors (in Section 4.3.3 it was shown that the majority of users' activity and consequently the main clusters of geo-tagged photos are spotted at the urban and popular places while the places with fewer or no photos are located in the rural and less popular areas). For example, for the Greater London area there have been 53,079 matches recorded in 1,602 1km^2 tiles. The average positional error is 5.47m with St. Dev. 4.80m.

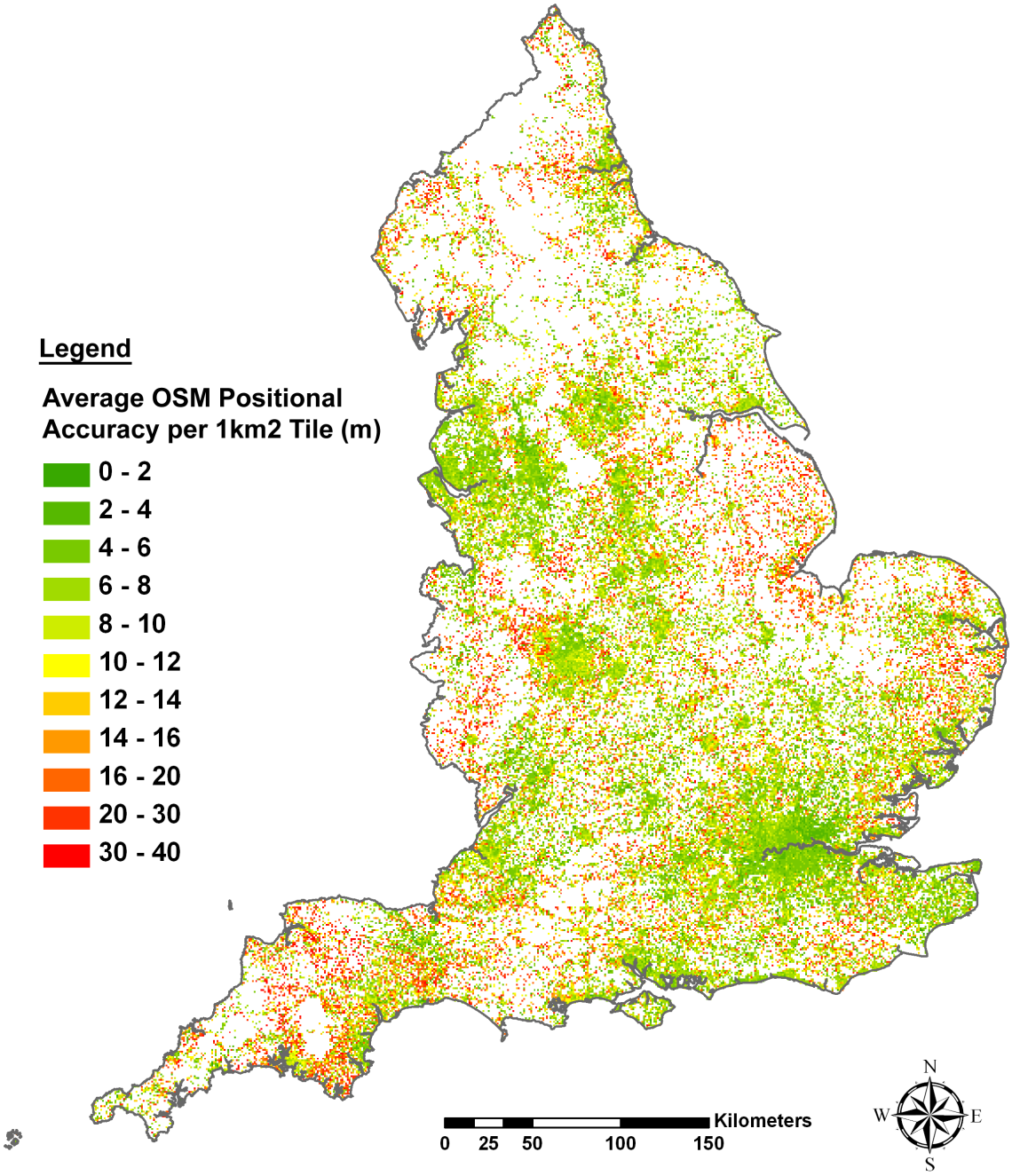


Figure 45. The evaluation of the OSM positional accuracy against the OS Meridian 2 intersections.
(For the legend's values the (a,b] rule applies).

On the other hand, the rural area around and north of Dartmoor National Park (which appears mainly white in southwest England as there are few or no intersections) up until Exmoor National Park has a considerably larger positional error. Figure 46 shows the district boundaries (in blue) of North Devon, Teignbridge, West Devon, Mid Devon, South Hams, Torridge and Caradon that cover this area. There have been recorded 8,015 matches in 2,911 tiles for these 7 districts. The average positional error is 13.34m with

St. Dev. 9.61m. The latter positional error is almost 70% larger than the average positional error for England and approximately 2.5 times the positional error recorded in the Greater London area. In turn, in Greater London area OSM has a positional error that is more than 30% smaller than the average.

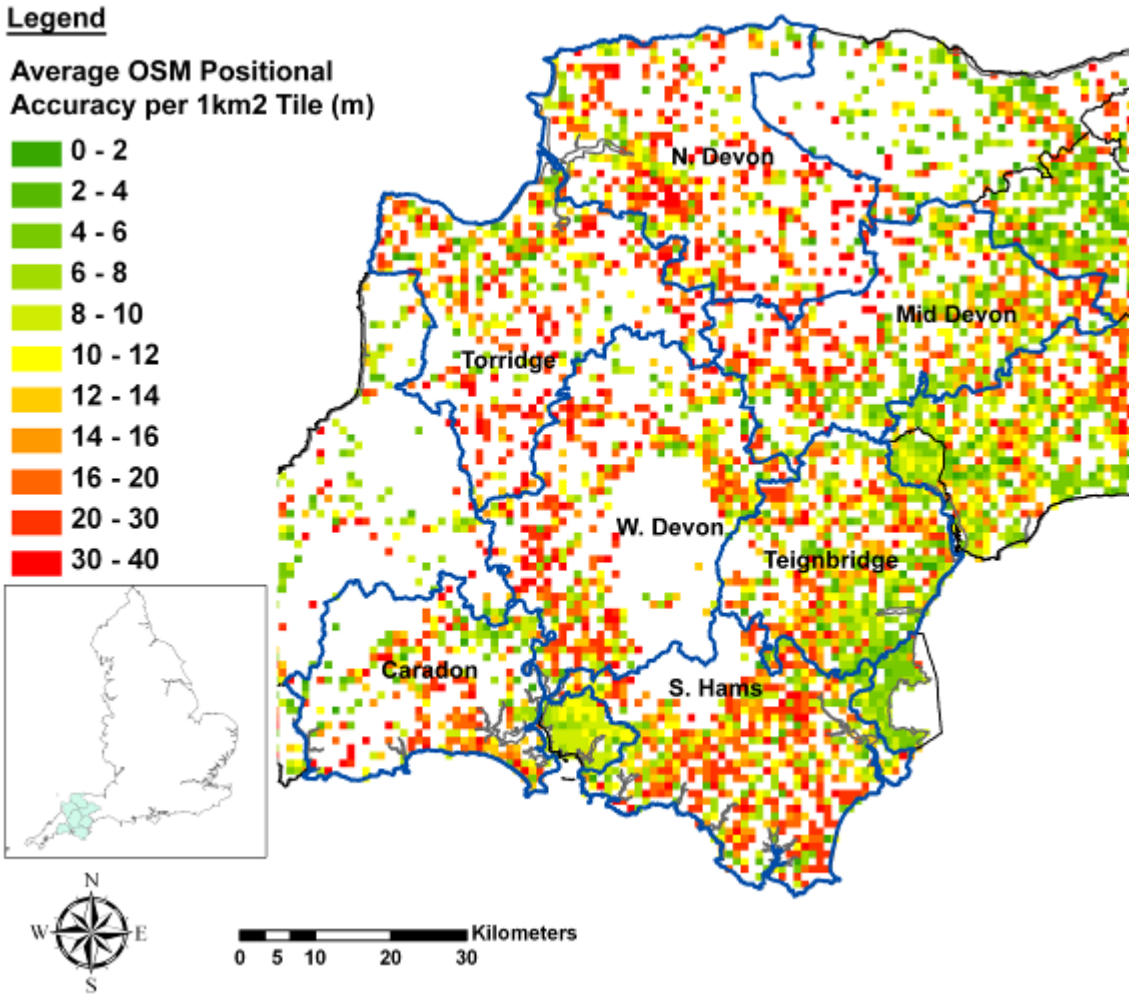


Figure 46. An area where OSM has large positional error (Devon, England).

Nevertheless, returning to the entire study area, given the means that users have at their disposal for the OSM data collection (i.e. hand-held GPS devices and digitisation from geo-referenced satellite images) the overall positional accuracy of OSM is considerably high (Figure 47). More than 70% of the intersections have a positional error smaller than 12m, another 10% is between 12 and 15m and 20% of the tiles have a positional error more then than 15m.

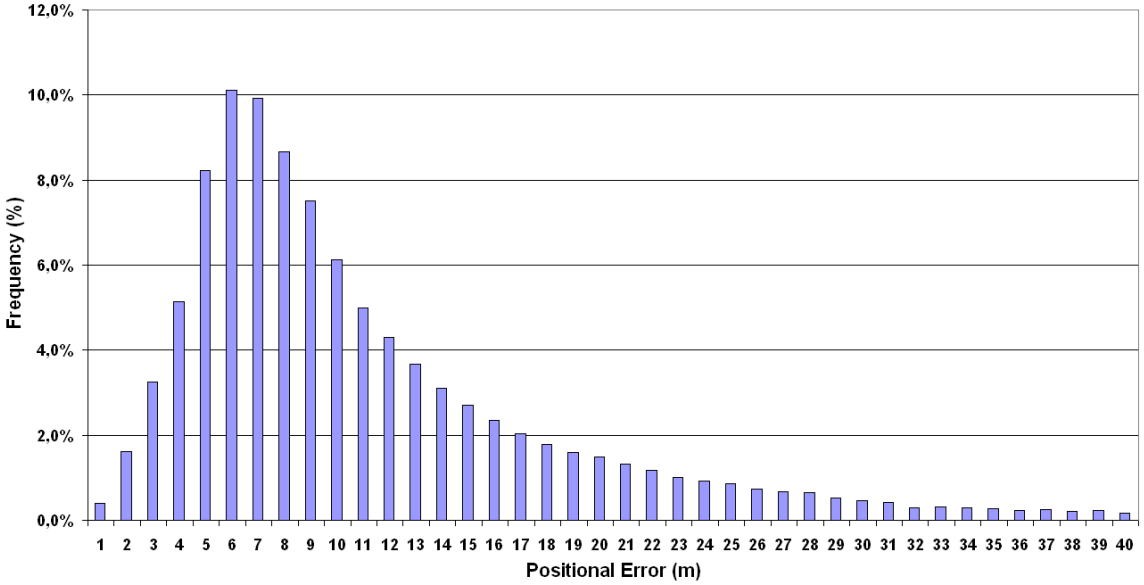


Figure 47. Frequencies of the positional errors of OSM data against the OS Meridian 2.

The calculation of the OSM positional accuracy for England enabled to perform two separate user-centric analyses in a national-like level.

5.4.6 Positional accuracy and users' participation

The first analysis involved the comparison between the number of OSM contributing users (the data was provided by Dr. Muki Haklay) and positional accuracy (Figure 48)

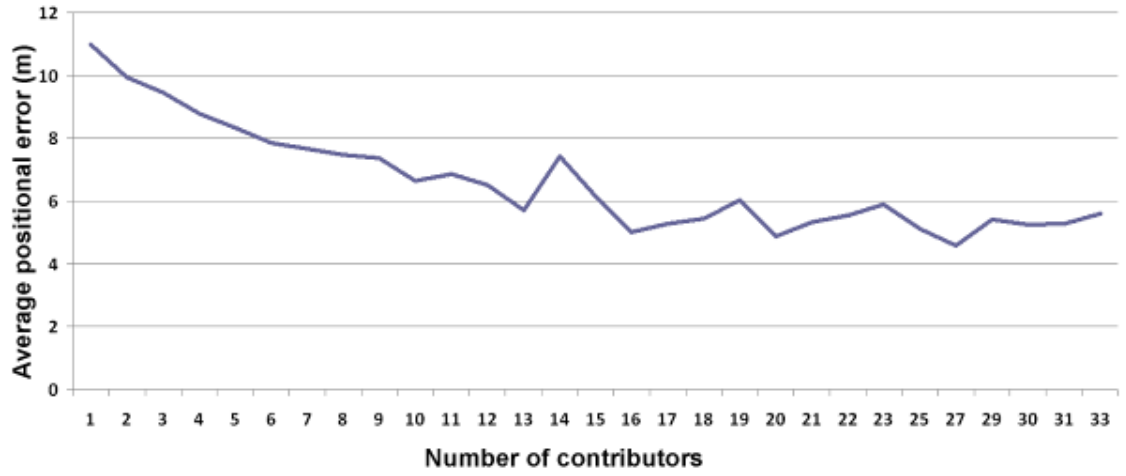


Figure 48. Average positional error vs. number of contributors to OSM road network for England.

Previous experiments have shown (see Basiouka 2009 and Haklay et al 2010) that there is no clear pattern of dependency between number of contributors and improved quality. However, the authors noted that the evaluation was conducted for limited urban areas and for 5 or more OSM contributing users.

In this evaluation the results yielded when examining the phenomenon in a national-wide level clearly show that the number of users contributing to an area is affecting the positional accuracy of the OSM dataset. Yet, this statement is valid up to a certain extent. It can be seen that positional accuracy remains fairly the same when the number of contributors is approximately 16 or more. On the contrary though, up until the number of contributors reaches approximately 13, each one of the users added in the pool of contributors, considerably improves the dataset's quality. Consequently, it can be supported that a similar phenomenon to the one described in Linus' Law (which dictates that "given enough eyeballs, all bugs are shallow", Raymond 1999) appears here. Indeed, up to a certain extent, the positional accuracy improves as more users are contributing and thus are active in an area. Also, it stands to reason to support that a similar behaviour could be observed to other UGSC sources or to other quality elements beyond positional accuracy.

5.4.6 Positional accuracy and completeness

The second experiment used the completeness data presented by Haklay (2010). The author showed that the level of completeness is linked with socio-economic factors (i.e. the Index of Deprivation 2007 that was created by the Department of Communities and Local Government, DCLG) as OSM contributors are providing less coverage to poor and marginalised areas compared to richer ones.

Here, the segregation line between complete and incomplete areas was used in order to group the positional accuracy in a similar way. Figure 49 presents the frequencies of the average OSM positional error for both complete and incomplete areas for the $1km^2$ tiles. With blue are painted the frequencies that correspond to tiles where OSM completeness level is higher than the OS Meridian 2 dataset and with purple those that are lower. It is

shown clearly that the positional accuracy in the complete (i.e. richer) areas is considerably better than the accuracy of the incomplete (i.e. poorer) ones. Indeed, the average positional error for the former is 9.57m with the St. Dev. equal to 6.51m. On the other hand, the average positional error for the incomplete tiles is 11.72m with a greater St. Dev. of 7.73m. This means that the OSM data for the incomplete/poor areas is almost 22.5% less accurate.

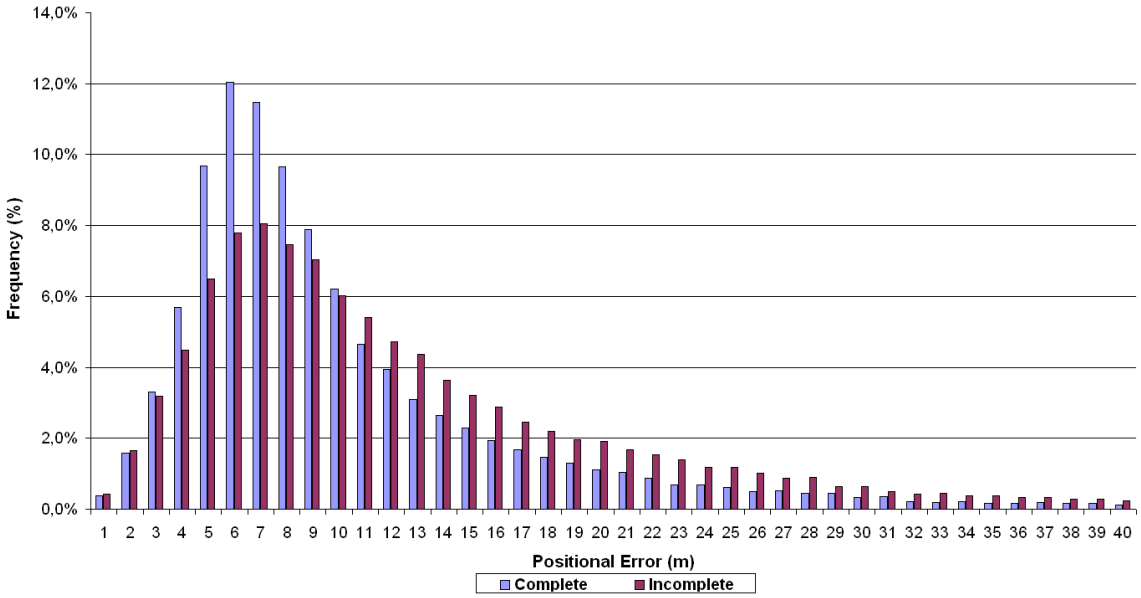


Figure 49. Average positional error vs. number of users for OSM road network for England.

Thus, it is shown that Haklay's findings regarding the effect of the social-related factors apply also to the positional accuracy quality element.

5.5 Tags analysis[10]

As explained in Section 3.4, apart from the geometric accuracy of the OSM's spatial entities, of high importance is also the attributes (and their quality) assigned to these entities. The analysis focuses initially on the statistics of the Highway's and POIs tags for Great Britain.

[10] Parts of this section have been adapted from:

Antoniou, V., Haklay, M., Morley, J. (2010a). A step towards the improvement of spatial data quality of Web 2.0 geo- applications: the case of OpenStreetMap. *Proceedings of the GIS Research UK 18th Annual Conference*. pp.197-202.

However, before analysing further the attribution of the OSM data, it will be helpful to fully understand the process of tag creation for OSM. The openness of a crowdsourced project like OSM has saturated most of the basic functionalities of OSM, including the attribution process of the spatial entities recorded. Neither the process of deciding the real world objects that need to be recorded to the OSM database nor the attributes that should describe those objects are controlled centrally. Instead both are commonly decided by the OSM users as described in Section 3.4. Following a voting process, the OSM community decides on the necessary tags in an effort to meaningfully describe the collected geometry. The analysis that follows examines the outcome of this process in an effort to realise both the development of the user's contribution and, most importantly, to gain the basic knowledge that will help to proceed with the evaluation of the attributes quality.

5.5.1 Initial tags analysis

The first step of the analysis was to collect the tags submitted for the OSM Highways of Great Britain. As explained, the OSM Highways describe the road network of Great Britain which is the most popular category with 1,286,992 spatial entities (Table 8). The tags associated with each one of those entities were collected. To complete this task the unique OSM IDs contained in the Geofabrik provided shapefiles were used. The outcome of this process was to collect 2,276,449 tags for 25 different OSM Highway categories (Table 8).

Num	Layer	Num. of Entities	Num. of Tags	Examined
1	Residential	514381	915547	Yes
2	Footway	198247	275376	Yes
3	Unclassified	165384	270421	Yes
4	Service	115541	124322	Yes
5	Tertiary	58334	118830	Yes
6	Primary	44034	132688	Yes
7	Secondary	34761	93957	Yes
8	Track	30408	58265	Yes
9	Trunk	30257	99046	Yes
10	Cycleway	23310	45642	Yes
11	Bridleway	14395	33899	Yes
12	Path	14373	22333	Yes
13	Steps	10351	10504	Yes
14	Road	9231	11439	Yes
15	Pedestrian	5830	11129	Yes
16	Trun_Link	5538	12072	Yes
17	Motorway	5077	21881	Yes
18	Motorway_Link	3541	9079	Yes
19	Byway	1626	5441	No
20	Primary_Link	1575	3062	No
21	Living_Street	542	977	No
22	Raceway	105	315	No
23	Sevices	94	123	No
24	Secondary_Link	40	59	No
25	Bus_Guideway	17	42	No
	Total	**1,286,992**	**2,276,449**	
	Total Examined	**1,282,993**	**2,266,430**	
	Percentage Examined	**99,69%**	**99,56%**	

Table 8. The number of spatial entities and the number of tags for each one of the OSM Highway layers. The first 18 layers where used in the analysis.

From the 25 categories available in the OSM Highways layer, the research focused on the first 18. This was deemed necessary to have enough data for examination and thus the research to lead to sound conclusions for a national level analysis. Yet, these 18 categories account for the 99.69% of the spatial entities and the 99.56% of the tags collected. The population of the spatial entities and of the tags, for the selected datasets, ranges from just few thousands (e.g. Motorway_links) up to more than half a million spatial entities and more than 900,000 tags for the Residential roads category.

The analysis started with the calculation of the average number of tags per spatial entity. It is interesting to note that despite the huge volume of tags submitted by the users to the OSM database in an effort to describe the spatial entities recorded (i.e. more than 2.2m tags), still the average number of tags per entity is relatively small (Figure 50). The OSM features with the highest average are the Motorways (4.3 tags per feature). The lowest average is recorded for the Steps (1 tag per feature). The majority of the OSM features have between 1 and 3 tags per feature in average.

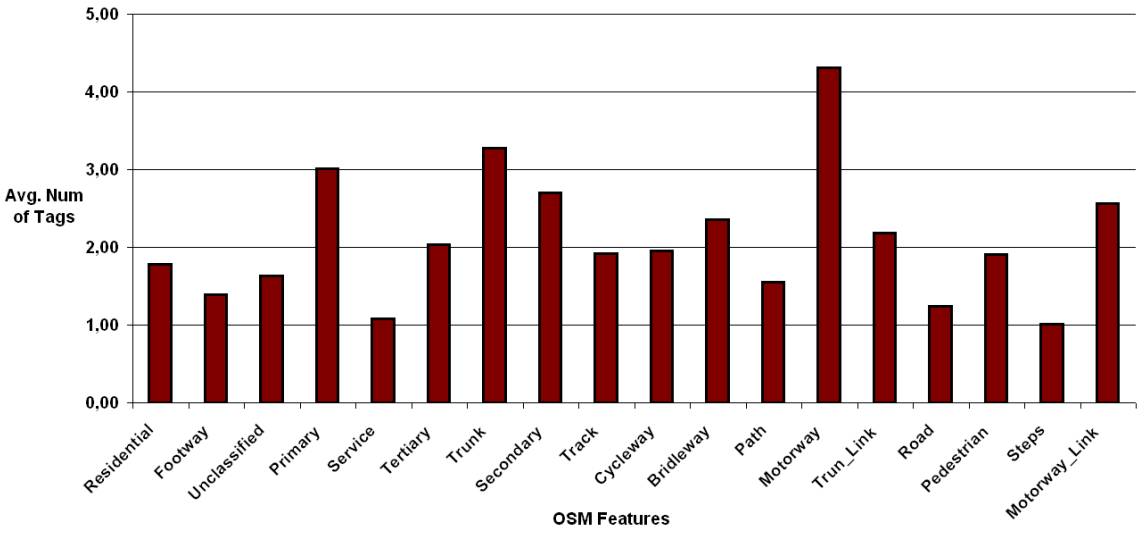


Figure 50. The average number of tags per OSM feature category.

Figure 50 gives a first indication of the completeness in terms of entity attribution and certainly indicates that the population of the tags will increase significantly in the future because new entities will be contributed to OSM and also because the average number of tags per entity will possibly grow.

This indication is further strengthened when examining the total number of tags for each category versus the unique tags recorded for each category. Figure 51 examines simultaneously two issues. The first one is the number of tags recorded for each of these 18 categories for Great Britain (as shown in Table 8). However, the most important point in this phase is the number of unique tags recorded for each OSM feature category. It can be seen that for the Motorways category there have been recorded more than 300 unique tags by the users in their effort to fully describe this specific spatial entity. Even more interesting is the fact that even for the least populous category (i.e. the Motorway_Link),

both in terms of spatial features and tags recorded, there are 47 unique tags. Here it must be noted that the connecting lines do not indicate continuity rather the possible evolution/trend of the phenomenon. For example, the lines show that when the Secondary roads reach the population of the Unclassified or Footway entities, the number of unique tags might increase in a similar way (the hypothesis of the increase in the population of each category is a valid one and is based on the results presented by Haklay 2010 regarding the completeness of the OSM dataset).

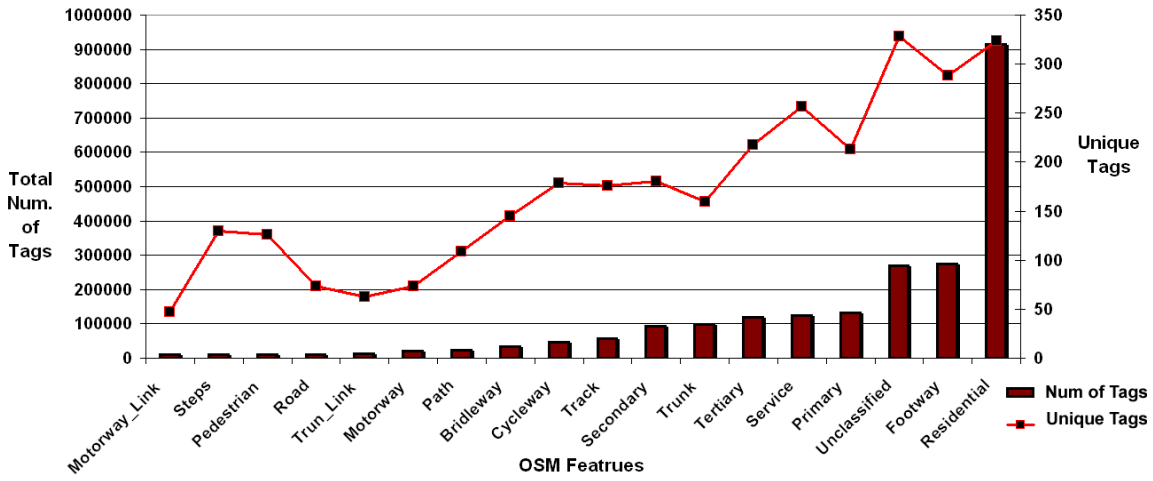


Figure 51. Unique tags vs. total tag population for each OSM features category in Great Britain.

An obvious observation here is that 300 unique tags (i.e. attributes) to describe the spatial entity of a motorway or even 47 to describe a motorway link are well beyond the actual tags needed.

Thus, two questions emerge from this observation. The first is how often a unique tag is introduced for an OSM feature category. The second is how many tags are actually enough to describe a spatial entity in each category (at least according to the OSM community).

Regarding the first question, in Figure 52 it is shown that the introduction of a unique tag depends in the total tag population of each category. So, for example, for the Residential roads, in average, there is a unique tag introduced for approximately every 3000 tags submitted where as for Primary roads a unique tag appears for every 600 tags.

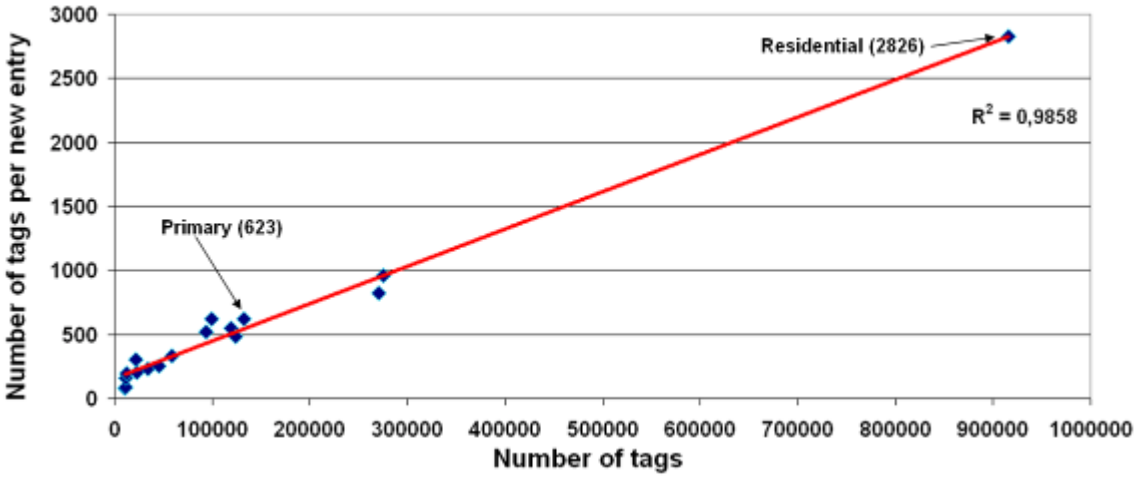


Figure 52. New tag introduction per OSM category versus the population of each category.

Regarding the second question (i.e. how many tags are actually enough to describe a spatial entity) Figure 53 shows the total number of unique tags per OSM category and the number of tags that cover the 95% of the total tag population in each category. For example, only 10 different tags account for the 95% of the total population of the tags submitted to describe a Residential road. For Unclassified this number is higher and reaches the 31 tags. Still, these figures are considerably smaller than the 324 and 328 unique tags recorded for each category respectively.

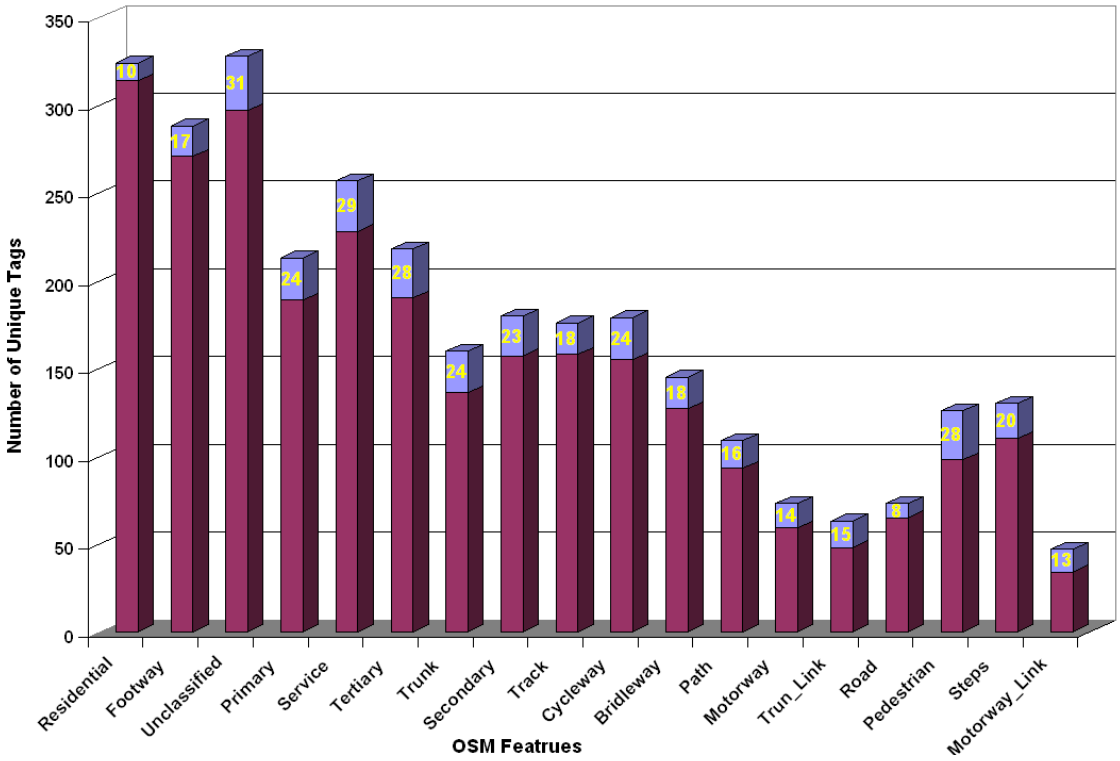


Figure 53. The total number of unique tags and the number of unique tags that account for the 95% of the total tag population for each category.

5.6 Tags' quality evaluation

The findings of the previous analysis showed that the attribution process of the OSM entities results in significant inconsistencies regarding the variety of the tags recorded by the users. The analysis showed that the open process of data creation introduces uncertainties. Such uncertainties can be presented in a dataset as errors (i.e. tags that are wrong as they do not correspond to an actual entity attribute), as noise (i.e. tags that do not add any significant information in an entity's description) and omission (i.e. tags that have not been submitted by the users and thus leave an entity's description incomplete). This fact deteriorates the overall quality of the datasets and raises the level of difficulty for third parties when it comes to using raw OSM data. It must be noted here that there must be no confusion between OSM raw data and Geofabrik or Cloudmade shapefiles as the former is the raw UGSC and the latter a processed product.

Therefore, it was deemed necessary to develop an evaluation process that would be able to quantify the uncertainty of the OSM attributes. The area of scope in this final stage of

the analysis was England. The first step of this evaluation process was to analyse the capturing instructions for spatial entities that are provided through the OSM wiki pages. These instructions were then transformed into rules that were used to evaluate the level of conformity of the data generated. The evaluation took place in three levels. First, a general attribution evaluation was conducted mainly focusing on the automatically generated OSM attributes that are assigned to spatial entities (Section 5.6.1). Also at this level, the focus was on the effect that the changes in the capturing instructions (both systemic changes and tag deprecation) have on the dataset quality. In the next level (Section 5.6.2), the formation of a conceptual schema based on the OSM wiki-pages for the categories Motorway, Unclassified, Residential and Path, took place. This conceptual schema was used to evaluate the spatial entities that belong to each category. In the final level a closer evaluation of the attribute domain consistency for the Motorway and Unclassified categories was conducted (Section 5.6.3).

There are multiple gains from this process. Firstly, it will reveal the types of errors introduced in the OSM datasets. Secondly, it will quantify the uncertainty generated by the users during the attribution of the spatial entities. Finally, this process will provide an unambiguous and consistent way to communicate the results of the quality analysis to any interested party.

5.6.1 General attribution evaluation

The methodology of data quality examination has been described in Section 3.4.1.4. The start was made by evaluating the OSM Highways category as a whole (there are 960,255 Highway entities for England). Initially, the evaluation tests were run for the attributes that are system-generated (i.e. these attributes are created by the OSM application and not by the OSM contributors). The entire presentation of the evaluation tests can be found in Appendix B. It can be seen that when examining the completion of the "osmuser" attribute (as this attribute is obligatory according to the OSM rules) the evaluation test shows that 0.55% of the Highway entities have not an OSM User ID recorded. In another test the domain consistency of the "rec_time" (i.e. recording time)

has been examined. The evaluation test shows that there are no domain inconsistencies for this system-generated attribute.

However, when examining a UGSC dataset the quality of the user-generated attributes is the most important element. Table 9, for example, shows the evaluation test that examines whether there are any spatial entities that have been assigned to a category other than the 25 Highway categories described in the OSM wiki pages (also presented in Table 8). Indeed, 0.21% (i.e. 2,009) of the spatial entities have been attributed with non-agreed category tags.

Data quality component	**Component domain**
Data quality scope	All items classified as Highways
Data quality element	2 - Logical Consistency
Data quality sub-element	2 - Domain consistency
Data quality measure	
Data quality measure description	Percentage of violating items
Data quality measure ID	N/A
Data quality evaluation method	
Data quality evaluation method type	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the domain of “highway” tags by count of highway features in the dataset
Data quality result	
Data quality value type	4 – Percentage
Data quality value	**0.21%**
Data quality value unit	Percent
Data quality date	2009-09-14
Conformance quality level	Not specified

Table 9. OSM Highways data quality: attribute domain consistency measure ("highway" tag).

At the next step the evaluation process moved to the examination of two main OSM categories: Highways and Points (in the Points category belong 238,636 spatial entities).

Three tests were conducted. The first was the evaluation of the conceptual rule that dictates that all OSM spatial entities should have at least one tag assigned to them. As

explained, the existence of tags is the catalyst that enables the proper use of the OSM data, and this explains why the OSM community has set this specific rule. Table 10 shows that the 0.16% of the Highway entities violates that rule. Only the 0.01% of the Points is violating the same rule (Table 11).

The other two evaluation tests examine the existence of the "created_by" tag, which although it is a system-generated feature tag (i.e. generated automatically by the OSM editors) is an interesting example as it eloquently describes how the changes in the specifications can affect the overall quality of the OSM datasets. A relatively recently (i.e. 30-04-2009) introduced OSM rule dictates that tags should not have the 'created_by' key. In the first evaluation test the data scope is all items in each category (Highways and Points) whereas in the second evaluation test the data scope is only the entities created after the introduction of the rule. As it can be seen in the Tables 10 and 11 the 63.76% of the Highways and 35.16% of the Points violate the OSM rule. However when the examination is contained in the spatial entities created after the introduction of the rule, the violation percentages fall to 25.60% for Highways and just 6.22% for Points.

In another similar example the examination focused on a deprecated tag that its use should be avoided by OSM contributors (Table 12). The use of the "highways = unsurfaced" key-value pairs for Highways was examined as this combination has been deprecated since March 2008. It can be seen that in total there are 1,248 spatial entities that violate the OSM rule and approximately half of them (601 entities) have been created after the publication of the rule. Moreover, shortly after the benchmark evaluation date of these tests the OSM community decided to deprecate the Highways' Byway category altogether. Consequently, 1,626 spatial entities (that have in total 5,441 tags) violate the domain consistency of the Highways dataset.

These evaluation tests show clearly that since the data quality changes whenever there is a change in the data (e.g. due to a transformation), in the ground truth or in the specifications of the product, the change of the OSM wiki-based rules does not leave the data quality unaffected.

Data quality component	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items classified as Highways	All items classified as Highways	All items created after the 30-04-2009 and have been classified as Highways
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	1 - Conceptual consistency	1 - Conceptual consistency	1 - Conceptual consistency
Data quality measure			
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID	N/A	N/A	N/A
Data quality evaluation method			
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the OSM rule: "all elements should have at least one tag" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the OSM rule: "tag should not have the created_by key" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features that have been created after the 30-04-2009 which violate the OSM rule: "tag should not have the created_by key" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result			
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0.16%**	**63.76%**	**25.60%**
Data quality value unit	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified

Table 10. Logical consistency evaluation tests for Highways

Data quality component	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items with Point Geometry	All items with Point Geometry	All items created after the 30-04-2009 and have Point Geometry
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	1 - Conceptual consistency	1 - Conceptual consistency	1 - Conceptual consistency
Data quality measure			
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID	N/A	N/A	N/A
Data quality evaluation method			
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the OSM rule: "all elements should have at least one tag" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the OSM rule: "tag should not have the created_by key" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features that have been created after the 30-04-2009 which violate the OSM rule: "tags should not have the created_by key" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result			
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0.01%**	**35.16%**	**6.22%**
Data quality value unit	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified

Table 11. Logical consistency evaluation tests for POIs.

Data quality component	**Component domain**	**Component domain**
Data quality scope	All items classified as Highways	All items created after the 19-03-2008 and have been classified as Highways
Data quality element	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - Domain consistency	2 - Domain consistency
Data quality measure		
Data quality measure description	Number of violating items	Number of violating items
Data quality measure ID	N/A	N/A
Data quality evaluation method		
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Count features which violate the OSM rule: “the key-value combination highway=unsurfaced should be avoided”.	Count features which violate the OSM rule: “the key-value combination highway=unsurfaced should be avoided”.
Data quality result		
Data quality value type	2 – Number	2 – Number
Data quality value	**1248**	**601**
Data quality value unit	Spatial Entities	Spatial Entities
Data quality date	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified

Table 12. Logical consistency evaluation tests for Highway deprecated tags.

5.6.2 Conceptual schema evaluation

In the next level of the analysis the focus turned to the conceptual evaluation of the entities that belong to the Motorway, Unclassified, Residential and Path categories. The total population of each category is presented in Table 13.

Num.	OSM Category	Number of spatial entities
1	Motorway	3,725
2	Unclassified	121,285
3	Residential	354,105
4	Path	5,845

Table 13. The population of the OSM categories evaluated.

It is easily understood that the important step of this process is the construction of a conceptual schema for each category based on the instructions presented in the OSM wiki pages. The compilation of the conceptual schemas was made through the study of both the main wiki page that describes each category (e.g. Motorways) and the wiki pages that this main page further links to (e.g. the tags "name", "bridge" or "tunnel" are further explained in dedicated OSM wiki pages). Additionally, the conceptual schema was completed with tags that are applicable for the category under examination (e.g. access restrictions for Motorways). Finally, the conceptual schema formation was concluded with tags that although are not explicitly stated in the description of a certain category, yet they are clearly implied as a good practice for entity attribution (e.g. the "URL" tag that is used from the OSM community to associate further information for each captured spatial entity, or the "FIXME" tag that is used to notify the rest of the OSM community that there is a need for further work on the specific spatial entity).

Table 14 presents the conceptual schema constructed for the four selected OSM categories (see Section 6.2 for further discussion on the conceptual schema and the attribute domain formalisation with the use of XML Schema). This means that only the tags listed in the table are in conformance with the conceptual schemas of the OSM entities that belong to Motorway, Unclassified, Residential or Path category. However, Table 15 presents a number of associated tags that can be used by the OSM contributors

to describe the spatial entities, without violating the conceptual schema, as long as the basic corresponding tag is already present.

Conceptual Schemas

Num.	Motorway Tags	Unclassified Tags	Residential Tags	Path Tags
1	access	abutters	access	access
2	attribution	access	attribution	attribution
3	bridge	attribution	description	bicycle
4	description	bridge	FIXME	bridge
5	FIXME	description	highway	description
6	highway	FIXME	image	FIXME
7	image	footway	lit	foot
8	lanes	highway	name	highway
9	lit	image	note	horse
10	maxspeed	lit	oneway	image
11	minspeed	maxspeed	smoothness	lit
12	name	name	source	name
13	note	note	source:name	note
14	oneway	oneway	source:ref	sac_scale
15	ref	smoothness	source_ref	ski
16	smoothness	source	surface	snowmobile
17	source	source:name	URL	source
18	source:name	source:ref	website	source:name
19	source:ref	source_ref	wikipedia	source:ref
20	source_ref	surface		source_ref
21	surface	traffic_calming		surface
22	traffic_calming	tunnel		tunnel
23	tunnel	URL		URL
24	URL	website		website
25	website	wikipedia		width
26	wikipedia			wikipedia

Table 14. Tags that describe the conceptual schema of the OSM Motorway, Unclassified, Residential and Path categories.

Num.	Basic Tag	Associated Tags
1	bridge	maxweight, maxspeed, maxheight, height, length, layer
2	tunnel	layer, maxheight
3	name	name_*, int_name, nat_name, reg_name, loc_name, alt_name, official_name
4	Ref	int_ref, nat_ref, reg_ref, loc_ref, old_ref, source_ref
5	minspeed	Units
6	maxspeed	Units

Table 15. Tags associated with the basic tags of the Motorways' conceptual schema.
*(The * is used to denote that any form of that tag is acceptable as the OSM is using tags like name_de to refer to German names)*

Once the conceptual schemas were constructed, it was possible to measure the violation/conformance of the OSM entities against the guidelines published by the OSM community itself. A spatial entity violates the conceptual schema of its category whenever it has tags that are not included in Tables 14 and 15. Table 16 shows the evaluation tests and the results yielded.

Generally, the conformance level of the categories is between 77.56% and 87.29%. Clearly, these are hardly acceptable conformance levels for a mapping agency. However, taking into consideration the loose OSM coordination model (e.g. wiki-pages and Mapping Parties) this level of conformance is a quite encouraging indicator for the phenomenon's evolution. More specifically, for Motorways, more than 22% of the total entity population violates the conceptual schema. This is the largest percentage among the categories examined. In contrast, only the 12.71% of the spatial entities that belong to the Unclassified category violate the conceptual schema. Another observation is that the conformance/violation level is independent of the number of tags that form each category's conceptual schema. Furthermore, the fact that a user needs to constantly consult the wiki pages to implement accurately the numerous guidelines was a factor originally expected to negatively affect the conformance level of the OSM spatial entities. However, judging by the conformance levels recorded it appears that this need has not affected severely the overall conformity of the OSM datasets. However, the level of the guidelines' adoption needs to be further researched taking into consideration at

least the users' commitment to the OSM project (i.e. regular or occasional users) as an independent variable. Another factor that explains this observation is the low tag average per entity seen in Figure 50 (Section 5.5.1). In other words the conceptual schema conformance comes from the fact that OSM users are submitting only few main tags for each entity and thus they stay inline with the community guidelines.

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items classified as Motorways	All items classified as Unclassified	All items classified as Residential	All items classified as Paths
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	1 - Conceptual consistency	1 - Conceptual consistency	1 - Conceptual consistency	1 - Conceptual consistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the conceptual schema of Motorways by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the conceptual schema of Unclassified by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the conceptual schema of Residential by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the conceptual schema of Paths by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**22.44%**	**12.71%**	**15.20%**	**13.40%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Table 16. Conceptual schema conformance evaluation

Up to now the focus was on the tags that the OSM users should use in their effort to describe the real world without violating the conceptual schema created by the OSM community. However, equally interesting is to examine the user-generated tags that fall outside the conceptual schema. The concept of the tag-cloud was used to visualise those tags (see Figures 54, 55, 56 and 57). The use of this visualisation shows instantly the tags that are more regularly used.

This examination of the tags that are not included in the conceptual schema serves three purposes. First, it enables a self-evaluation of the evaluation method followed in the Thesis. By visualising the tags that were left outside the conceptual schema it is easily understood if there is a number of tags that should have been embodied into it. Second, it provides a method to evaluate the suggestions of the OSM community and thus examine whether the behaviour of the OSM contributors follows a different pattern, regarding the choice of tags, other than those agreed by some members of the OSM community. Finally, using tag-clouds it is possible to instantly visualise if there are any common mistakes that the users are repeating during the entity attribution.

Figure 54 shows the tag-cloud for Motorways. A number of observations can be made. For example, ref:carriageway is the most common tag used by the OSM contributors although its use with the Motorway spatial entities is not suggested in the OSM wiki pages. Another example is the "Layer" tag. The use of this tag without concurrent existence of either the "tunnel" or "bridge" tag is vague. Similarly, the use of the "motorcar" tag does not add anything to the description of a spatial entity already characterised as a Motorway. Moreover, the use of a "horse = no" key(tag)-value pair for a Motorway does not really add anything. On the other hand though, it is worth considering the incorporation of the "toll" attribute to the description of a Motorway or even a tag to clarify the state of the road, such as "incomplete", despite the fact that there are only 2 such tags assigned to the entire Motorways population (i.e. 3725 spatial entities).

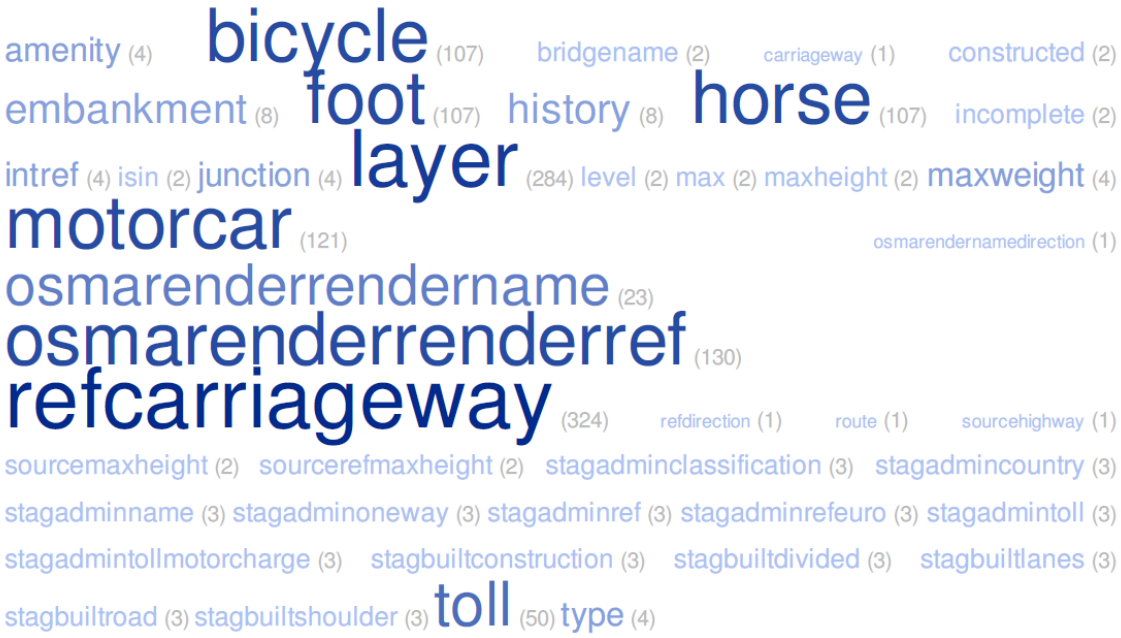


Figure 54. A tag-cloud formed by the OSM tags not included in the Motorway's conceptual schema.

Similarly, by examining the Residential's tag-cloud (Figure 55) it can be concluded that the "postalcode", "abutters" and "maxspeed" tags are popular among the OSM users. Again, the presence of the "motorcar" tag is not adding any crucial information as the spatial entities have been already characterised by the users as Residential roads. Moreover, the presence of certain tags, although not popular, should prompt further consideration regarding their adoption in the conceptual schema of the category. For example, the presence of tags such as: "area", "isin", "isinicity", "isincountrycode", "isincounty" and "isintown" shows an effort on behalf of the OSM users to clarify the administrative hierarchy of the spatial entities that belong to the Residential category. Another such example is the "width" and "noexit" tags that provide important information that can be used by applications (e.g. routing) and considerably enhance the overall value of a dataset.

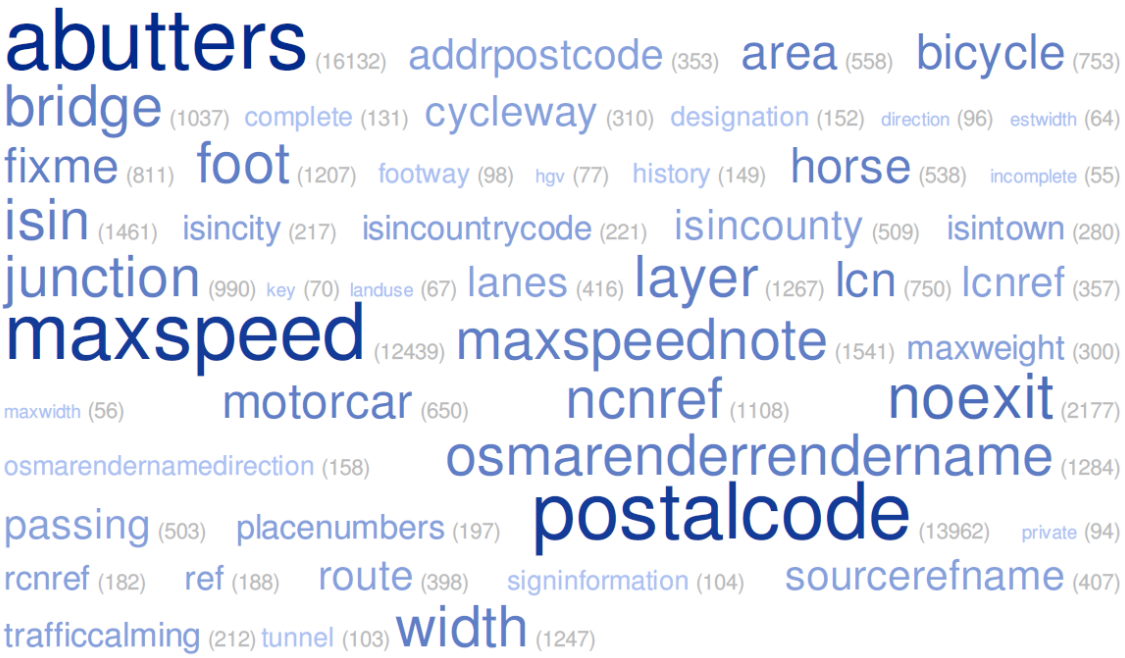


Figure 55. A tag-cloud formed by the OSM tags not included in the Residential's conceptual schema.

Similar observations can also be made for the other two OSM categories (i.e. Unclassified and Path). For example, the "postalcode" tag appears again to the Unclassified's tag-cloud while the "layer" tag (without properly connected to corresponding tags - i.e. "bridge" or "tunnel") appears again in the Path's tag-cloud.

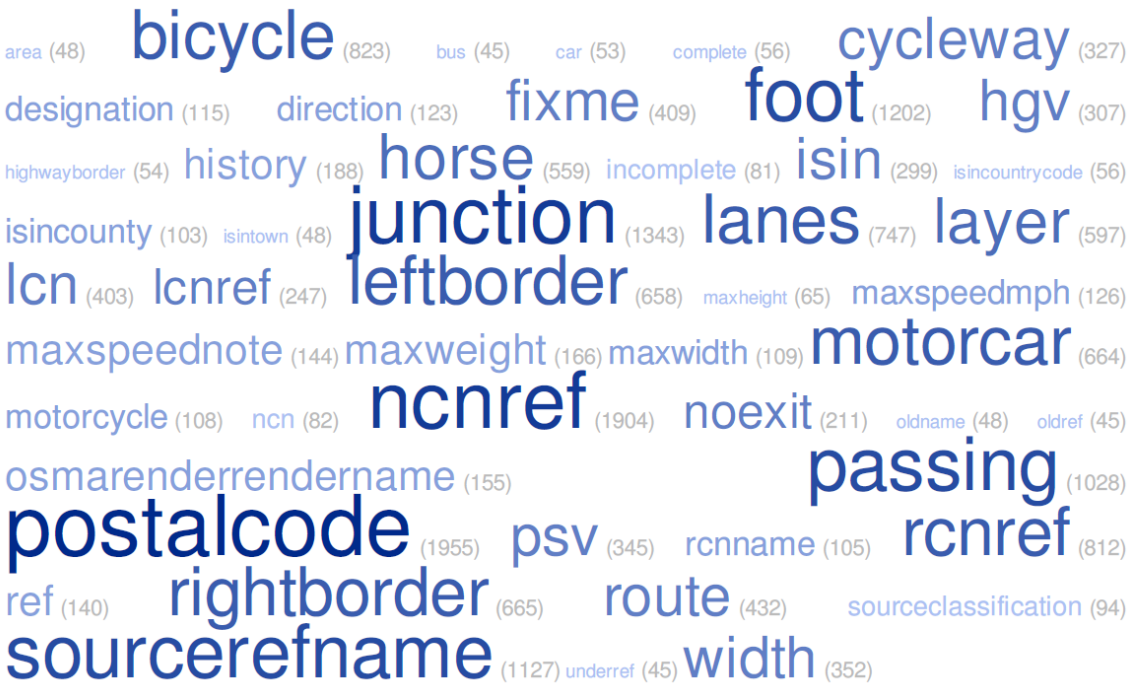


Figure 56. A tag-cloud formed by the OSM tags not included in the Unclassified's conceptual schema.

Figure 57. A tag-cloud formed by the OSM tags not included in the Path's conceptual schema.

Finally, it must be clarified that not all tags included in the conceptual schemas are needed for the description of every category's entity. Yet the absence of a firm conceptual schema makes it unclear for the tags missing whether they are not applicable to a certain spatial entity or they just have not been recorded by the OSM users. This uncertainty in the OSM data is deteriorating the overall quality. A methodology to face the challenge of improving the UGSC capturing process through the minimisation of the uncertainty (that appears as possible error, noise or omission) is discussed in Section 6.2.

These evaluation tests lead to three important observations. The first one is about the relationship between the data quality and the behaviour of the OSM community. Although, the OSM contributors adopt the rules introduced through the voting process, the level of adoption is not sufficiently high as more than one out of ten entities is not following the commonly agreed rules. The second observation is about the datasets' attribution quality compared to the needs of a mapping agency. For a mapping agency's standards the violation percentages are far too high and thus such datasets cannot be used "as is". Although a tangible conformance quality level that would cover the needs of a mapping agency has not been specified here, the OS level of data currency that dictates 99.6% completeness is indicative of the difference between the two types of datasets. The final observation is about the relationship between the data quality and the voting system itself. The change in the OSM rules leads to the deterioration of the dataset's quality. Interestingly, regarding the latter observation, it could be supported that such types of errors can be easily corrected in a database administration level. Yet, this would mean that a third party (i.e. the database administrator) would have to tamper the contribution of the OSM users. However, the adoption of such a policy is in direct contrast with the openness and freedom of the OSM project. Are the OSM administrators authorised to alter the contribution of the OSM community? Apart from some obvious system-created tags, where does exactly lies the red line that the administrative intervention cannot cross? In a sense, the OSM data quality is somehow trapped by the project's principles of openness, equality and freedom.

5.6.3 Tag's domain evaluation

The final stage of the analysis regarding the tag evaluation will focus on the tag's domain evaluation. The tag evaluation started with a general evaluation of the systemic and user-generated tags and continued with a more detailed evaluation of the OSM spatial entities against each category's conceptual schema. Here, the focus will be to realise whether the users are following the guidelines published in the wiki pages regarding the domain validity of the tags inside an OSM category. For example, the OSM guidelines dictate that whenever the OSM contributors use the "layer" tag they should add a value that

ranges from -5 up to 5 (e.g. "*layer = 3*" and not "*layer = +3*"). Similarly, the domain of the "access" tag is defined by the following enumeration: *unknown*, *yes*, *designated*, *official*, *destination*, *agricultural*, *forestry*, *delivery*, *permissive*, *private*, *no*. Thus, any "access" tag that has a different value violates the domain consistency as it has been suggested by the OSM community. The spatial entities of the Motorway and Unclassified categories were examined using the ISO evaluation methodology explained earlier.

The entire set of the results is presented in the Appendix B, However, here the discussion will focus to a few indicative cases presented in Tables 17 and 18 for the Unclassified and Motorways respectively.

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items classified as Unclassified with an "image" tag	All items classified as Unclassified with a 'maxspeed' tag	All items classified as Unclassified with an "access" tag	All items classified as Unclassified with a "surface" tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - Domain consistency	2 - Domain consistency	2 - Domain consistency	1 - Domain consistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID	N/A	N/A	N/A	N/A
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the rule: "*image tags should have an image url*" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the rule: "*Values are assumed to be in km/h unless units are explicit*" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the "*access*" enumeration constraint by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the "*surface*" enumeration constraint by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**94.22%** (=929/986)	**35.20%** (=1393/3957)	**1.88%** (=23/1221)	**1.41%** (=34/2418)
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Table 17. Domain evaluation for the Unclassified OSM category

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items classified as Motorways with a 'maxspeed' tag	All items classified as Motorways with an 'oneway' tag	All items classified as Motorways with a "layer" tag	All items classified as Motorways with a "lit" tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - Domain consistency	2 - Domain consistency	2 - Domain consistency	1 - Conceptual consistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID	N/A	N/A	N/A	N/A
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the rule: "*Values are assumed to be in km/h unless units are explicit*" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the "*oneway*" enumeration constraint by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the "*layer*" enumeration constraint by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the "*lit*" enumeration constraint by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**57.31%** (=192/335)	**23.18%** (=862/3719)	**0.54%** (=8/1471)	**0.00%** (=0/439)
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Table 18. Domain evaluation for the Motorways OSM category

As it can be seen, for both categories the violation/conformance levels of the OSM tag domains range significantly. For example, there are tags that present very high levels of domain violation like the "image" with 94.22% and the "maxspeed" with 57.31% for Unclassified and Motorways respectively. In contrast, there are tags that present complete (or almost complete) compliance with the OSM community-introduced domain constraints. For example, from the 2,418 Unclassified entities with a "surface" tag only 34 (i.e. 1.41%) have a value that is not listed in the OSM enumeration domain. It is interesting to note here that there are 22 valid values that the OSM community has adopted when it comes to describe a roads surface. For Motorways, the compliance of the values assigned to the 'lit' tags with the enumeration domain is 100%. All 'lit' tags attributed to the 439 Motorway features are in accordance with the wiki-based guidelines.

5.7 Summary

Chapter 4 focused on the analysis of geo-tagged photos. This Chapter concludes the empirical research as it focuses on the analysis of the second big family of UGSC on the Web: vector data.

The analysis started with a preliminary examination of the phenomenon's evolution through the comparison of three different OSM datasets. This initial step helped to realise both the potentials and strong points of a project like OSM as well as the possible erroneous processes or challenging issues that are inherent to OSM datasets and possibly to UGSC in general. Two particular issues stood out: the positional accuracy and the features' attribution. Both issues are very important quality elements for a spatial dataset and thus any discussion about a mapping agency engaging with UGSC without this fundamental knowledge is fairly shallow. Therefore, closer analyses were conducted for these issues in national-like levels.

For both issues the results yielded are twofold as there are both promising and challenging issues. Regarding the analysis on the positional accuracy, it was realised that the OSM accuracy is such that can be possibly used by mapping agencies. However, the

effect that socio-economic factors can have on the positional accuracy should be an issue of concern for the OSM community. If proper attention is not paid, a two speed spatial dataset is possible to emerge. Regarding the attribute's quality the analysis showed a mixed picture as well. There are many cases where features' attribution is in high conformance with the community's guidelines. However, equally many are the cases of low conformance. This volatility in the results can be attributed to the freedom and the loose coordination that exists throughout the data capturing process. The formalisation of this process in the context of an open crowedsourced project remains a major challenge.

Chapter 6

Challenges and solutions

6. Challenges and Solutions

6.1 General

This Chapter will focus on two of the most important challenges revealed during the empirical research. These challenges are closely connected with the development of Web 2.0 geo-applications that aim to function as UGSC sources.

For mapping agencies, when it comes to engaging with UGSC, one option is to co-operate with existing Web 2.0 geo-applications. To do so effectively, mapping agencies should be in a position to understand the nature and the main characteristics of such sources and to have a clear view on the data quality issues. Thus far the Thesis' centre of gravity has been in these issues. However, another option could be the development of their own Web 2.0 geo-applications so as to create in-house UGSC able to fuel their spatial repositories or become the basis for new products. In this case on top of the requirements mentioned earlier there is a compelling need to tackle challenging issues that can affect the overall outcome. Obviously, the combination of these two options is not excluded. For example Ordnance Survey has built its own Web 2.0 geo-application (OS explore) while at the same time is co-operating with Geograph.

The first challenge met here stems from the findings of the OSM attributes quality evaluation analysis: the need to formalise the data produced from UGSC source (Section 6.2). The second challenge met is the interactivity enhancement of the Web 2.0 geo-applications (Section 6.3).

6.2 Challenges and solutions: Data formalisation and quality improvement

6.2.1 The context in the OSM case

Given the results of the OSM tags analysis, it is clear that there is a need to integrate quality-evaluating and quality-preserving procedures in Web 2.0 geo-applications. This should not be just another impressive technological feature of the Web 2.0 applications but rather a conscious effort to use technological advances to build quality-aware geo-applications and to train lay users to embrace the spatial data quality principles. However, the development of a mechanism that will be able to reveal errors and prompt users for corrections is not a straightforward task. As shown in the course of this Thesis, the high-level end of such a mechanism involves the general evaluation of the applications' data repository in order to reveal social-generated content imbalances (see more on Section 7.3 for that). The focus now is on the low-level (i.e. entity level) functionality that Web 2.0 geo-applications should have.

As explained, the attributes' evaluation can be performed without external help as long as there is an unambiguous attribution process or a clear product's specification (e.g. specific name and number of attributes for each spatial entity, attributes' types and domains etc. - see also the discussion on Section 5.6). Embodying this case of quality evaluation in the functionality of a Web 2.0 geo-application and communicating it to the users so to act on it, is a challenge that all interested to the UGSC phenomenon parties should tackle.

The analysis of the OSM attributes gave a number of empirical examples that clearly show the level of inconsistency introduced in OSM datasets. Indeed, the second stage of the OSM analysis (i.e. the tag analysis in Section 5.7) made clear that the open process of entities' attribution adopted by the OSM community, has inherent faults that lead to the introduction of noise in the OSM datasets. The wiki pages created to help users, by communicating to them the correct way of attribution, have not proved to be enough in leading them to use the commonly agreed tags. A fair share for this heterogeneity in the

tags used can be attributed to the simple fact that the users are not (fully) aware of the entire range of wiki pages before they start digitising or collecting data in the field. Thus, their individuality affects the data contributed as their conceptualisation of the real word differs. Another important reason is the functionality of the available OSM editors. When this research was conducted, the most popular editors (i.e. Potlach and JOSM) were just beginning to integrate logical rules in their functionality (a strategy that is only expected to continue) and consequently much of the data attribution process was introducing errors to the OSM data repository. In other words, the necessary quality rules are disassociated from the data generation process: the quality rules are described in the wiki pages and there is little or no linkage between them and the OSM editors. On the contrary, in mapping agencies mechanisms like database schemas, attribute domains, topological rules and spatial features' extraction guides (i.e. methodology for data digitisation from satellite imagery) are everyday practices that are followed to the letter.

The OSM community by not following the existing practices failed to manage efficiently the spatial entities' semantic heterogeneity that was expected to appear from such a motley crowd. An important issue here is that the uncertainty introduced can be propagated to the neighboring entities although other quality elements might not be affected (e.g. positional accuracy). Figure 58 illustrates such an example.

Figure 58. Attribute's discrepancies between OSM users.

Suppose that although the geometry of two adjacent road features is accurate, there might be discrepancies in the tags contributed to each part of that entity. Such discrepancies in features' tags affect the neighbouring entities and thus multiply the uncertainty.

6.2.2 XML schema

Thus far, it has been made clear that the uncertainty in the UGSC needs to be minimised. The first step to achieve this is to provide a clear specification of the data sought to be produced. This is not to say that there is a need to have a strict and inflexible product specification but rather the opposite. However, at each given point there must be a clear view of how the data should be structured. In other words, the universe of discourse is possible to change over time but its evolution should be clearly defined and mapped.

A discussion about the specification's change mechanism is out of the scope of this Thesis not least because it will have to be defined according to the particularities of each Web 2.0 geo-application. For example, it could be decided centrally by the application's administrators like in Geograph or a bottom-up approach (through a voting system) like in OSM. Another interesting option would be the actual UGSC to have a more active role in this mechanism. For example, in the case of OSM the voting process could start not only from a user's proposal but also from the actual tags' generation process. New tags could be introduced automatically for voting once they reach a critical threshold (or become deprecated if their percentage is less than a given threshold).

Apart from the change mechanism, another important issue is the methodology that could be used to describe UGSC's specification. Here, the use of the XML Schema was chosen for the development of the OSM specification. As discussed, OSM is already using XML-encoded files for data transfer, manipulation and sharing.

The first step for the development of the formalisation mechanism is the analysis of the rules included in the OSM wiki pages. The results of this analysis were used to translate the user defined OSM rules into an XML Schema that would formally describe the OSM

data specification. This part took place in parallel with the process of the OSM's tags evaluation analysis. In fact, XML Schema fragments were used for the tags' quality evaluation presented in Section 5.6. This resulted in the formation of an XML Schema that modelled accurately the commonly agreed rules by the OSM community (this work did not cover the entire OSM data range due to time limitations).

The XML Schema creation process started from the OSM categorisation of the real world entities into physical (man-made) and nonphysical ones. Further categorisation was implemented for all the main entity categories included into the OSM wiki pages (Figure 59). As discussed in Chapter 5, the focus was on the Highway category and more specifically on selected entity types from the Roads and Paths sub-categories.

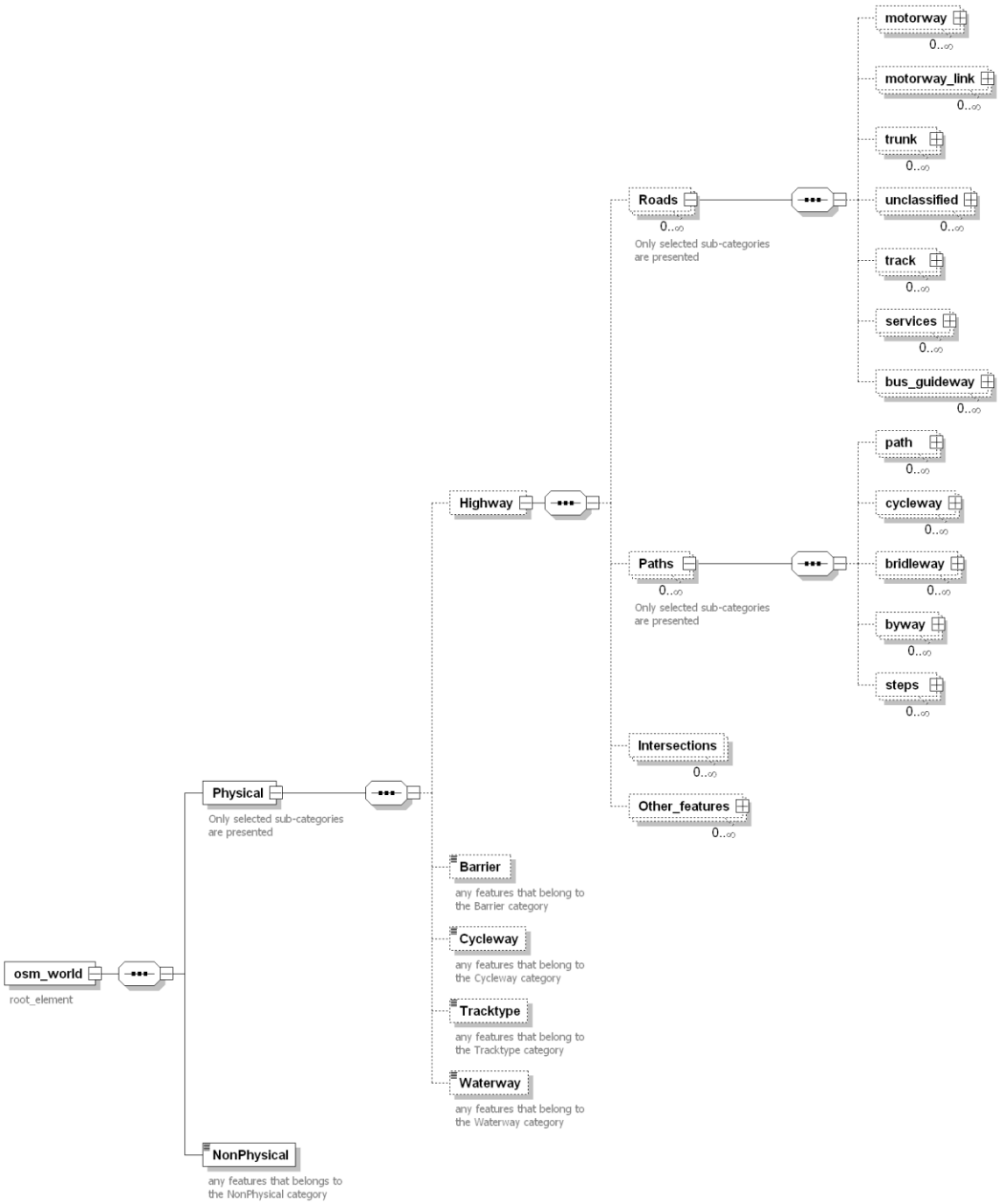


Figure 59. XML Schema for modelling the formalisation of the rules included in the OSM wiki pages.

The methodology adopted for the modelling of the OSM data was helped by the findings of the tags' analysis (see also Section 5.5) and follows a hierarchical model. More specifically, the analysis showed that a number of tags is repeated for all OSM recorded entities. For example, according to the suggestions in the wiki pages, there are annotation-related tags that can be attached to all spatial features (e.g 'fixme', 'description', 'URL' etc.). Apart from the user-generated tags there is also a number of

attributes that are assigned by the OSM application automatically (e.g. ‘osm id’, ‘user id’, ‘user’ etc.). Combining all this, an OSM object type was created that included all the attributes that can define an OSM spatial entity and each spatial entity was assigned the OSM object type (Figure 60).

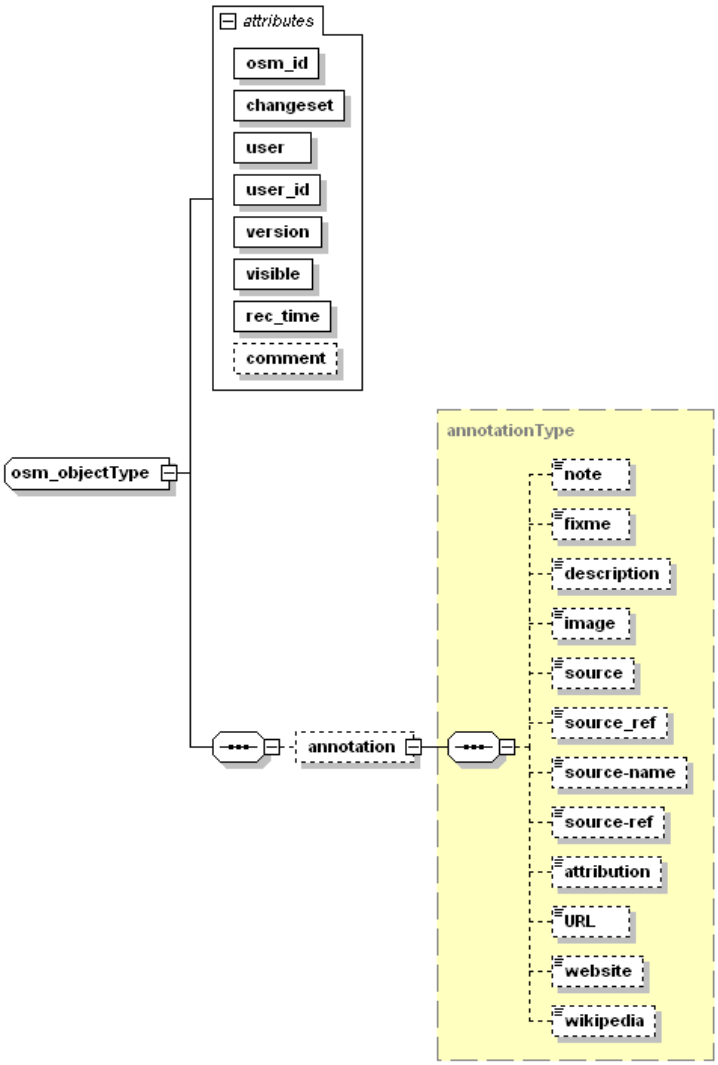


Figure 60. The XML Schema fragment of the OSM object type (the continuous line denotes an obligatory attribute whereas a dashed line means an optional one).

Next, each OSM entity was created by extending the basic OSM object type and completing it with further attributes from two sources. The first source was the OSM category that each entity belonged to. The second source was the attributes relative to each specific entity. For example, Figure 61 shows that the motorway entity is defined by the general OSM object type (see also Table 19 and Table 20), by the attributes inherited by the OSM Roads category (Table 21) and by the attributes that are specific to

the OSM Motorway object type (Table 22). Finally, the proper geometric type was assigned to each entity using the GML grammar (although the OSM geometry encoding is out of the scope of this Thesis). The XML Schema constructed is presented in Appendix C.

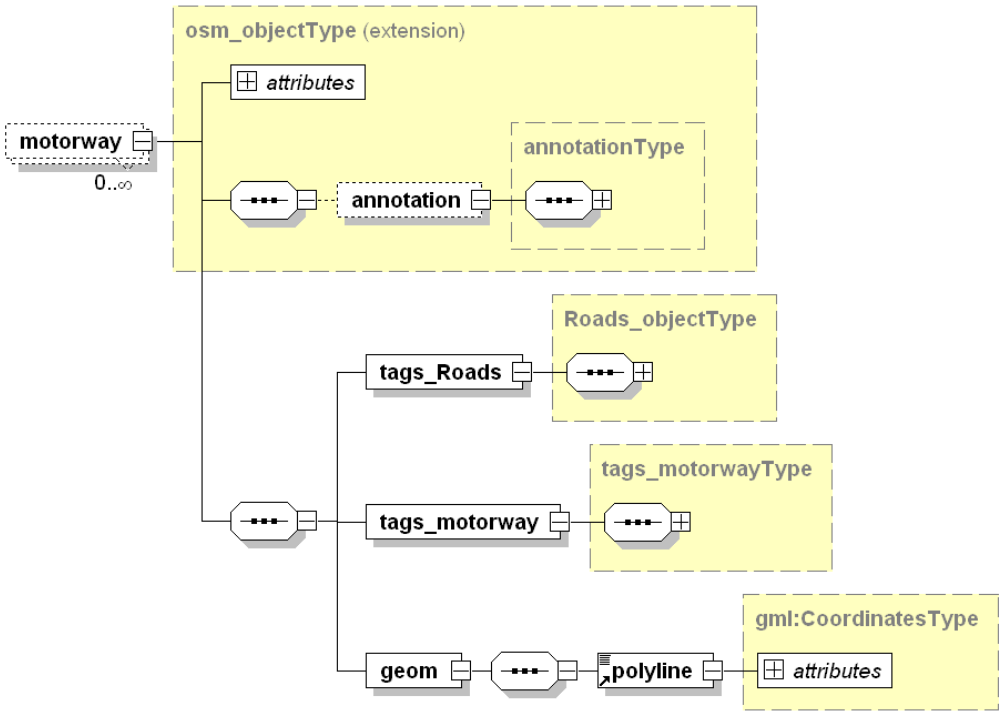


Figure 61. The diagram of the XML Schema fraction of the motorway OSM entity.

```
<xs:element name="motorway" minOccurs="0" maxOccurs="unbounded">
  <xs:complexType>
    <xs:complexContent>
      <xs:extension base="osm_objectType">
        <xs:sequence>
          <xs:element name="tags_Roads" type="Roads_objectType"/>
          <xs:element name="tags_motorway" type="tags_motorwayType"/>
          <xs:element name="geom">
                <xs:complexType>
                  <xs:sequence>
                        <xs:element ref="polyline"/>
                  </xs:sequence>
                </xs:complexType>
          </xs:element>
        </xs:sequence>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
</xs:element>
```

Table 19. The XML fragment of the OSM Motorway definition.

```
<xs:complexType name="osm_objectType">
	<xs:sequence>
		<xs:element name="annotation" type="annotationType" minOccurs="0"/>
	</xs:sequence>
	<xs:attribute name="osm_id" type="xs:integer" use="required"/>
	<xs:attribute name="changeset" type="xs:integer" use="required"/>
	<xs:attribute name="user" type="xs:string" use="required"/>
	<xs:attribute name="user_id" type="xs:integer" use="required"/>
	<xs:attribute name="version" type="xs:integer" use="required"/>
	<xs:attribute name="visible" type="xs:boolean" use="required"/>
	<xs:attribute name="rec_time" type="xs:dateTime" use="required"/>
	<xs:attribute name="comment" type="xs:string"/>
</xs:complexType>
```

Table 20. The XML fragment of the OSM Object type definition.

```
<xs:complexType name="Roads_objectType">
	<xs:sequence>
		<xs:element ref="smoothness" minOccurs="0"/>
		<xs:element ref="surface" minOccurs="0"/>
		<xs:element ref="access" minOccurs="0"/>
		<xs:element ref="traffic_calming" minOccurs="0"/>
		<xs:choice>
			<xs:element ref="tunnel" minOccurs="0"/>
			<xs:element ref="bridge" minOccurs="0"/>
		</xs:choice>
		<xs:element ref="lit" minOccurs="0"/>
		</xs:sequence>
</xs:complexType>
```

Table 21. The XML fragment of the Roads type definition.

```
<xs:complexType name="tags_motorwayType">
	<xs:sequence>
		<xs:element ref="name"/>
		<xs:element ref="ref"/>
		<xs:element ref="oneway"/>
		<xs:element ref="lanes"/>
		<xs:element ref="minspeed" minOccurs="0"/>
		<xs:element ref="maxspeed"/>
	</xs:sequence>
</xs:complexType>
```

Table 22. The XML fragment of the Motorway type definition.

Finally, as discussed earlier (Section 5.6.2) not all attributes included in the XML definition of an OSM entity need to be completed in order to sufficiently describe it.

Thus, only the system-generated and the attributes explicitly stated in the OSM wiki pages are required. The rest are optional and might not appear at all (i.e. minOccurs="0"). However, as it has been noted it is useful to know whether an attribute is not applicable to a certain spatial entity or it has not been recorded by the users.

This process can give a solid base for defining the entities included in the OSM data repository. The absence of an OSM product specification considerably hinders the understanding of the data's spatial value as it is difficult for anyone to compile the now scattered, yet necessary information.

6.2.3 Quality evaluation mechanism (proof-of-concept prototype)

In the previous Sections the need for creating product specifications for the datasets that come from UGSC sources was discussed. A mechanism to build such a specification using the grammar of the XML Schema was presented for the case of OSM. This Section will discuss a prototype application that serves as a proof of concept regarding how the sense of data quality can be engraved on the functionality of a Web 2.0 geo-application and thus provides a prototype solution to the challenge of data formalisation and quality improvement. This step builds upon the discussion about interactivity and the formalisation of the UGSC. Thus, this is not simply a technology-based, programming-oriented approach, but it is founded on top of the theoretical concepts of interactivity, spatial data quality and quality communication. A unique element of the process presented here is that it does not fragment the spatial data quality information from the content creation process as is mainly the case with OSM editors or the theoretical suggestions of Goodchild (2008c) and Grira et al. (2010) (see also Section 2.4.3). Instead, here, the merge of all involved phases (i.e. quality evaluation, communication and improvement and data creation) is implemented for the data attribution process, using OSM data as a case study.

The basic concept behind this prototype is fairly simple, yet efficient. The aim is to achieve a direct communication of the data's quality conformance level to the untrained

users, and thus to provide a clear picture of the state that the data at hand is. Furthermore, by doing so there is a direct motivation for the users to put extra effort for the improvement of the map presented to them. As explained in Literature Review, the challenge of communicating quality has been addressed using the metadata mechanism. However, as discussed, metadata is not a suitable method when it comes to lay users and even the theoretical concept of user-centric metadata might not have the desired results. Instead, a more Web 2.0 generic process is presented here. In parallel, the users' intention to correct the map should be undistracted and their possible actions should be hosted directly by the map. Therefore, the map content should have the necessary interactivity to support such user actions.

The prototype's architecture is shown in Figure 62. The current data specifications are clearly modelled in the back-end of the applications and describe the data that reside in the application's repository. The users are presented with an application that serves interactive spatial content that can be evaluated against the specifications. As explained, the specifications' change mechanism should be examined for each UGSC source separately and it is outside this Thesis' scope. However, it should be noted that an orderly evolution of the specifications enables data transition and harmonisation among specifications' versions.

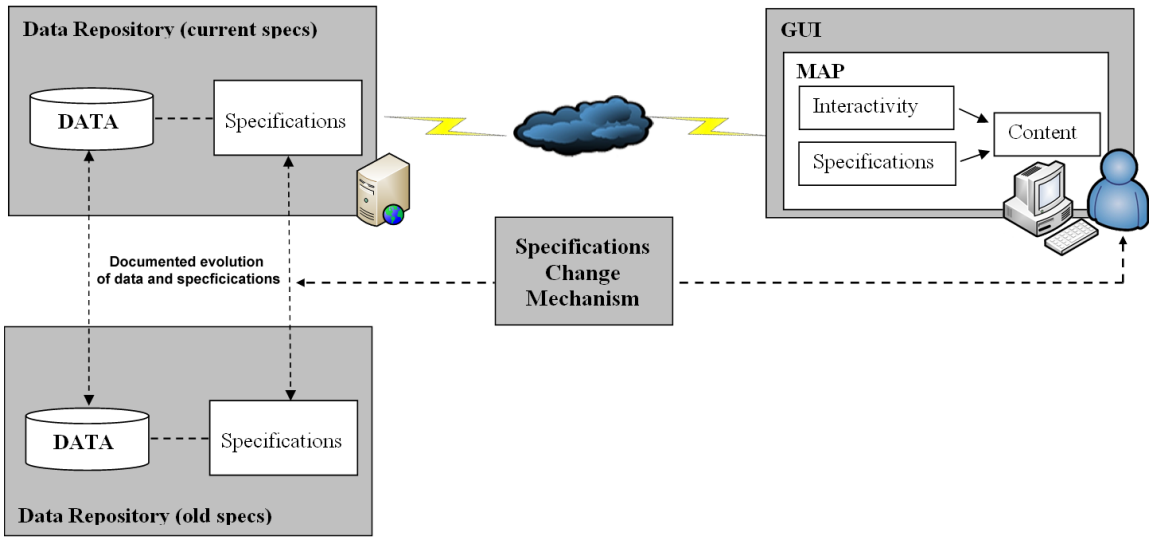


Figure 62.Quality communication and improvement mechanism for UGSC sources.

The prototype application consists of a vector map that handles points (OSM POIs), lines (OSM Highways) and polygons (OSM Buildings). Vector encoding enables spatial content to be interactive and thus each entity is associated with further information (in this case the user and the system assigned tags) and each spatial entity is responsive to users' actions. Thus far, the prototype functions as a common vector-encoded map (Figure 63).

Figure 63. The basic functionality of the prototype's Graphical User Interface (GUI).

However, the interesting point in this application is that both the map content and the tags are evaluated against the OSM XML Schema specification. More specifically, the user, upon request, can instantly see a thematic quality map (Figure 64). Each one of the spatial entities presented on the map is evaluated against a set of rules. In Figure 65 two simple rules are examined for buildings: each entity should have its *name* and *type* tags completed. The red colour is used to show that a building has none of the two attributes, the green that both attributes are completed and the orange that there is a missing element.

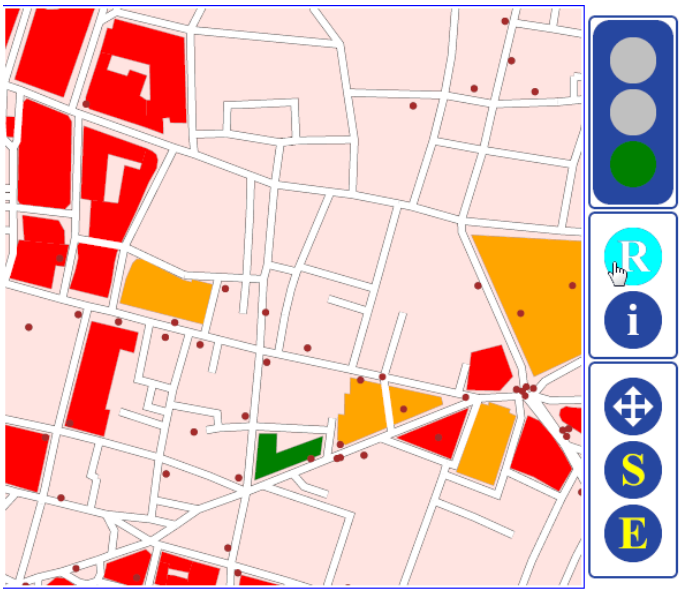


Figure 64. Quality evaluation results.

Further details of the map content's quality evaluation are presented for each entity. This is achieved by evaluating the user-assigned tags of each entity against the XML Schema specification and presenting to the users the tags and the evaluation's results. For example, Figure 65a shows the user assigned tags for an OSM building. In this case both required tags are completed properly and thus the entity is coloured green. In contrast, Figure 65b shows the tags of an orange-coloured building where the tag *type* is not completed and the user is prompted to fill it in. Finally, Figure 65c shows a selected road segment for which its tags' evaluation against the XML Schema showed two violations. The user is prompted to disambiguate the tags and thus to improve the entity's quality.

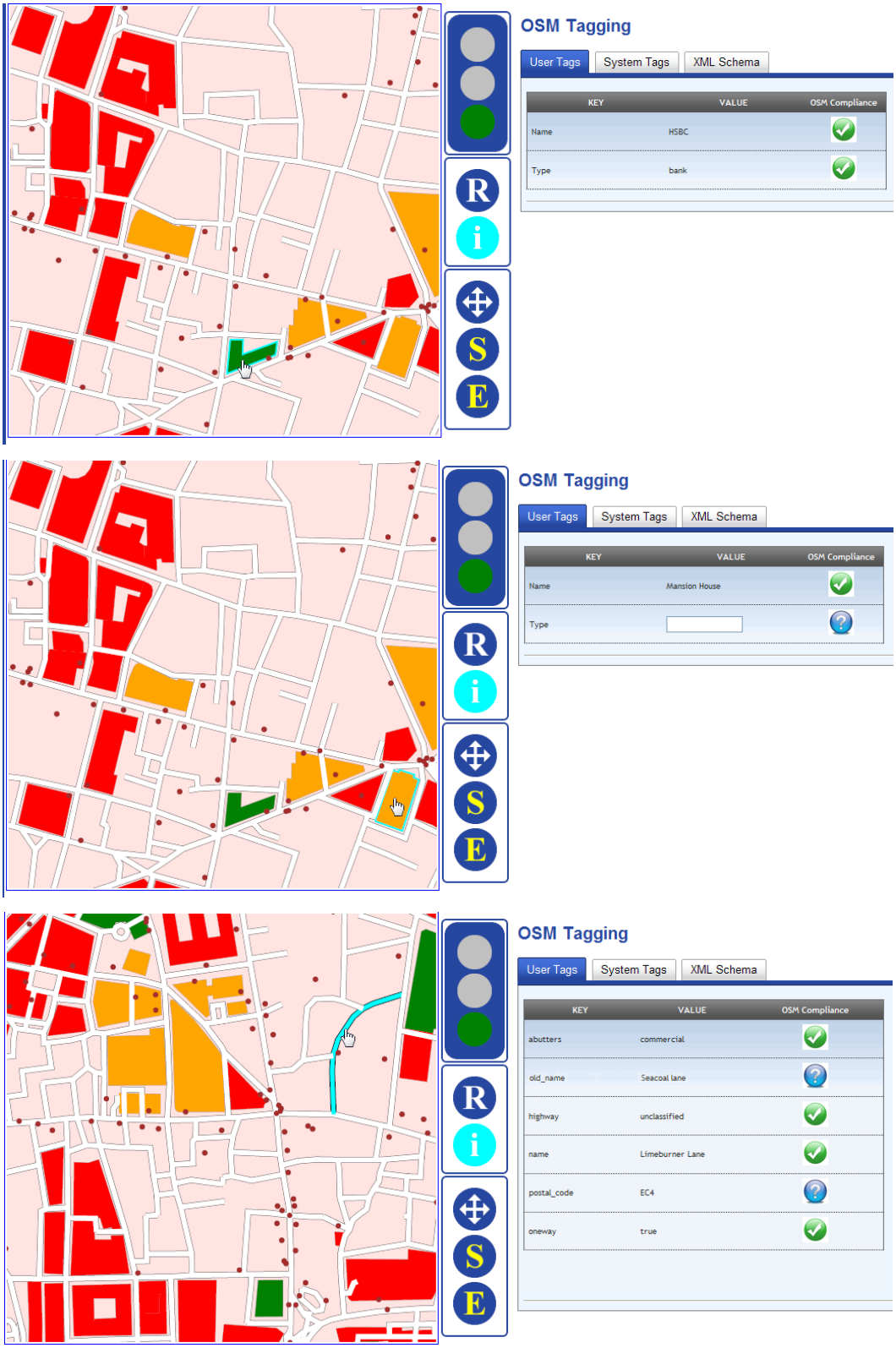


Figure 65. Screen-shots from the quality evaluation and communication functionality of the prototype.

Finally, the prototype application gives to the user the option to fill in all the attributes that fully describe the entity according to the XML Schema (Figure 66). Ideally, filling (completely or not) this form would be obligatory when creating a new spatial entity instead of spontaneously submitting tags.

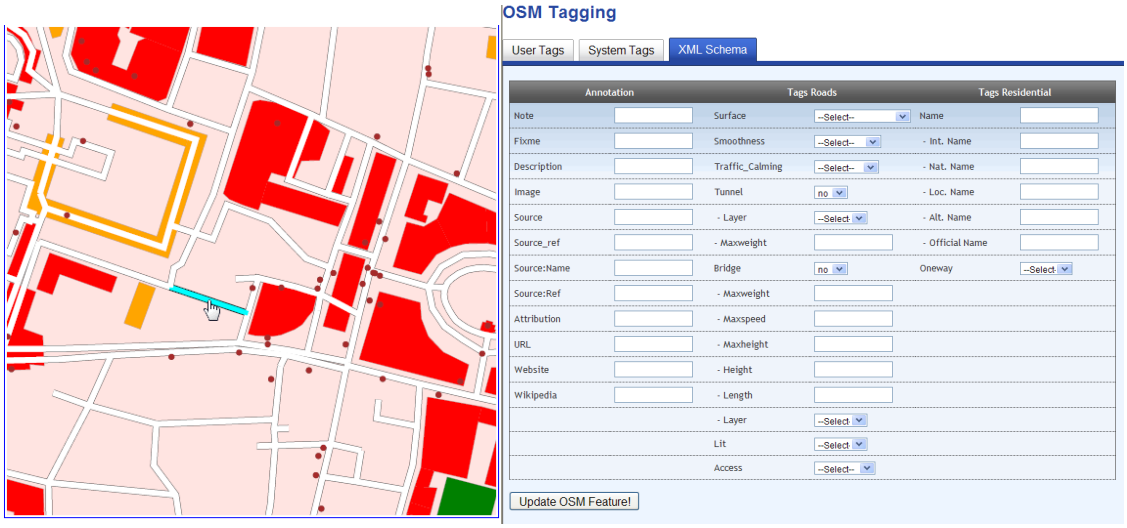


Figure 66. The completion of an entity's tags that are required by the XML Schema.

6.2.4 Discussion

As seen in Literature Review, scholars recognise that spatial data quality will be an important factor and a major challenge in the evolution of UGSC phenomenon and a number of them provided theoretical suggestions on how this challenge should be tackled. Indeed, the empirical research conducted during the course of this Thesis showed that the data producers' heterogeneity in combination with the loose co-ordination and the structural flexibility of the data, negatively affect the overall data quality. A central point of the Thesis was that UGSC sources need to establish a formalisation process for the data sought and to engraft quality-awareness to the Web 2.0 geo-applications used by the crowd to create UGSC. Moreover, it has been supported that this quality information needs to be efficiently communicated to the lay users in such a manner that will generate their reaction for improving the data presented to them. This section provided a case study of how this challenge should be met using OSM as a case study. The functionality presented in this prototype can be further expanded to the improvement of existing entities' geometry (as is the case of the OSM editors) or to the

creation of new ones (by presenting to the contributor a Schema-compliant form for the entities attribution).

The effort to meet this challenge has been based on the prototype's ability to present to the users a content-level interactive map. The importance of interactivity in Web-based geo-applications have been already discussed (Section 2.3). It is understandable that raster-only maps cannot support such functionality as they cannot provide the entity-level interactivity needed to host the quality improvement users' actions. However, it has been seen in the Literature Review that efficient vector data transmission over the Web remains a challenge. The vector data transmission method that this prototype used to overcome the vector data limitations is discussed in the next Section.

6.3 Challenges and solutions: Vector data transmission over the Web[11]

Thus far the importance of interactivity in the evolution and the acceptance of the Web 2.0 applications have been discussed (Section 1.2.4). Also, the effect of interactivity in the Web 2.0 geo-applications has been shown. In Section 2.3 the discussion focused on the existing methods for spatial data transmission (vector and raster) over the Web and the strong points as well as the limitations of each case were analysed. Based on this discussion, interactivity was recognised as a fundamentally important factor for the evolution of the Web 2.0 geo-applications and consequently a key factor for the UGSC phenomenon. Therefore, one of the Thesis' objectives was to achieve content-level interactivity for Web-based geo-applications. From the analysis so far it has been made clear that if it was possible to overcome the problem of vector data transmission, Web 2.0 geo-applications would be able to easily use vector data to model spatial entities and thus infuse interactivity in the application's content resulting in increased user participation and content generation. Moreover it has been discussed that interactivity could also be beneficial to the improvement of UGSC quality (Sections 2.4.3 and 6.2).

[11] This Section has been adapted from:
Antoniou, V., Morley, J. & Haklay, M.M., 2009a. Tiled Vectors: A Method for Vector Transmission over the Web. In J. D. Carswell, A. S. Fotheringham, and G. McArdle *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pp. 56-71.

This can be done by an interactive mechanism that first makes users aware of the UGSC underlying quality and then is able to host their contribution/reaction for data quality improvement as discussed in Section 6.2.4.

This challenge was met in the course of the Thesis. A methodology for vector data transmission over the Web that overcomes the existing problems has been developed.

6.3.1 Methodology's overview

Here, a new method for XML and text-encoded vector data transmission over the Web is presented (the methodology was tested using SVG as the map data encoding format). Instead of trying to implement a progressive transmission technique (as seen in Section 2.3.2.1) tailored to vector data needs, the method presented here follows the tile-based approach. In brief, according to the proposed methodology, asynchronous data requests are submitted to the server only if the data has not already been sent to the user, otherwise data are read from the browser's cache. When a user's request reaches the server the data are cut into tiles and then send to the user's browser. At the user's machine the tiles are merged and the final map is presented to the user (see Table 23 for the steps followed after a typical map request). This approach provides a method to transmit vector data to the client using AJAX, but it does not solve the problem of on-the-fly vector generalisation. Thus, this approach is applicable in the case of pre-prepared multi-resolution spatial databases (which is the common case for mapping agencies). It is understandable that the effective implementation of a methodology based on the client-server architecture needs the coordination of all engaged parts (spatial database, server, user's browser and the map document itself). In what follows the architecture of the methodology with an emphasis on the structure of the map document, the interaction of map document with the browser and the server and the map preparation (i.e. tile merging) are described.

6.3.2 The map document

The role of the map document's structure is central for the methodology. The map document has three different layers (Figure 67). The first layer consists of a 5x5 grid of tiles (background area). This layer is not visible to the user but is used to hold the tiled data sent to the map document either by the server or the browser's cache memory. Also, this layer provides data to the next layer of the map document. The second layer (viewable area) consists of a 3x3 grid of tiles. These tiles are cloned from the background area (first layer). Its role is to hold the data that are going to be merged and then assigned to the thematic layers at the next layer of the document. The third layer (map area) consists of one tile. This layer is the actual map requested by the user and is compiled from the thematic layers according to cartographic rules.

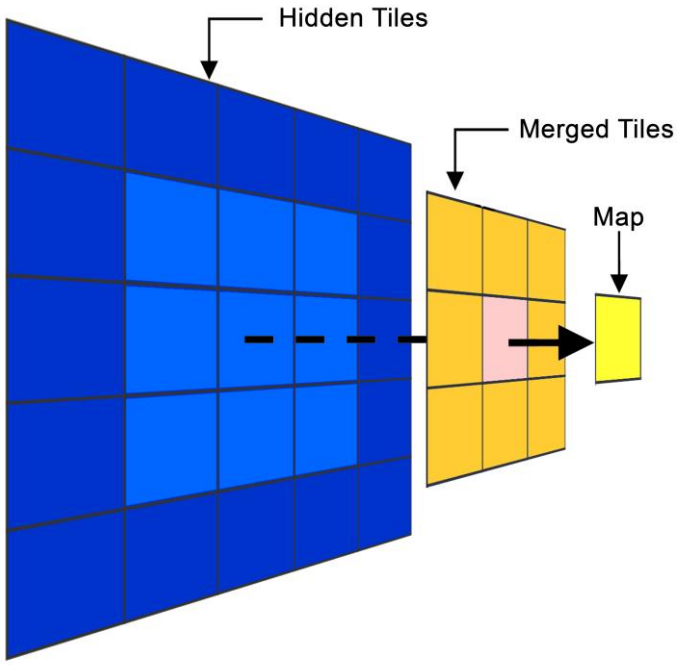


Figure 67. The structure of the map document.

The cloning of data from one part of the document to another is triggered by the user's actions. When the user requests to view (by panning the map) more data, the tiles that are already stored in the hidden area of the document are cloned to the merging area, where the merge of geometries takes place (see next paragraph), and then are assigned to the correct thematic layers which are presented to the user as the final map. At the same

time, the map document prepares itself for the next user actions by requesting new data either from the browser's cache memory or from the server.

6.3.3 Merging

An important step of the whole methodology is the merging process that takes place at the client. The geometries stored in each of the 9 tiles of the viewable area are merged before they get assigned to the correct thematic layer. This step is needed because the map entities presented to the user should be in logical accordance with the entities stored in the database. For example, if a single polygon is split into two polygons (each one stored in a different tile) during the extraction from the database, when presented to the user these polygons should be merged back into one entity. This will allow users to interact correctly with the elements of the map. Since thematic layers can hold either point, line or polygon geometry there needs to be a merging mechanism for every type of geometry.

Points: The merging of points is trivial, since the only thing needed is to clone all points from the tiles to the final thematic layers. Javascript can easily parse the Document Object Model (DOM) of XML documents and clone data from one part of the document to another.

Lines: The merging of lines is based on the use of the unique feature IDs. IDs are used as keys to search inside the 9 tiles of the viewable area. Line segment that have the same ID are grouped and then joined into single entities. Such a join may result either in a polyline feature when the segments have common points or in a multi-polyline feature when there are no common points among segments that have the same ID, depending on how the original line, stored in the database, was split into tiles (Figure 68).

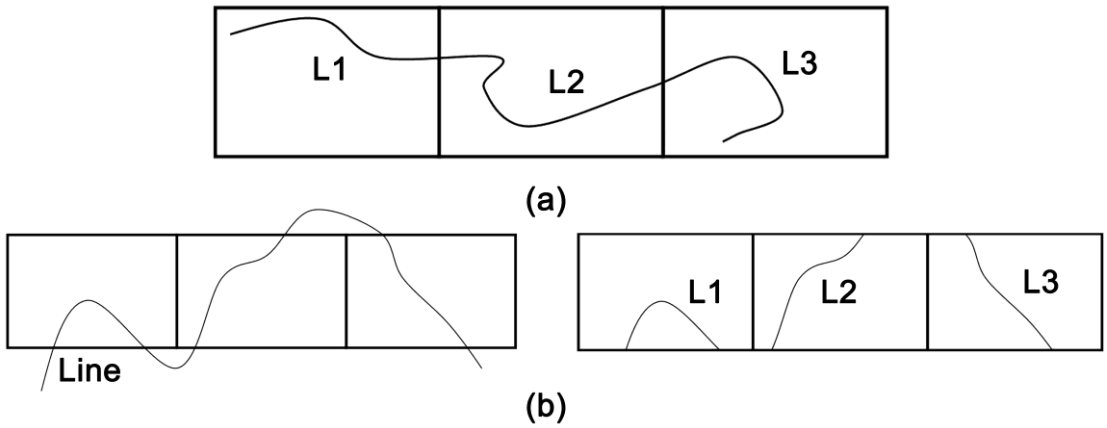


Figure 68. The merge of a line that resides in different tiles can lead to: (a) polyline element or (b) multi-polyline element.

Polygons: Unique IDs are also used for merging polygons. Once again, polygons that share the same ID are grouped and then merged either into polygon or multi-polygon entities. However, the case of polygon merging presents a greater degree of difficulty in order to disambiguate all possible cases. The main obstacle is the presence of unwanted border lines generated during the tiling process (see for example Ch. 3 of Rigaux et al. 2002 for details about spatial operations and how new line segments are generated through the tiling-intersection process). For example, Figure 69a and 69b show two different cases of a border line appearance. In Figure 69a a polygon border line generated during the tiling process is needed to achieve the correct colouring of the polygon presented to the user. In contrast, in Figure 69b that same border line is causing an unwanted visual effect. The correct way of rendering the merged parts of the polygon in 69b, is shown in Figure 69c.

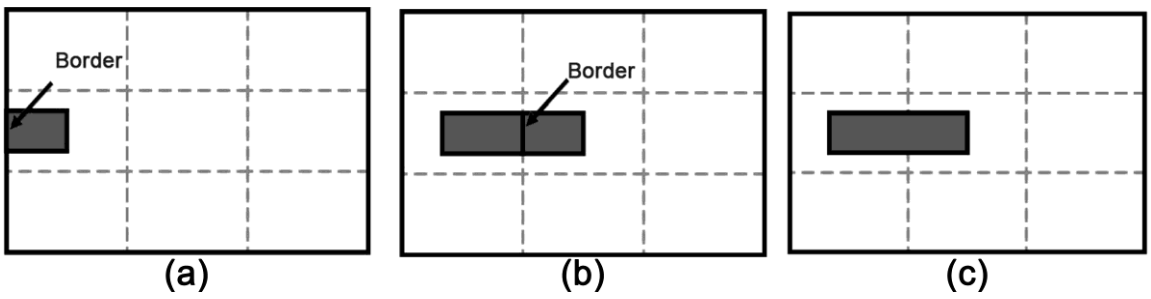


Figure 69. Dealing with border lines during the merging of polygons.

Still, the border lines generated during the tiling process are necessary during the merging phase since they help elucidate the rendering of polygons. For example, Figure

70 shows that when border lines are absent there is no indication of how a polygon should be coloured.

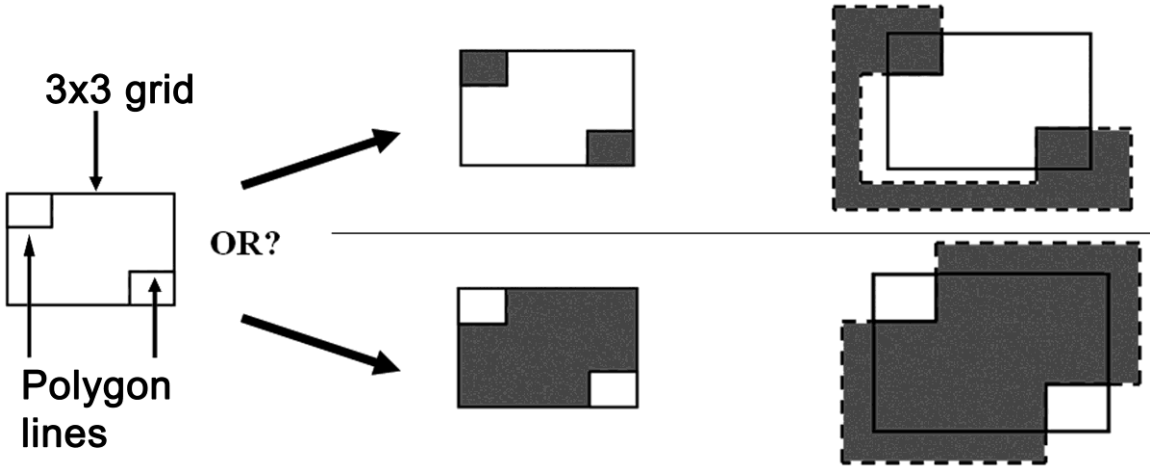


Figure 70. A vague case of polygon coloring.

To tackle that problem we need to have both the border lines of the polygons and an indication regarding when each border line should be used in the merging process. Since border lines are generated by the intersection of the polygon and the tile it is obvious that these border lines will coincide geometrically with the outline of the tile. We need to record at which side (1, 2, 3 or 4) of the tile (see Figure 71) there was a border line created. This element allows us to create a set of rules regarding the inclusion or not of the border line in the merging process. For example, border lines that coincide with the number 1 side of a tile should be included in the process only if the tile is placed in the (0, j) places of the 3x3 grid. In any other case the border line should be excluded by the process.

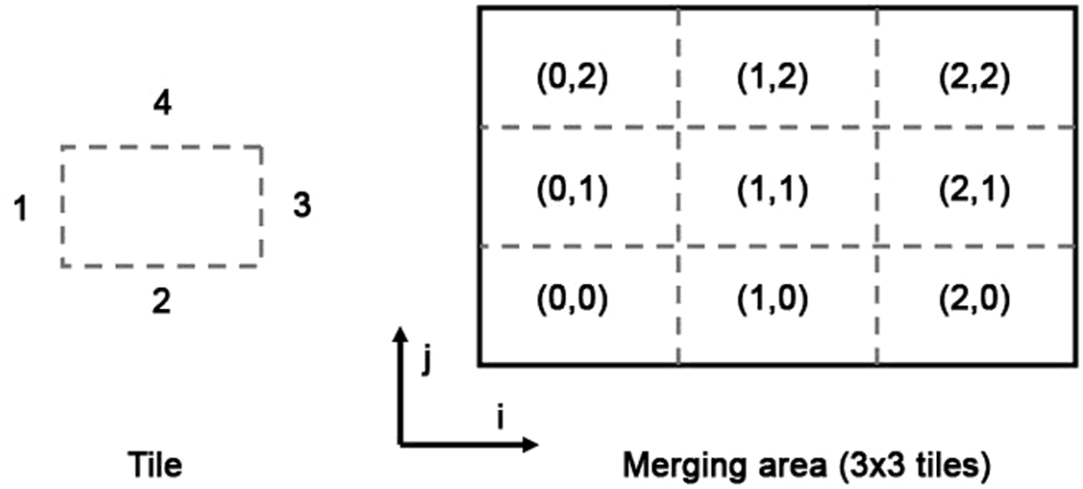
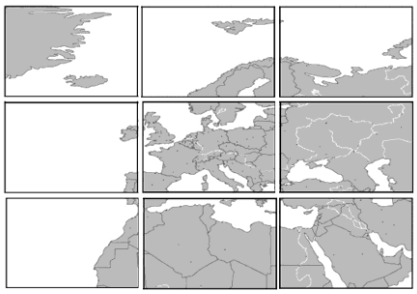


Figure 71. The method to assign an indicator to border lines.

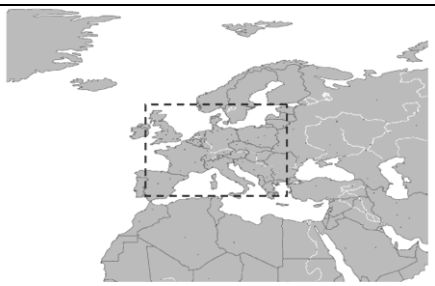
By doing so, in the case shown in Figure 69a the border line would have been used to form the outline of the polygon but in case shown in Figure 69b the border would have been excluded in the merging process. This approach has been successfully tested in all possible polygon cases like ring polygons, island polygons or multi-part polygons. Finally, Table 23 describes the steps that take place inside the map document.



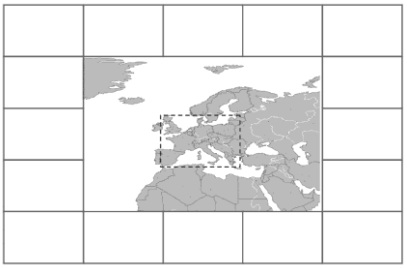
1. The method starts with a request by the user. A number of tiles is extracted from the database, covering a broader map area and send to the user's browser (the client requested only the central tile).

2. At the client's browser, for each layer, the geometry of the tiles is merged creating a continuous and correctly compiled map.

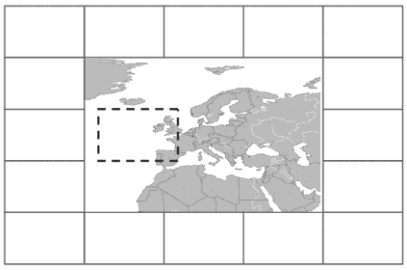
3. The user is presented with the requested map



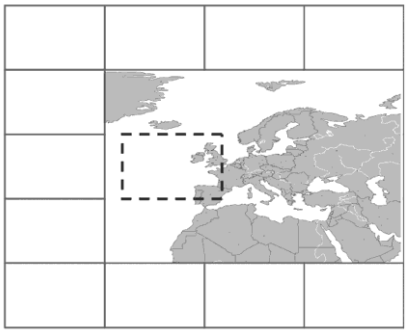
4. So, in fact, while the user sees a restricted map area (the one that originally requested), a much larger area is available to accommodate the next panning gesture without the need for further client-server communication



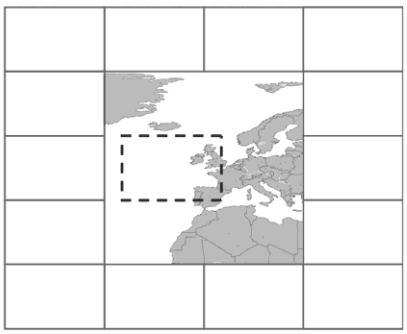
5. Additionally, another set of tiles is also extracted from the server and send to the user. These tiles are not merged but are kept in the background section of the map which is not visible to the user. These tiles work as cache memory inside the document to accommodate the next panning gestures without making the user wait for new data.



6. When the user makes a panning gesture the map document is able to respond without delay. Nevertheless, the map document starts the process to prepare itself to accommodate the next user's moves by re-arranging the tiles in the map document and by getting the missing data, either from browser's cache memory (if the needed data is already there) or from the server (see how in the next steps)



7. First the column (or row) of tiles in the background section that is on the opposite from the pan direction is dropped from the document and stored in browser's



8. The tiles that exist in the viewable area and are not needed any more are dropped, but remain in the browser's cache too.

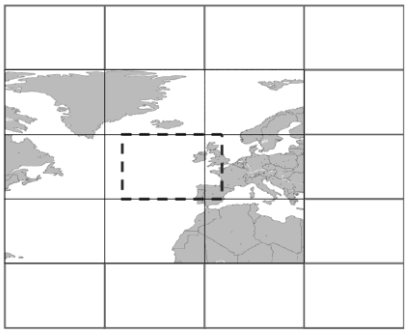
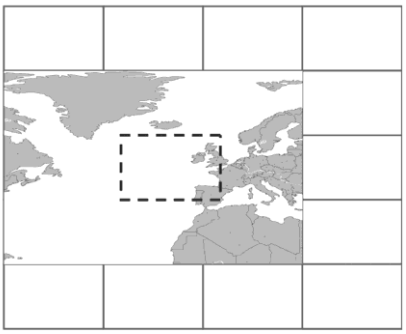
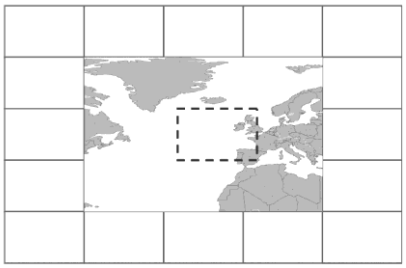
cache.	
 9. The tiles that reside in the background area and are needed to accommodate the new map are cloned to the viewable mapping area.	 10. The geometry of the tiles is once again merged for each thematic layer.
 **11.** At the same time, a new set of tiles, needed to prepare the map document, is requested from the browser. If not found there, the request is propagated to the server. These tiles will be stored in the background area of the map document and the process is ready to start all over again**.**	

Table 23. The steps of map preparation.

6.3.4 Map document's interaction with browser and server

Client side caching is a common programming technique. This technique allows browsers to hold locally for later use data that have been sent from the server without any further interaction. Thus, reduced network latency and enhanced user experience is achieved. In this case, browser's cache memory is used to hold tiles of data sent from the server to the user but not any more stored inside the map document. Thus, whenever new tiles are required from the map document, the search starts in the browser's cache. If the tiles needed are stored in the cache memory, then the data is inserted into the map document. If the data has not been already sent to the user, then the request for new tiles is propagated from the client to the server.

This client-server interaction takes place asynchronously using AJAX requests. In this case the AJAX methodology is used to manage the client-server communication behind

the scenes. Given the fact that vector tiles have fixed dimensions (defined by the map window of the application and the scale), the construction of spatial queries to retrieve the missing tiles is a straightforward process.

Figure 72 shows the time periods involved in the whole process. Period A is the mouse movement; using AJAX a new map request can be triggered while the user is still panning in a direction. Period B is the time that the user observes the map. The tiles method coupled with AJAX requests provides data constantly to the user and thus there is no need for the user to wait for the map to be downloaded. Period C is the time the server needs to extract the data. The key point of the method is that it fully exploits the time period of user inactivity (i.e. does not interact with the map requesting new data). While the user observes or queries the map entities available, behind the scenes there is work in progress to prepare the map document to accommodate the next user's moves. If time period B is very small (i.e. the user is quickly panning towards the same direction) the process of requesting new data from the server can be suspended with the help of a time counter. This process is implemented only for those requests that cannot be accommodated by data stored in browser's cache.

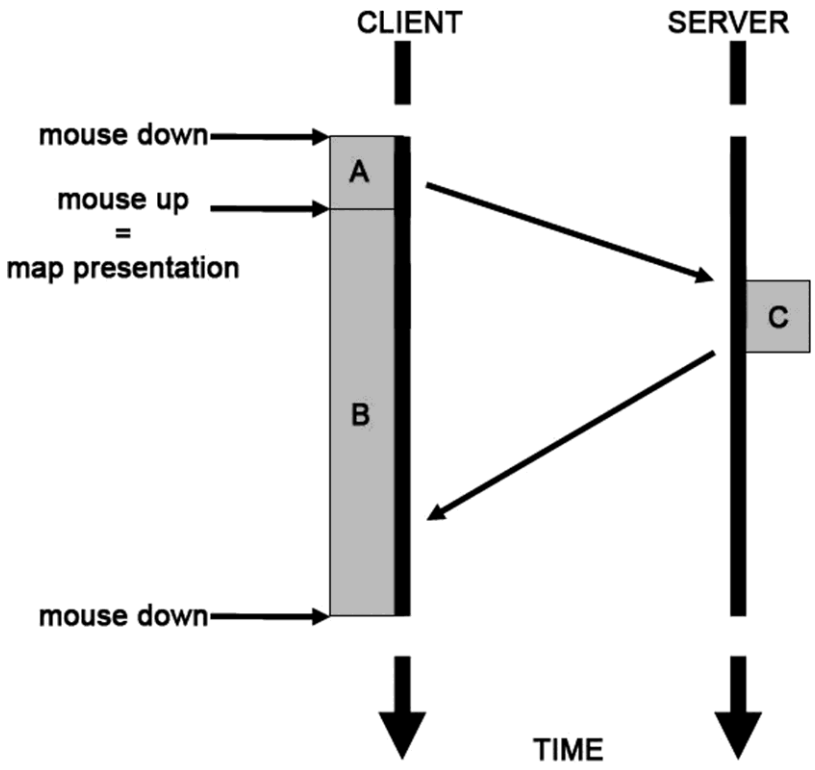


Figure 72. Steps and time periods in tile-based transmission of vector data.

Figure 73 shows a screenshot of a prototype built to test and improve the algorithms developed. The SVG was used as the format to build the map because it is an XML based format that supports scripting. Any practices and methods applied during the implementation can be implemented in other XML-based or text-based vector formats.

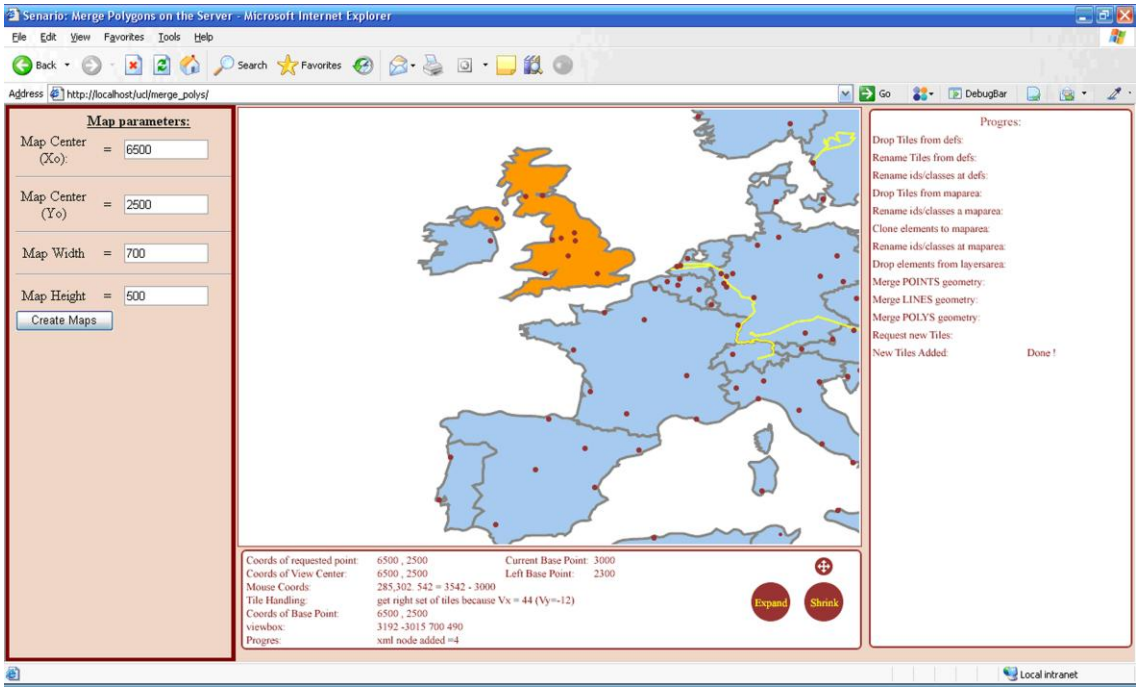


Figure 73. A screenshot of the prototype.

6.3.5 Performance

Usually the performance of progressive transmission methods for vector data are tested by examining the time needed to transmit a dataset of a certain size (for example see Yang 2005, Yang et. al 2007, Bertolotto 2007). This type of performance evaluation is not applicable in this case since the method aims to the exact opposite target: to avoid the transmission of bulk sets of data but instead to split data in small packets and exploit the real-life user behaviour to efficiently transmit vector data. Therefore, for the performance evaluation of this method a more user-oriented and thus more suitable method was used. The time periods for panning (time period A in Figure 72) and map observation (time period B in Figure 72) that will allow this method to present to the user a ready to use map immediately (i.e. not fail to fulfil the condition $t_{mouse\ up} = t_{map\ presentation}$) were examined.

The experiments were performed over the Web using a server with 2.13 GHz CPU double-core processor and 1 GB of memory stationed in Athens (Greece), connected to the Internet with nominal value for download 24Mb/s and for uploading 1Mb/s (the actual average values recorded during the experiments were 3.1Mb/s and 280Kb/s respectively) and a client with 1.66 GHz CPU double-core processor and 1 GB memory stationed in London (UK), connected to the Internet with 8Mb/s download and 1Mb/s upload nominal speed values (the actual average values recorded during the experiments were 3.8Mb/s and 682.9Kb/s respectively). The average ping time between client and server during the experiments was 112 ms.

For the evaluation two different sets of data were used: a world level dataset that contains countries, rivers and cities and one at a street level that contains parcels, parcels' registration points and street centrelines. Table 24 shows the analysis both of the entire datasets stored in a spatial database and of the data accessed by the user during the experiments.

World Level Dataset

Geometry Type	Entire Dataset			Data Accessed by the user		
	Num. of Entities	Num. of Parts	Num. of Points	Num. of Entities	Num. of Parts	Num. of Points
Polygons	147	249	28,162	140	228	24,983
Polylines	98	177	3,965	95	147	3,852
Points	606	606	606	550	550	550
		Total:	**32,733**		**Total:**	**29,385**

Street Level Dataset

Geometry Type	Entire Dataset			Data Accessed by the user		
	Num. of Entities	Num. of Parts	Num. of Points	Num. of Entities	Num. of Parts	Num. of Points
Polygons	5,879	5,879	37,213	2,360	2,360	15,122
Polylines	15,186	15,186	32,128	5,941	5,941	12,535
Points	6,245	6,245	6,245	2,447	2,447	2,447
		Total:	**75,586**		**Total:**	**30,104**

Table 24. The analysis of the datasets used and accessed during the experiments.

The performance of this method was exmined during the worst case scenario: the user was panning constantly in the same direction and thus no help from the browser's cache was provided. Instead, every user's request was propagated to the server so as to retrieve data not yet sent to the client. The time for the panning gestures was recorded as well as the minimum time that the user had to observe the requested map until the map document prepares itself for the next panning gesture without introducing further delays.

In every step of the experiments the user was presented with a correctly composed and fully functional map. This means that all entities were styled and assigned to the correct thematic layers which were used to compose the map. Additionally, a link was assigned to every entity presented to the user so to enable further AJAX queries (for example, request the attributes of an entity from the server). Table 25 shows the time needed for the presentation of the map after the first request and the minimum, maximum and average map observation times so the user to be constantly presented with such a map.

	First map presentation (sec)	**Average Pan Time (sec)**	**Min - Max - Average Observation Time (sec)**
World Level	7.8	0.7	0.4 - 3.5 - 1.3
Street Level	18.1	0.7	0.6 - 5.5 - 2.8

Table 25. Performance results of the proposed method.

As expected, the slowest part of the method is the time needed for the presentation of the map after the first request. Subsequent map presentations though, need considerably less time which varies due to differences in the volume of data transmitted and the number of entities that need to be merged each time. As explained earlier, in a real life scenario (i.e. the user does not pan only in one direction) the observation time is expected to be further reduced by using data stored in browser's cache memory.

6.3.6 Discussion

The proposed method provides a smooth user interaction with the map, overcoming the problem of long waits for the download of vector data. The efficiency of the method depends on the correct coordination and refinement of all the method's steps. So, in addition to what has been discussed so far, a new database structure can be developed that will hold directly pre-calculated tiles instead of extracting them at the time of request, based on global gridding similar to existing tiled raster services. This will considerably improve server performance and server-side caching.

The key point of the proposed method is that the map document should always have the necessary data to accommodate the next user's move; the tiled vector data must reach the browser before the user requests them in order to eliminate waiting times. In other words, the performance of the method is in close relationship with the user's behaviour as well as the efficiency of the method itself. The bigger the time periods that the user remains inactive in terms of data request, the more time is available for the preparation of the map document for the next request. In contrast with what takes place in the case of raster tiles, this preparation time varies in the vector case. Raster tiles have a standard dimension which also results in having almost a fixed size. In contrast, there is no guarantee of the size of a tile that holds vector data and this can introduce delays in the process. A solution to that could be tiles of variable dimension (i.e. different spatial coverage) that will hold data that has size below a chosen threshold. In this case specialised algorithms for tile indexing and manipulation are needed. Alternatively, progressive transmission techniques could be implemented for the data that each tile holds.

The proposed method has a number of advantages compared to the transmission methods available today. Following this strategy, a continuous vector map will be available to the user by sending only small pieces of data to achieve reduced network latency. The procedure does not need to interact with the server, and thus introduce network latency, for small panning gestures since the client will have enough data to accommodate such user actions. Additionally, when needed, the use of client caching and AJAX will avoid

the unnecessary client-server transactions but will still allow the missing tiles to be embodied in the map behind the scene without the user noticing it.

By testing this method over the Web it was proven that it can be implemented in real life applications. The approach builds upon the architecture that most mapping agencies use: multi-scale databases. While the on-the-fly generalisation problem remains unsolved, map providers serve maps from different LoDs on the Web. Exploiting that fact and in contrast with the existing limitations of progressive transmission techniques, this method disturbs neither the geometry nor the topology of the features presented on the map. Moreover, it has no restrictions in handling multi layered requests of any geometry and in applying all cartographic principles during map composition. Also, the method can be implemented to any XML or text-encoded format like KML or GeoJSON. It is interesting to note that, after the first map presentation, this method offers an efficient way so users can access unlimited volume of vector data with waiting/observation times that do not differ from the typical waiting/observation times that occur when they browse any other Web application. Finally, this tiling approach makes more difficult to compromise IPRs of vector data compared with methods that send the whole dataset to the client, progressively or not, as only tiles of small parts of data are sent any given time to the client's machine, and thus any attempt to reconstruct of the whole dataset would be a very difficult, if not impossimble, process. The compromise of IPRs has been a major factor that made developers and mapping agencies reluctant to publish vector data on the Web. While in the first step of the method more data than requested are sent, this disadvantage is balanced by reduced client-server interaction and network latency during the subsequent steps.

The combination of the evolution in the Web and the efficient mechanisms for raster delivery on the one hand and the need for vector data for enhanced interactivity and object manipulation on the client side on the other form a new environment for Web mapping. In such an environment the role of hybrid Web maps able to host both raster and vector data will be considerably increased (Antoniou et al. 2008). Such maps will use only the strong points of raster and vector data and is likely to be the most efficient way to deliver spatial data for complex Web mapping applications. Finally, the content-level

interactivity achieved can be the catalyst for increased user participation and content generation in Web 2.0 geo-applications.

6.4 Summary

In this Chapter the results yielded by the empirical analysis have been used as a stepping stone for a more in depth examination of two of the most important challenges regarding the development of UGSC sources.

The first challenge was the inclusion of the quality principles in the functionality of a Web 2.0 geo-application. The development of quality-intelligent geo-applications might be the way forward for the most important issue that the UGSC phenomenon faces. Here, a case study of how the quality awareness can be infused to a UGSC source has been presented. The second challenge was the enrichment of Web 2.0 geo-applications with content-level interactivity. Interactivity could be a multiplier for both content generation and quality improvement for Web 2.0 geo-applications. However, inefficient methods of vector data transmission hindered the expansions of such type of geo-application in Web 2.0. Here, a new method that overcomes known problems has been presented.

These challenges should be closely examined by both entrepreneurs and mapping agencies when designing/planning to build such applications as they affect both the functionality and the output of the UGSC sources. A second round of issues that need to be thoroughly examined follows in the next Chapter where a detailed discussion on the findings of the empirical research is presented.

Chapter 7

Discussion

7. Discussion

7.1 General

This Chapter will start with the discussion of the research's findings regarding the Web 2.0 sources of geo-tagged photos and the OSM. The aim is to focus on the important elements revealed from this analysis. The discussion will concentrate on the characteristics and the particularities of each Web 2.0 source type, on the lessons learned and on issues that might prove as sources of concern for the future evolution of UGSC.

In the second part of this Chapter the discussion's scope will be broader. The experience gained from the analysis will be used to place the UGSC sources in a more general, yet still challenging, background in the Web 2.0 world. More specifically, it will be explained what type of Web 2.0 geo-applications should be built to achieve the type of evolution that scholars and entrepreneurs in Geomatics aspired in the dawn of UGSC and what their fundamental characteristics should be. Then, the discussion will move to a broad examination of the ergonomics behind the acceptance of the UGSC from the Geomatics world and particularly from mapping agencies since UGSC is a phenomenon that has little or no credentials and it is difficult for institutional organisations to engage with it.

This Chapter, along with the presentation, analysis and the solutions provided in Chapter 6, builds a solid knowledge base that can be useful when engaging with UGSC and for future research on the subject.

7.2 Discussion on the geo-tagged photos analysis' results[12]

In Chapter 2 the enthusiasm and optimism that surround UGSC and the GeoWeb's evolution was briefly discussed. Scholars and entrepreneurs alike have supported that this phenomenon could radically change our conceptualisation of GI on the Web and the nature of mapping products available online.

Indeed this enthusiasm and optimism do not appear to be baseless. It could very easily stem from the volume of UGSC on the Web. The magnitude of the users' activity on photo-sharing websites is an impressive one. Recording user participation from just four Web sources provided over 4.4 million geo-tagged photos for Great Britain alone. This number indicates that people upload on the Web massive quantities of geographic information. Furthermore, it should be kept in mind that this huge user participation and content generation takes place with technological means that constantly evolve. So for example, data capturing devices such as GPS-enabled mobile phones and photo cameras are getting both more sophisticated and cheaper and thus more productive and more accessible to lay users (although the digital divide remains an issue). Thus, it stands to reason to support that the ubiquitous presence of constantly improving data capturing devices will lead to increased flow of UGSC. At the same time the increasing popularity of the social media will drive much of that UGSC to the repositories of geo-applications similar to the ones examined in this Thesis.

Another strong positive indicator regarding the potentials of the phenomenon can be based on the spontaneous behaviour of the users. As seen in Figure 33 (Section 4.6) more than 50% of the photos were captured and posted online on the same day and this figure jumps to approximately 80% for the photos posted in less than 6 months. Such up-to-dateness is unprecedented for most mapping agencies that follow the classical methods of spatial data collection. Thus, the currency of the data available could prove a major tool for the evolution of the Web mapping applications. Providing and consuming

[12] Parts of this Section have been adapted from:
Antoniou, V., Haklay, M. and Morley, J. 2010b. Web 2.0 Geotagged Photos: Assessing the Spatial Dimension of the Phenomenon. *Geomatica (Special issue on VGI)*, 64(1), 99-110

instantly data is not a strange concept: Atom, Really Simple Syndication (RSS) feeds or even Twitter are fairly established ways of information sharing on the Web but not yet for mapping applications. This UGSC's characteristic becomes more important if it is jointly considered with the findings regarding the familiarity of the users with the area in scope. It has been shown that about 40% of the users in explicit sources have a persisting relationship with the area captured as their activity in the area spans for more than 2 weeks. As seen in the literature review, familiarity with space and local knowledge is a valuable factor that could reveal aspects of the human activity that have been yet uncharted by the contemporary mapping products. However, the flip side of the latter argument is that 60% of the users for explicit sources and more than 70% for implicit ones are just visiting the area as they remain active for less than 2 weeks. This is the first of a series of issues that suggest a more moderate stance against the potential of UGSC geo-tagged photos as a robust source of information.

Perhaps the more significant issue of all is the type of the Web 2.0 source. Indeed, the analysis showed that when it comes to the evaluation of the sources' overall value a core issue is the type of the Web 2.0 geo-application. The findings painted a clear picture: the photo-sharing Web 2.0 geo-applications can be grouped into spatially explicit and spatially implicit ones. In the first group GI is the main characteristic of the application. Spatially explicit geo-applications urge their users to collect spatial information that comply with some sort of loosely defined specifications (e.g. "…submit a photo of a place…", "…cover every square kilometre…" etc.). This results in creating a dataset that is richer in GI and better positioned in space than the implicit sources. In implicit sources GI is not prioritised over the rest of the features offered by the Web application and thus GI is firmly connected with social behaviour patterns that lead in a clustered distribution of data around popular locations (both highly populated and attractive for social gathering, such as tourism attractions).

More specifically, despite the fact that in terms of spatial distribution the spatially explicit source of Geograph had the third largest data pool among the four sources, it provided considerably better coverage of the study area. In contrast, the spatially implicit sources of Flickr and Picasa Web failed to cover the study area with the exception of urban areas and tourism attractions. This is one of the primary differences between

implicit and explicit Web 2.0 applications. For Great Britain the space not covered by the implicit sources ranges approximately from 73% up to 85%. In other words for the overwhelming majority of the areas the social networking, photo-sharing Web 2.0 applications is like they never existed and this trend is not expected to change radically as seen in Section 4.7. In contrast, the popular and tourism areas are repeatedly covered by literary thousands of photos. A different picture is painted when observing the behaviour of the spatially explicit source of Geograph. Given the fact that the uncovered areas are less than 10%, it is obvious that many obscure areas have found their way in the UGSC. Interestingly, a similar observation of the differences between implicit and explicit sources can be made when looking the phenomenon in a larger scale. The analysis showed clearly that even in the areas where spatial implicit sources have a strong presence, there are still popular and unpopular sub-areas as their distribution is clustered around few popular spots. In contrast, Geograph's photo distribution is more scattered, thus covering the area more adequately even with considerably fewer photos. Thus, the differences revealed between the spatially implicit and explicit UGSC sources when the entire area of scope was examined were further intensified when the examination moved to a larger scale for 15 test areas.

An interesting approach here would be if these two types of sources were considered complimentary instead of antagonistic. In other words, taking into account the spatial distribution and the data flow of implicit and explicit sources, it can be suggested that explicit sources have the means to distribute data collection at a national level but they lack the power generated by the participation recorded and the UGSC volume generated in implicit ones. This argument gains further support when the popular areas of the two types of Web sources are examined. As seen in Section 4.4.1 (Figure 27) the sense of "popular area" is considerably different between the users of implicit and explicit sources. However, a counterargument would be that even when combining the popular areas of both types of sources, still a relatively small area is covered. This is one of the key points revealed in the course of this Thesis that mapping agencies should consider before deciding to engage with UGSC in the form of geo-tagged photos. Other important issues include the spatial usability of the data and the effect that social factors have on the creation of UGSC.

Regarding the former issue, it has been shown that the spatial distribution pattern for the implicit sources is not considerably changing in time. Thus, the discussion can move to the pursuit of the actual geographic information that can be extracted from implicit sources as we know them today. This is to say that further research is needed to develop efficient tools for GI retrieval suitable to surface the spatial usability of the geo-tagged photos. Although there are already early efforts to introduce geo-tagged photos into mainstream and commercial GIS (see for example Woolford 2008), it should be anticipated that the inequalities recorded in spatial coverage will affect the results of such efforts.

Regarding the latter issue, it would be useful to examine the subject from a broader point of view. As noted, one of the pillars of the Web 2.0 evolution is the Long Tail (O'Reilly 2005; Anderson 2004, 2006) and indeed the architecture of Web applications like Flickr is heavily based on that principle. The ability of any user to add any content about any subject, popular or not, is the cornerstone that enables the support of users' interest for small niches. That cumulative interest transforms these niches into important elements of the application. This research showed that in spatial terms, the Long Tail principle is not realised in the spatial distribution of the photos. Spatially speaking, the users are not interested in the small, relatively unpopular, niches of space but focus on the mainstream places. This is in contrast with early observations about UGSC regarding the ability of the phenomenon to record local places and activities that would be otherwise uncovered (Goodchild 2007a). This inability of implicit sources to cover space on the one hand and the magnification of the phenomenon in popular only areas on the other can also be corroborated from the examination of the expectation surfaces. It has been shown that implicit sources underperform in the majority of the urban areas when compared to the underlying population.

Finally, an equally important social factor is the similarity of the users' behaviour for both types of Web sources (see for example photo up-to-date-ness in Figure 32 and users' activity in Figure 33). Thus an interesting observation is that while the fundamental users' behaviour is the same for both sources, their behaviour regarding space is different. The crucial issue that makes the Web applications differ, is how related to space are the incentives given by the Web application to the users so as to

guide their participation towards a spatially-oriented direction. Thus, users need spatial explicit applications to provide distributed UGSC. A clear example of this necessity can be found in Panoramio's case. In contrast with the initial classification that considered Panoramio as an explicit Web source due to its apparent relationship with space, the analysis showed that Panoramio behaves as a spatially implicit source even if the aim is to submit photos of places. Apparently, the motivation to publish photos of the users' favourite places is not enough to provide extensive spatial coverage for Great Britain.

As the spatially explicit sources appear to be an important element for a possible engagement of mapping agencies with UGSC, a more detailed discussion on the subject will be presented in Section 7.4.

7.3 Discussion on the OSM analysis' results[13]

The next phase of the analysis was the examination of the OSM datasets. Initially, a basic analysis of Highways and POIs for England was conducted. This step helped to surface three important issues for OSM. The first one had to do with the data volume generated. Collecting, classifying and monitoring the evolution of OSM data for two consecutive quarters gave valuable insights of the intensity of users' productivity and their commitment to the OSM project. The second issue is the OSM datasets' positional accuracy. The results from the preliminary analysis showed that a fairly large number of OSM entities undergo positional alterations that probably do not correspond to real changes on the ground. This observation intensified the need to investigate the positional accuracy of UGSC. Finally, the issues of the attribution processes and attributes' quality emerged. By simply examining the Highways' name attribute it was made clear that the consistency and completeness of the attributes is a challenging issue for the OSM project.

[13] Parts of this Section have been adapted from the author's contribution to the following paper:

Haklay, M., Basiouka, S., Antoniou, V., Ather, A., 2010. How Many Volunteers Does It Take To Map An Area Well? The validity of Linus' law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4), pp. 315-322

Examining this first round of observations from a mapping agency's point of view it can be supported that the data volume produced is enticing enough for a mapping agency to pay a closer attention and further analyse the OSM project. Interestingly, the other two factors (i.e. positional accuracy and attribution) and their impact on the overall data quality, justify a closer examination of UGSC. In other words, mapping agencies need to clarify whether the UGSC is suitable to be used in their mapping procedures or the errors and thus the uncertainty that such datasets carry are prohibiting their use from an institutional point of view. The rest of Chapter 5 was devoted to shed light on this question by evaluating the positional and attribute quality of OSM.

The next stage of the research regarding OSM focused on the examination of the positional accuracy. An average positional accuracy of 7.9m is encouraging for further uses of such content. Particularly from the point of view of a mapping agency, there are many mapping products that could be based or supported by datasets of such positional accuracy. For example, a common working scale for many NMAs is the 1:50.000. This means that the spatial entities have a positional error less than 12.5m (based on the rule of ¼ mm of the scale). Combining this with the findings of the analysis it can be seen that approximately 78% of the OSM Highways are inside that threshold. Smaller scale mapping products can easily be based on OSM data. However, as seen in Figure 45 and discussed in Section 5.4.5, the OSM positional accuracy when examined in a national level cannot be treated as a homogeneous phenomenon. On the contrary, the segmentation of the area under examination is evident as there are clear formations of areas with high and low positional accuracy. For example, the percentage of the OSM spatial entities that have a positional error less than 12.5m jumps to 94% for the Greater London area.

Another important aspect of the OSM positional accuracy is its correlation with factors that have social origin. The first factor examined was the users' participation. By combining the positional accuracy and users' participation data it was made clear that there is a close relationship between the number of users that actively participate in spatial content generation and the positional accuracy of that content in a certain area (Section 5.4.6). The second factor examined was social deprivation. It has been proven by previous research that the spatial quality element of completeness (particularly for

OSM and possibly for other UGSC sources as well) is directly affected by this factor. In this Thesis it was shown that positional accuracy is affected in a very similar way. Poor and generally deprived areas are covered by less accurate OSM data compared to affluent ones. An interesting observation emerges from these two experiments. In the first experiment regarding user participation it was shown that the Linus' Law ("given enough eyeballs, all bugs are shallow") applies also to OSM data. In the second experiment the quality of the 'eyeballs' was examined. It has been shown that in order an area to be completely and accurately covered does not only need more 'eyeballs' to look at the problem but also 'eyeballs' that are not deterred by an areas' reputation and are willing to look through the curtain that socio-economic barriers draw.

Thus it has been made clear that the heterogeneity in positional accuracy is an intrinsic OSM characteristic. It has been shown that the OSM overall quality is affected by a mixture of spatial and social factors. Apparently the new breed of spatial data apart from special is also social. This realisation should be examined by the OSM community and mapping agencies alike with a due share of reflection. If proper attention is not paid, a two speed OSM data is likely to emerge.

The final stage of the research focused on the OSM attribution process. This phase of the research was split into two parts. The first part looked into the existing situation regarding the OSM attribution. The first observation is that despite the huge volume of user assigned tags the average number of tags per spatial feature is very small (Section 5.5.1, Figure 50). Another interesting point is that there is a very large diversification of the unique tags recorded for each OSM Highway category. Thus, the high recorded levels of tags' conformity against the OSM conceptual schema (Table 16) can be attributed mainly to the fact that OSM users are submitting only a few main tags for each entity and thus they stay inline with the community guidelines. It is reasonable to expect that diversification will increase as the average number of tags per entity increases or as more spatial entities are added to the data repository. Therefore, an apparent problem is to recognise which tags are truly necessary to describe a spatial entity and separate those from the noise created by the rest of tags' population. Interestingly, this is a repetition of the Long Tail phenomenon but here the focus is on the head whereas the tail contains the noise.

These observations regarding the user-generated tags helped to realise an intrinsic erroneous process adopted by the OSM community. While the effort was to establish an open and free of restrictions Web-based geo-application for the OSM community the outcome (i.e. the UGSC) suffers from a series of limitations that threaten to deteriorate the overall data quality. Furthermore as the tag generation is based on a set of good practice guidelines, issues are raised regarding the level of adoption of these commonly agreed OSM rules.

What these arguments show is that after the first wave of truly amazing achievements (i.e. the impressive volume of UGSC) the OSM project has started to display signs of malfunction. For example, what is still largely unclear regarding the attribution process adopted is what the final goal is. What will be the final form of the OSM datasets after some years given the current situation? Probably each OSM category will have hundreds or even thousands of different tag values to describe the same spatial entity. If this will be the case, how easy will be such datasets to be delivered to data consumers (e.g. mapping agencies) for further use? OSM community has started to understand these limitations and although it has not withdrawn the over-optimistic phrase of not posing any content restrictions on tags, in fact the major OSM editors have started to implement tag domains in may cases, thus forcing users to select among predefined tags.

The observations regarding the erroneous OSM attribution process were also corroborated in the second part of this phase which focused on the tags' quality evaluation. As mentioned earlier, the physical spatial entities are not clearly defined in the OSM world as there is no firm conceptualisation. Instead there are numerous wiki pages that include scattered guidelines. This part of the research has shown that the lack of conceptualisation regarding the physical spatial entities in the back-end of OSM, resulted in poor formalisation during the data generation process by the users in the front-end that negatively affects the overall data quality. Moreover, it has been shown that even changes in the administrative level can possibly deteriorate the quality of OSM data. In that context what remained largely elusive was a tangible proof of the OSM attributes' quality. To achieve that, there was a clear need for a product specification. The effort to combine the available OSM guidelines created the closest available thing to

an OSM product specification. The conformance of OSM data against the wiki-based guidelines was found to vary from 77.56% up to 87.29% for the examined categories. For a wiki-based, crowd-sourced and loosely co-ordinated project, such level of conformance is a great achievement and encourages mapping agencies to further engage with the phenomenon. However, the percentages remain well under the acceptance standards of mapping agencies. This means that although the potential exists, still OSM is not a product ready to be used as it is.

This Thesis has examined whether it is possible to successfully implement the accumulated knowledge and experience of the GI community regarding the spatial data quality on the loose form of UGSC. The findings of the research showed clearly that the up to now established methods (such as the ISO suggested methodology) are not only applicable but are also necessary to elucidate the quality status of UGSC. However, one of the most important lessons learned from the OSM analysis is that the evolution of UGSC brought along new uncertainty sources for the spatial data available on the Web. For example, the importance or the popularity of the space examined, the land use (e.g. urban vs. rural) the number of spatial content contributors and the underlying socio-economic factors of an area (e.g. deprived or privileged) have been recorded as influential to quality factors. Thus, UGSC opens up new areas for further research in the subject matter of spatial data quality.

One such issue is the need for a theoretic background to document and explain these new uncertainty sources along with empirical research that will provide firm models of how the UGSC quality is affected. This Thesis has made the first step to set a knowledge base that has been supported by empirical findings and sets the context to meet the next important challenge: the evolution of the spatial data quality evaluation processes so to be able to assess and measure the UGSC quality. This can be achieved through the modification, completion or adjustment of the existing quality evaluation processes so as to enable them to take into account the new generation of spatial data and their sources of uncertainty. Alternatively new quality evaluation processes could be developed that will combine both old and new knowledge around UGSC quality evaluation.

Another issue is the ability to communicate the UGSC quality. As the established methods of quality evaluation need to adapt to the new reality so does the established way of communicating spatial quality. The metadata mechanism needs to be reconsidered in the context of a wiki-based, online world of produsers. The issue of quality information sharing is further discussed in Section 7.5.

In this part of the discussion the focus was on the examination of the particularities in the UGSC quality evaluation. These particularities mainly stem from the collaborative environment of spatial data generation and the social factors affecting this process. It has been discussed that these novel to Geomatics issues need further attention. Possible errors and misconceptions during the development of either real-world projects like OSM or theoretical approaches to handle these issues are both expected and justifiable. However, what is not justifiable is the negligence of the established knowledge in Geomatics domain. In Geomatics there are already known erroneous processes that have been identified, received extensive coverage and documented by scholars and researchers that have developed the appropriate methods for their remedy. There is no need to reinvent the wheel in the name of Web 2.0. OSM project, in an effort to provide free spatial data by mobilising the crowd, fell into the trap of trying to re-invent everything around the subject matter of building a spatial repository. There is a very large pool of knowledge built by the GI community that has not been embodied efficiently in OSM. For many years numerous GI experts have been building spatial databases and mapping the globe. Most of the times the poor results come from the fact that such efforts are based on underfunded efforts or highly bureaucratic procedures. What OSM has shown the world is the strategy to overcome such problems, how to break the barriers that an institutional organisation poses and how to mobilise active citizens to map the globe. This is a unique achievement both in social and Geomatics terms. Scientifically though, OSM has not achieved the same results. OSM project has only partially taken advantage of the existing knowledge around the issues of cartography, spatial databases and quality control.

* * *

As discussed in Sections 7.2 and 7.3, UGSC is realised through the creation of geo-tagged photos and vector-encoded data. Given the delayed national mapping programs, both the variety and the volume of these datasets might prove to be of great help in the future mapping efforts. However, there is still quite a large distance between the outcome generated by the Web 2.0 geo-applications and the standards that mapping agencies are setting for themselves. Thus, the fact remains that if an engagement between mapping agencies and Web sources of UGSC was to take place, then both parties need to adjust to the new reality. Definitely the UGSC sources have to make all the necessary steps to realise the erroneous processes used, find the causes that introduce errors into their data and develop methodologies to tackle them whereas mapping agencies need to investigate the possibilities that such sources hide and adjust accordingly their future strategy. Thus far, the discussion concentrated on the findings of the empirical experiments and included the advantages, the potentials and the issues of concern related with the UGSC phenomenon. Another round of discussion will follow but here the aim is to highlight two new challenges that the UGSC phenomenon hides. These challenges, along with the ones discussed in Chapter 6, have been surfaced during the empirical experiments conducted on the course of the Thesis but here theoretical solutions will be provided. Nevertheless, these challenges need to be addressed proactively both by mapping agencies and the Web sources that provide UGSC. It is interesting to note that mapping agencies need to do so whether they decide to create in-house UGSC Web applications or to use external sources to get such content.

7.4 Spatial explicit sources

One of the most important issues revealed during this research is the development of Web 2.0 geo-applications that will be able to support internal to mapping agencies procedures or to help in the creation of new mapping products. From the point of view of a mapping agency this is a very crucial issue. Either when interested in building in-house Web 2.0 geo-applications or when planning to establish close co-operation with existing ones, the factors of content productivity and the importance of the content's spatial distribution are of high importance. In that context, and based on the analysis of geo-

tagged photos, an important observation was the dichotomy between spatially implicit and explicit, Web 2.0 sources.

The analysis shows that spatially implicit sources are not able to provide the spatial distribution needed by mapping agencies. Implicit sources provide coverage for very few popular places in both national and local level. However, this was not the initial theoretical conception of the phenomenon's potentials. It has been discussed in Chapter 2 that both scholars and entrepreneurs have evangelised in favour of a social phenomenon that could bring a revolution in the Geomatics domain. Such was the enthusiasm and optimism that ideas were put forward for the creation of totally new and unconventional mapping products or the opportunity to monitor human activities that were not possible to be mapped up to now by traditional methods. In this research the validity of such approaches was challenged and it was examined whether this enthusiasm is justifiable. By examining the spatial aspects of popular photo-sharing websites the analysis both justify a moderate optimism and raise some concerns about the overall direction of the phenomenon. It has been empirically proven that neither the phenomenon has evolved in such a level nor the potentials, even of the most popular Web 2.0 sources, justify the initial optimism. This is because most of the Web 2.0 sources that use photos as their prime medium for users' socialisation are spatially implicit. These sources use the element of space as another interesting element for further user interaction and socialisation with no significant preference or support over the other elements of the Web application. The analysis showed that common social networking application cannot provide the spatial coverage needed. Nevertheless, social-oriented Web applications can be proved to be an immense pool of spatial information when the GI retrieval scope is limited to urban areas and tourism attractions as they provide a large and fairly updated pool of spatial content for such areas.

On the contrary, it has also been shown that spatially explicit sources have the potential to serve as universal sources of spatial content in terms of data volume and spatial distribution. However, explicit applications need to have a number of fundamental characteristics that should govern the user-application-space relationship to achieve this goal (on top of the charecteristics discussed in Chapter 6 – i.e. interactivity and formalization):

Space. The aim of the geo-application should be clearly related to space. Users' participation and contribution must directly interact with spatial entities in a conscious effort to better describe our world.

Space equality. The Web 2.0 geo-applications should make every effort possible not to marginalise any areas or allow spatial imbalances created by the social nature of UGSC. On the contrary, particular efforts should be put in order to lift any known barriers that could discriminate an area over another.

Quality evaluation mechanism. The Web 2.0 geo-application should have a proper mechanism in place to discover where spatial imbalances occur and be able to confront them. Such mechanism can work both at high-level (i.e. entire data repository) and low-level (i.e. spatial entities). In the former case, the mechanism will examine the entire or large parts of a dataset in order to find and reveal content imbalances. The methodology needed and its empirical implementation has been shown in the course of this Thesis (Chapters 4 and 5). In the latter case, the mechanism described in Section 6.2 is able to examine each spatial entity separately and prompt the users to act in order to restore possible errors. Interestingly, as seen in the Literature Review, the majority of the researchers focus on that latter case (i.e. entity level quality evaluation) failing to highlight the importance of an overall balanced data repository and thus missing the importance of the social factors regarding UGSC's quality.

User motivation. A systematic effort is needed to communicate and explain to the users the initial and possibly the updated goals of the application (without excluding the case these goals to have been raised by the users themselves). This will have as a consequence to motivate the lay users to participate in the common effort.

Perpetual effort. As has been discussed in Section 2.2.1.2 there is great concern regarding the sustainability of UGSC phenomenon and thus measures need to be taken to prevent users' fatigue or boredom. To avoid this, there needs to be a clear view of how the geo-application should evolve and incentives and motives should drive users' contribution towards that direction. One way to achieve this is to frequently update or re-

adjust the initial goals according to the needs of the application each given time. This should be a perpetual effort.

For a Web 2.0 geo-application to gain these characteristics the combination of the application's aims and a series of design and functionality choices to serve these aims are needed.

For example, Geograph uses the National Grid to divide space into $1km^2$ tiles and has based the entire application's functionality on this tessellation. Moreover, Geograph actively treats equally every square tile through a points awarding system and has clearly declared its aim to cover the entire area of Great Britain and Ireland with geo-tagged photos (this aim is even written in the application's logo). All these form an excellent example of how a Web 2.0 geo-application can communicate its space-related aims and at the same time design the overall application's functionality to support them. Furthermore, the initial aim of submitting one photo per tile has now been re-adjusted as the users are encouraged to provide more extensive coverage of space by submitting more photos, starting by the tiles that have less than four photos submitted to them.

Interestingly, a similar approach can also be applied to the vector-based applications. For example, it has been shown that even projects like OSM, that do belong to the spatially explicit family, need further enhancement in the communication of their goals especially when it comes to the coverage of either marginalised or rural areas. This aim can be achieved by Geograph because is in position to monitor the $1km^2$ tiles and spot the tiles with no or few geo-tagged photo submitted to them. Consequently, using that knowledge, Geograph can motivate its users to submit content for specific areas. Similarly, the OSM project can improve its quality by adopting such a strategy. OSM is not in position to realise where spatial content imbalances occur (in terms of completeness, positional accuracy, attribution quality etc.) as it has not built the functionality to monitor that. Consequently, there is no organised user motivation or when there is (e.g. through the Mapping Parties) it is not based on empirical evidence but rather on subjective criteria. Thus, a self-checking mechanism able to trace and reveal such imbalances should be in place for UGSC explicit sources.

Although in the course of this Thesis the focus was on the most successful example of such sources (i.e. OSM), there are numerous Web 2.0 applications that urge their users to submit vector-encoded data (see relative discussion in Section 2.1). Not all Web 2.0 applications that do so belong to the family of spatially explicit applications. For a Web 2.0 geo-application it is imperative its user to participate in a targeted effort to describe space otherwise the limitations of the social-networking implicit sources will emerge. On the other hand, this does not mean that content from implicit sources is of no spatial or even commercial value. For example, a Web 2.0 geo-application that simply provides the mechanism to its users to upload a number of GPS trails to demonstrate their favourite everyday running path although is not a spatially explicit application it can provide the basis for the creation of niche mapping products.

7.5 Quality information sharing

The methodology followed, the research conducted and the up to this point discussion provided answers regarding the available types of UGSC, the evolution of the phenomenon, the fundamental characteristics of the UGSC phenomenon and the functionality of the crowdsourced mechanisms of spatial data collection. Moreover, issues from different aspects of the UGSC quality elements evaluation process have been analysed using empirical studies. The findings of this evaluation, which has focused on the vector-based (OSM) datasets made evident the potentials of the data to fit a number of purposes. Early, yet important, efforts towards this direction can be found in the applications portfolio of private companies such as Cloudmade and Geofabrik.

Apart from these early signs though, a question remains whether this research's (or any other similar research for that matter) results are enough to drive mapping agencies in the use of UGSC.

The information provided by empirical research in an academic context can be proved fundamental to facilitate a series of initial steps: raise the awareness on the potentials of UGSC and stimulate the interest of an institutional organisation like a mapping agency to examine the usage of such data in its mapping procedures. This is a very important phase

that needs to be completed to move forward. But raising awareness and stimulating the interest covers part of the road until a fully active, and most importantly a fully productive, engagement of a mapping agency with UGSC sources comes into place. To cover the missing part it will need the Web 2.0 sources to provide signs that these crowdsourced spatial-oriented applications have well escaped the immature phase; they have recognised and cured erroneous processes and are taking the necessary steps to enhance their applications both internally through the formalisation and the standardisation of their processes and externally through the advances in interactivity. Although the popularity and thus the effectiveness of the sources examined has clearly been proved in the Web 2.0 world, it is crucial to emphasise that these elements are of high importance when a Web 2.0 application is examined through the Geomatics domain prism. The challenges recognised and the solutions provided in this Thesis, aim to contribute towards this direction. Yet, an environment where the interested parties will be clearly informed of the benefits that can come out of such a co-operation in order to take the necessary decisions will be needed.

Interestingly enough, a recently published, joint survey from the Association for Geographic Information and PriceWaterhouseCoopers (AGI and PwC 2010), showed eloquently that the Web 2.0 and crowdsourced geo-data have not yet captured the interest of public and private sector business leaders and practitioners. The survey reports that the Web 2.0 providers gather a mere 5% when it comes to evaluating the data areas where the most benefit resides for them, while the leading positions are occupied by pan-European mapping agencies' data (27%), national public sector data holders (22%) and from their own data stores (20%).

The remainder of this Section provides an explanation for this low acceptance of UGSC and discusses possible solutions.

All factors considered, the answer to the question whether, and under which circumstances, a mapping agency will use the UGSC in its mapping procedures boils down to the quality and uncertainty that UGSC carries. Looking the issue from a broader point of view, this choice is not much different from any other transaction between two potential interested parties. On the one hand is the administrating body or the owner (let

us call this the Data Providers) of the Web platform that could play the role of a spatial data repository that would reap some benefits from this transaction and on the other is the mapping agency with a potential interest for this data in order to improve its spatial products. In a more abstract conceptualisation of this, in the position of the mapping agency could be any other interested party (from private or public sector) that wishes to use the crowdsourced data (let us call this the Data Users). What changes with who is the second party involved in such a transaction is the benefits that each party gets and not the transaction or the product itself.

The uncertainty mentioned earlier, is not the statistical figure that refers to the difference between the measured and true value, rather the situation that each of the two involved parties is in, due to either mutual lack of information or imbalanced information regarding the true state of the data in a crowdsourced repository. So for example, if no extra effort has been put forward, none of the two parties actually knows if the UGSC is of any real spatial value and therefore both parties are in mutual ignorance. It does not stand to reason to support that the interested parties will come to an agreement and thus a transaction will not take place in such an uninformed environment. Therefore acquiring information on the quality of UGSC is paramount if such content is to enter the databases of institutional organisations like mapping agencies, or any other user for that matter. On the other hand, assuming there is effort to measure quality, it is reasonable to support that due to the dynamic nature of UGSC and the direct access that one of the two parties has on the actual data, the Data Users is not possible to have the same information about the crowdsourced data as the Data Providers. Thus there is an imbalance in the information available; information needed to make the final decision on the realisation of the transaction. Interestingly enough, both parties have the incentive to try to fix that problem. On the one hand the Data Providers want the data to be used by others (if not, why bother to build APIs?) and on the other the Data Users want to know that they get data of known quality that fit their purposes and promote their aims.

Without going into details of how such a transaction could take place regarding financial or IPRs issues, it is hard to imagine that this particular type of problem regarding imbalanced information has firstly occurred in the Geomatics domain. It is reasonable to suspect that the ergonomics behind transactions based on similar contexts have been

already studied by other disciplines. As a matter of fact, the 2001 Nobel Prize in Economics was awarded to Akerlof, Spence and Stiglitz for their analyses of markets with asymmetric information (Akerlof 1970, Spence 1973, Stiglitz and Rothschild 1976 – these references are used until the end of the Chapter). In brief, Akerlof's argument is that in a situation where a certain product in the market exists in both high and low quality but this information is only known to the party that sells the product and not to the buyer, the consequences are damaging both parties. Akerlof points out that the buyers' awareness of their ignorance make them suspicious and force them to treat any product as being of low quality and consequently bid down their offers for the products. Akerlof proved that the immediate outcome of such environments is either the high quality products will stop being offered or the market will collapse altogether and there will be no transactions at all. It is evident that this affects negatively both the sellers and the buyers that are willing to offer a fair price for a product of known quality. Thus, Akerlof has shown that when there is an environment with asymmetric information this can lead to what is known as an adverse selection. This analysis, accurately describes the environment of imbalanced information that Data Providers and Data Users of UGSC are in.

While Akerlof's contribution was the realisation and the modeling of a problem that can be observed in many fields, Spence and Stiglitz offered two different solutions focusing on the possible actions of sellers and buyers respectively. As it will be discussed later on, both of these solutions are applicable to the environment generated by the imbalanced information regarding the UGSC quality.

Spence suggested that one way to solve the problem of the asymmetric information environments is the party that has better information should act upon it and signal that information to any interested party. The important point here is that the signals themselves need to have certain characteristics so to act as a quality certification for the recipient. In other words, the signals should enable the recipient to make an informed decision in a previously uncertain environment. In order that to be true, Spence explained that the necessary signaling cost (including effort, expenses or time) should not be the same for everyone. In fact, there should be a negative correlation between the signaling cost and the product's quality. That is, the lower the product's quality the greater the

signaling cost should be and thus only sellers of high quality products would be able (and willing) to bridge that gap in order to signal such information. For example, a costly advertising campaign or a firm's extravagant corporate headquarters building can be signals for high quality products. Interestingly enough, Spence supports that this also applies to sellers that want to signal their aim to have a permanent presence in the market in contrast with those that are planning on an infrequent or a one-off presence. Spence himself focused on the job market and the signals that employees should transmit in order to inform employers of their quality. Simply described, one such signal is education as the overall cost for obtaining higher education is less for capable, and thus possibly more productive, employees than for the less capable ones.

Stiglitz's contribution transposes the ability of improving an asymmetric information environment to the uninformed party. Stiglitz and Rothschild suggested that the uninformed party can make the informed one reveal information through specific incentives. This method is called self-screening. So, by putting the sellers in a self-screening process and thus making them reveal information regarding their product's quality, the buyers can make an informed decision in an otherwise uncertain environment. The authors used the personal insurance market as an example where an insurance company provides a list of possible insurance policies that differ in the price and quantity. By doing so, the individuals that know that they belong in a high-risk category (for example because they are aware of a hereditary condition) will voluntary choose the expensive policy whereas the low-risk individuals will be happy to settle with a cheap one.

Bringing the discussion back to the UGSC issues, the question is if and how these economic theories can be implemented to balance the available information between the involved parties and whether this balance can serve as the catalyst that will enable the institutional and UGSC co-operation.

Starting with the conceptual lens that Spence provided, it would be useful to examine what a UGSC provider can do in order to signal the necessary information needed in order to facilitate the uninformed parties come to a decision. Obviously, the first step needed in this process is the Data Provider to make that extra effort to acquire the

information regarding its data quality. As this research has shown, this is a cumbersome process. Not all Web 2.0 applications can actually deliver on that task. The average and fairly obscure Web 2.0 initiatives will not bother to collect that kind of information as the overall cost of doing so will be well beyond their means. Furthermore, collecting, documenting and communicating that information in a formal and highly reputable way, such as the methodology described in the ISO specifications will make that goal more distant for the obscure players. On the other hand though, a fairly popular and with high pool of contributors Web 2.0 source, will find it easier to collect that information, even through the Web 2.0 way: motivate the crowd to do the job.

Another signal that can be sent is investing in the fundamental factor of the Web 2.0 era: the user. A Web 2.0 source that prioritises its investments in issues like usability and interactivity, apart from the obvious goal of attracting users, can send a solid sign of the existence of an endogenous high quality mechanism. At the same time, such long term investments can signal the intention of the Web 2.0 source to have a permanent presence in the field. A similar signal that can be beneficial in more than one way is the investment in the data capturing and data sharing infrastructure. As in the case of usability and interactivity, facilitating the data flow in a Web 2.0 application, apart from enabling the gathering of huge volumes of data, signals the intention of the Web 2.0 application to remain active. In contrast, the cost to invest in, develop and maintain such technological advances (which will be provided to the users free of charge in order to be in line with the Web 2.0 era) will prohibit obscure sources to follow that route. Such examples can be found in the decision of Flickr and Picasa to develop desktop applications that further facilitate the photo sharing or the development of different OSM editors.

Finally, a more conventional (meaning that is not spatial data specific) way for a UGSC source to signal its strength is through the public relationships / marketing strategy. For example, part of this strategy is organising and hosting international conferences. An interesting point in case is the State of the Map Conference organised annually by the OSM community. It is difficult to justify why a community that has categorically proven in practice that location and distance between its members plays absolutely no role in their effort to map the globe, needs an international conference that requires physical

presence. There are plenty of technological alternatives to support any co-ordination or ideas exchange that might be needed. Such conferences are primarily credentials of the strength and the potentials of the specific Web 2.0 source that other obscure sources are not able to acquire.

As explained, these initiatives are expected to originate from the Data Providers with high quality UGSC in an effort to signal their differentiation form the obscure sources. The flip side of this statement is that exactly for such (or similar) signals the Data Users should be scanning the world of GeoWeb in an effort to recognise the data sources that could match their needs. Nevertheless, for Data Users this is a passive, yet necessary, stance towards the problem. In order to examine what an active stance would be to balance the information asymmetry Stiglitz's conceptual apparatus can be used. The point here for the Data Users is to provide the necessary incentives to make the Data Providers enter a self-screening process in regards with the quality of the data that they can offer. One such solution could be for the Data Users to offer a higher premium for datasets that can prove the fulfillment of certain standards. This will automatically motivate Data Providers to provide insights regarding the state of their data in order to get the higher possible premium. Consequently, the Data Users will gain information that was previously unknown to them. An example could be for a NMA to provide lower premium for data that are not accompanied with sufficient and formally structured metadata. Web sources that are not keeping a good track of their data contribution process and thus will not be able to compile the necessary metadata will ask for a contract or a co-operation schema that offers low premiums. In contrast, sources that are functioning in a well organised and carefully designed environment will be able to provide the metadata needed and thus get the higher premiums. A step further towards that direction could be for NMAs to provide various incentives (e.g. higher premiums, closer co-operation, technical support etc.) to Data Providers to reveal information regarding the degree of their implicit or explicit nature by providing datasets' statistics, similar to those analysed in the course of this Thesis. Overall, following this self-screening method will provide NMAs the needed information to understand the range of incentives that should be provided, develop a balanced value for price strategy against the UGSC and make secure decisions regarding a minimum threshold of data quality acceptance that will possibly needed to be in place.

An interesting twist in the whole argument, regarding the role of NMAs, is that such organisations in their everyday practice are the Data Providers and not the Data Users as it was suggested during the development of the argument. This change in roles takes place only when NMAs need to get data from another source; in our case from a Web 2.0 source. In other words, NMAs are the institutions that provide spatial data to their users and therefore are the ones that need to establish a communication channel that will help them to signal their quality to their potential users and clients. It can be assumed that this communication channel is already in place and works, more or less fine for their in-house, traditionally created spatial datasets. However, when it comes to data collected from Web 2.0 sources though it is a whole different story. In other words, this might prove a quality trap for NMAs. NMAs that will engage in a co-operation with external Web 2.0 sources, must not allow their users to have second thoughts about UGSC affecting the overall quality of the data provided. Both NMAs and the Web 2.0 sources should be able to send the proper signals to any interested party that the outcome of this co-operation is of certified quality. This is even more important when the Web 2.0 source is an in-house build application (e.g. the 'OS explore' application where users can upload their recorded routes). In this case, building the credentials of their Web 2.0 application is a task that NMAs should by no means neglect or underestimate.

7.6 Summary

Initially, the discussion focused on the empirical analysis' findings from geo-tagged photos. Both the positive and negative issues of UGSC were examined. The main characteristics of the phenomenon in terms of data spatial distribution (both in small and large scale), data flow, currentness and users' behaviour were discussed. Furthermore, the effect that these UGSC's characteristics could have to mapping agencies has been highlighted.

The discussion then moved to the findings from the OSM analysis. The subject matter of UGSC's quality was the main point in focus. The issues of the OSM positional accuracy and the attribution quality of OSM datasets were examined. In parallel, the danger for

Web 2.0 geo-applications to end up with two speed spatial datasets has been discussed. The analysis showed that this can occur mainly due to the impact that social factors have on the UGSC phenomenon's spatial dimension.

The next level of the discussion focused on an important observation revealed by the empirical research: the need for spatially explicit, Web 2.0 applications. It has been shown that such applications are able to provide both the data volume and the spatial distribution necessary to support GI retrieval at a national level. More specifically, the discussion focused on the basic characteristics that a spatially explicit, Web 2.0 geo-application should have. One of the most important elements is the conscious effort to interact with spatial entities. Other important issues that have been raised were the ability of the application to reveal possible errors or content imbalances and a mechanism of motivations and incentives that urge the users for their co-operation.

The final part of the discussion focused on the fundamental ergonomics that govern a possible closer co-operation between UGSC sources and mapping agencies. The existing knowledge in the Economics domain was used in an effort to explain the barriers that deter the interested parties to get involved in such a co-operation and the possible solutions. The information asymmetry around UGSC's quality is the major obstacle. However, it has been shown that there are methods to mitigate such imbalances if proper strategies are followed by the interested parties.

Chapter 8

Conclusions and recommendations for future directions

8. Conclusions and recommendations for future directions

8.1 General

It is unorthodox to claim that the first big achievement of a doctoral Thesis is its subject. However, back in 2007 when this effort started the term VGI, coined by Goodchild (2007a), was counting few months of life and the UGSC phenomenon was still in the phase of denial from private companies and mapping agencies. Thus, it is fair to support that one of this effort's achievements consists in foreseeing the value and the importance of this newly born phenomenon.

It should be noted that this subject has a fundamental advantage, too. The fact that the UGSC was a fairly uncharted phenomenon gave a relative freedom in heading the research to selected issues. At the same time though, the unknown was the Thesis' Nemesis. Indeed, one of the most important difficulties was the fact that there were very few references for this subject compared to other Geomatics issues. Worse, the existing bibliography was still in the level of theoretical approaches and assumptions around the phenomenon with almost no empirical evidence. Thus, there was an objective difficulty in adopting or rejecting ideas and suggestions found in the literature which frequently were contradicting each other.

However, all these painted a very interesting research environment. On the one hand, there was a largely unknown phenomenon for which the interest (or hype) was increasing and on the other, an academic community that had just started to realise the phenomenon's potentials and dangers but there were no clearly documented and empirically supported views of its nature and value. In that context, this Thesis set as its primary target to create a knowledge base around specific aspects of the phenomenon that would be useful to mapping agencies but also to UGSC sources and to lay users as well.

In the course of this Thesis a series of empirical examinations took place that covered data from the main UGSC types available today. The data collected and examined (approximately 6.5m geo-tagged photos, 1.4m OSM intersections, 1.3m OSM entities and 2.3m OSM tags) and the results shed light to many areas of UGSC. Consequently, this helped to corroborate or contradict theories previously supported by researchers. Moreover, the findings showed that there are issues of concern regarding the evolution of the phenomenon and challenges that need to be faced effectively by those who wish to engage with UGSC.

Relative to this last point, an important achievement of this Thesis is that it did not repose in the discussion of the experiments' results but rather used them to highlight erroneous processes, disadvantages and characteristics that needed further investigation. For some of these challenging issues practical solutions were presented. For others theoretical approaches were used (with references to empirical data whenever that was applicable).

However, the overall direction of the Thesis was to provide sufficient answers to the research questions and thus to fulfill the objectives set at the beginning of this effort. Indeed, now all elements are in place to reflect on the Thesis' research objectives.

8.2 Research objectives revised

8.2.1 Understand the nature of the UGSC phenomenon

An important contribution of the Thesis is that it provided enough evidence to realise the nature of UGSC. The aim was not simply to describe the phenomenon in a conceptual or theoretical level, but to empirically examine its nature and discover its distinctive characteristics. By empirically examining at a national level the main sources of UGSC available today it was made clear how the phenomenon is realised in practice. The Thesis showed that UGSC is a social as much as a spatial phenomenon. More importantly though, it showed how space and geography affect the social footprint of the users on the Web. The work on geo-tagged photos showed clearly how type, reputation and

popularity of space affect the spatial content submission by lay users. It showed that if no extra effort is put forward space is an overwhelming factor that acts against spatially balanced content creation. These results were also corroborated by the work on OSM. There it was shown that spatial differentiation of the underlying socio-economic reality affects directly the quality of the spatial content submitted.

Another issue relevant to the nature of the UGSC phenomenon revealed by this Thesis is that not all applications are suitable to serve as universal sources of GI despite the magnitude of users' contribution. More importantly, the Thesis has highlighted the fundamental characteristics that a Web 2.0 geo-application should have in order to play such a role. The infusion of specific characteristics on the applications' aims and functional designs can create spatially explicit applications (see also Section 8.2.3.2).

These findings have also contributed to the evaluation of theoretical assumptions that had been introduced by the academia in an effort to describe the phenomenon. For example the fears that the phenomenon will prove to be a passing fad is not corroborated. On the contrary, the examined data flow to the photo-sharing Web 2.0 geo-applications proved that although there are seasonal fluctuations the overall trend shows an increase in content submission. Similar conclusions source from the analysis of the OSM entities. However, on the flip side, the optimism about a citizens' network that will act as living sensors has not been corroborated either. The data clusters observed in the social networking applications are not allowing the support of that argument. Interestingly, both options are still there for the UGSC's future. UGSC as a phenomenon is at a crossroad, one of many that will probably follow: it can succumb to the social sirens of space popularity and affluence, and become a heterogeneous, two speed data repository or defy them and become the evolution that will provide an unprecedented map of the globe.

8.2.2 Evaluation of UGSC quality

The second objective of this Thesis was to evaluate the quality of the spatial content generated by lay users. As explained in Chapter 2, spatial data quality is a complex issue that consists of many components. In an effort to complement existing research on data

completeness, this Thesis focused on the examination of the quality element of positional accuracy and on quality elements relative to the attribution process of OSM.

The important point revealed is that while there was a fundamental change in the data origin (i.e. from mapping agencies and GI professional to lay users) the quality evaluation mechanisms remained fairly the same. Thus, the quality evaluation of UGSC used methods that were designed for institutional data. The Thesis has shown that although existing evaluation processes are still applicable, they are not enough to provide a fully effective description of the UGSC data quality. Factors such as data clustering or quality heterogeneity that source from users' participation, space and geography or the underlying socio-economic reality are not detected by the traditional methodologies of quality evaluation. This is primarily because the institutional spatial data creation processes are not affected by such error sources. Thus the quality evaluation methodologies need to adjust accordingly so as to be able to document effectively errors that are generated by the social nature of the UGSC phenomenon.

Closely related to the previous conclusion is the one that results from the analysis of the positional and attributes quality evaluation. More specifically, it has been shown that taking into account the technology used (e.g. GPS or satellite imagery) sufficient positional accuracy has been recorded. Thus it has been shown that the positional accuracy is in accordance with the technological means available and it is reasonable to conclude that it will improve according to the technological evolution. However, the Thesis showed that the conceptualisation of space and of the physical or man-made entities by the lay user proves to be fundamentally more challenging. This conceptualisation, as realised by the tags submitted for the OSM entities, shows that encloses the individuality of each user and thus ultimately introduces an unwanted diversification in the data created. Thus, for UGSC there are quality elements that are technology-related and they are expected to improve in the future but also there are quality elements that are user-related and introduce noise in the data repository that, if no further action is taken, are expected to deteriorate as more users join the Web 2.0 UGSC sources.

Finally, as discussed in the Literature Review, there has been a shift in the responsibility to interpret the quality of a spatial dataset. It has been explained that the data producer is not making any judgments about the usability of its products, rather reports the results of a series of tests that the products undergo and it is up to the data consumers to determine the fitness-for-purpose of a specific spatial dataset. Similarly, this Thesis has not made any such concrete judgments (although a few examples where given in Section 7.3). What this Thesis has achieved through the empirical quality evaluation, is to build a tangible profile of UGSC quality. This significantly contributes to the data consumers' role of determining the fitness-for-purpose of UGSC. The knowledge built around the quality of UGSC can be used by any interested party to conclude on the fitness-for-purpose of such content.

8.2.3 Highlight the challenges of UGSC and possible solutions

As expected for such a new and evolving phenomenon, in the course of the research a number of challenging issues were recognised. The confrontation of some of these challenges had been set as a research objective. Four issues have drawn the attention of the Thesis: data formalisation and quality improvement, interactivity, spatial explicit geo-applications and quality information sharing. For the former two, practical solutions have been presented. For the latter two, the Thesis contributed to the theoretical understanding of UGSC.

8.2.3.1 Data formalisation and quality improvement

An interesting conclusion is revealed when both the positional and attribution accuracy are considered. In the former case, the positional accuracy (i.e. the geometry of GPS traces or on-screen digitisation) reaches the technological possible accuracy whereas in the latter case heterogeneity and consequently noise is obvious in the data created. This can be attributed to the fact that in the former case the users have very specific drivers regarding what they should collect (i.e. spatial entities on the ground or on a satellite image). In the attribution process though such drivers do not exist (with the exception of the attribute domains applied in the OSM editors) and the wiki pages are too loosely

structured to provide a firm users' guide for such a complex endeavor as to map the globe. Indeed, the Thesis has clearly shown that the users' heterogeneity is reflected in the data submitted. For this challenge a practical solution was provided. As poor formalisation reduces the overall spatial data quality it has been concluded that the main step towards the improvement of the data quality is to conceptually formalise the data sought. This step will enable UGSC sources to create the necessary architectural and functional environment both in the front and the back-end of the application so as to guide contributors during their data submission. In turn, this will improve data quality by diminishing errors and inconsistencies in the dataset. Such a mechanism has been described in detail and a prototype application has been developed as a proof of concept.

8.2.3.2 Interactivity

From the beginning of this research the role and the importance of interactivity has been discussed for the Web 2.0 applications in general and particularly for the geo-applications.

On the one hand there was the dominance of raster-based Web maps and consequently the limitations introduced in content-level interactivity. On the other, there was the need for increased user participation and interaction that could lead to increased content generation. In that context, achieving efficient content-level interactivity for the Web 2.0 geo-applications was set as an important objective of the Thesis.

The method to infuse content level interactivity to the Web 2.0 geo-applications is to provide support for effective vector data handling on a Web client. The limitations of such efforts have been discussed extensively in Chapter 2 and it was realised that in order to meet this objective the effort needed to be concentrated to the creation of a methodology that would enable effective vector data transmission over the Web. Indeed, such a methodology has been developed in the course of this Thesis and it has been tested in a real-world environment with promising results. The same methodology was used in the prototype application built for the UGSC formalisation process. The prototype used the content level interactivity to enable quality improvement by hosting users' reactions on the view of a quality aware vector map.

8.2.3.2 Spatially explicit geo-applications

Another objective of the Thesis was to allow the interested reader, based on the knowledge built from this research, to be able to understand and describe the fundamental characteristics that a Web 2.0 geo-application should have so as to serve as UGSC source for universal GI retrieval purposes.

The research has shown that the new breed of Web-based applications that aspire to have such a role should clearly set goals that are primarily related with space. These geo-applications should declare to their users that their aim should be to directly contribute spatial entities in a conscious effort to better describe the world. This effort should not be deterred by spatial or social inequalities. Furthermore, it has been shown that these applications should be designed in such a way so as to detect content imbalances generated by the social nature of the phenomenon. When such imbalances occur the geo-application should be in position to motivate users to fix them. This motivation should belong to a broader strategy that will aim to constantly communicate and explain to the users the goals of the application each given time. Such Web 2.0 geo-applications are characterised as spatially explicit.

8.2.3.4 Quality information sharing

The current situation regarding the level of UGSC acceptance from the established institutional mapping agencies (public or private) is a challenging issue. Web 2.0 with all of its particular characteristics (i.e. long tail, bi-directional flow of data, enhanced role of the user etc.) is expected to play an increasingly important role in the future. Consequently, phenomena (such as the UGSC) that have sprung from it are expected to follow suit. Therefore, UGSC has the potential to be a phenomenon with increasing importance in the Geomatics domain. Thus, it is expected that the degree of interaction between institutional organisations and crowed-sourced initiatives will steadily grow. It has been supported that, given the differences of these two parties, this interaction will not be a straightforward process and it is in the best interest of all involved parties their co-operation to be based on healthy and solid basis of mutual understanding. Theoretical

foundations from the discipline of Economics were used to provide a strategy towards that direction. It has been concluded that quality signaling strategies from both parties will help the removal of the information imbalances and they will facilitate future co-operation.

8.3 Recommendations for future directions

When this effort started, UGSC was a fairly new phenomenon that was mainly attracting the interest of academic circles. Today the attitude of the Geomatics world towards UGSC has changed radically. Mapping agencies are trying to grasp the importance and the benefits of the phenomenon, private companies are investing in its evolution and services from all around the public sector are relying on the user's contribution to map activities and phenomena that affect the citizens. More importantly, informed and now accustomed to the value of maps and GI citizens expect or even demand the presence of such spatial routes of communication with the authorities.

As this Thesis reaches the end, the question is how UGSC will evolve and what the role of GI experts in this evolution is. Perhaps the starting point on this quest should be the exploitation of the spatial usability of UGSC. Taking as example the case of geo-tagged photos, methodologies should be developed for separating the noise from the actually useful information. Pockets of spatially usable information are available among the huge volumes of data generated constantly, but the GI retrieval methods have not evolved accordingly to isolate and extract them.

Furthermore, future efforts should aim to build quality intelligent geo-applications. The lay users need to become accustomed not only to the spatial data creation but also need to realise that the quality of the data created is paramount. However, it is neither fair to simply blame the users for the quality of the data produced nor realistic to expect that knowledge about spatial data quality will come out of nowhere and it will succeed to suppress users' individuality. Unlike other cases in the Web, UGSC is helped by technological drivers (e.g. GPS-enabled devices, satellite imagery etc.) during its creation process and thus some of its quality elements are kept at high level. For the rest,

the GI experts should provide the means to improve them. Also, closer attention should be paid to enrich the existing quality evaluation methods with new processes that will be able to provide sufficient results on the quality of UGSC by taking into account sources of errors unknown to the traditional data capturing methodologies.

At the same time, it would be interesting to further examine whether there is a pattern regarding how error-prone different classes of users are, depending on various factors such as their productivity or the overall time of user activity. For example, it would be enlightening to realise if the most productive users are also the most scrupulous or if the older members of the OSM community have gained a level of trust in their data capturing abilities and thus have stopped consulting the wiki pages or if error generation is an independent phenomenon from the users' background. Furthermore, it would be interesting to examine if the 80-20 Pareto rule is applicable not only in content generation but in error creation too. In other words, are 20% of the users responsible for the 80% of the errors introduced in UGSC, and if this is so, what is their distribution in the contributor's population and is it possible to isolate them? Are the incautious users randomly distributed? Do they belong to the old and seemingly experienced users or are they part of the new/occasional contributors? As repeatedly has been stated in this Thesis, UGSC is a social as much as a Geospatial phenomenon. Thus the 'user' factor needs to be constantly examined in any analysis and there is plenty of room for further research in the social aspects of the phenomenon.

All these recommendations could be part of a broader discussion on the evolution of UGSC. The next generation of UGSC sources does not have to consist from applications that are either geo-tagged based or vector-based. As some of the key private players have started to explore, an interesting case for an open project would be a hybrid effort that could include the vector-based map of an area complemented by descriptions, geo-tagged photos or video. In other words, the next level of expansion in the Geomatics could come by the convergence of multiple forms of data (e.g. text, geometry, photographs, videos etc.) in an effort to describe spatial entities. Finally, the GI community can help to move the Web-based geo-applications into the next level: transform them from data collecting applications to results producing ones by setting the

correct basis and showing the road for the development of geo-applications that could conduct and present spatial analysis' results based on UGSC.

8.4 Final thought

The evolution of the bi-directional Web 2.0 has created the phenomenon of UGSC. This Thesis by anlysing different aspects of UGSC showed its positive and negative points and highlighted the important issues and challenges that lie ahead. Issues like the volume of data produced and the crowd mobilisation for mapping the globe are impressive achievements. However, quality issues and the creation of spatially explicit geo-applications able to guide lay users in a conscious interaction with space need further effort.

The UGSC phenomenon has the potential to be the starting point of the next big thing not just in Geomatics but in the entire world of the Web. A lot needs to be done in order for the next version of geo-applications to be able to generate the necessary data, and the role of GI experts in this effort will be crucial. The spatial dimension of information, whether this is geographic, social, economic, cultural or any other type, is there. What is missing is the way to reveal it and interconnect it. Evolutions like the UGSC and the Linked Data (Berners-Lee 2009) are paving the road for a true GeoWeb.

References

Ajzen, I., 1989. Attitude structure and behavior. In Pratkanis, A.R., Breckler, S.J. and Greenwald, A.G. (eds.), *Attitude Structure and Function. Lawrence Erlbaum Associates*. Inc, Hillsdale NJ, pp. 241–274.

Akerlof, A.G., 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3), pp. 488-500.

Anderson, C., 2004. The Long Tail. *Wired*, Issue 12.10, 2004.

Anderson, C., 2006. The Long Tail. Croydon: Random House Business Books.

Andrienko, L. G., Andrienko, V. N., 1999. Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, 13(4), pp.355-374.

Antoniou, V., Tsoulos, L., 2006. The potential of XML encoding in geomatics converting raster images to XML and SVG. *Computers & Geosciences*, 32 (2), pp.184-194.

Antoniou, V., Morley, J., 2008. Web Mapping and WebGIS: do we actually need to use SVG? *SVG Open 2008*. [Online] http://www.svgopen.org/2008/papers/82-eb_Mapping_and_WebGIS_do_we_actually_need_to_use_SVG [Accessed 25 October 2009]

Antoniou, V., Morley, J. and Haklay 2008, Is your Web map fit for purpose? Drawing a line under raster mapping. *AGI GeoCommunity '08*, Stratford-upon-Avon, UK, 24-25 September 2008.

Antoniou, V., Morley, J. and Haklay, M.M., 2009a. Tiled Vectors: A Method for Vector Transmission over the Web. In J. D. Carswell, A. S. Fotheringham, & G. McArdle *Lecture Notes in Computer Science*. Berlin/Heidelberg: Springer, pp. 56-71.

Antoniou, V., Haklay, M. and Morley, J., 2009b. The role of user generated spatial content in mapping agencies. *Proceedings of the GIS Research UK 19th Annual Conference*. Durham, pp.251-255.

Antoniou, V., Haklay, M. and Morley, J., 2010a. A step towards the improvement of spatial data quality of Web 2.0 geo- applications: the case of OpenStreetMap. *Proceedings of the GIS Research UK 18th Annual Conference*. London, pp.197-202.

Antoniou, V., Haklay, M. and Morley, J., 2010b. Web 2.0 Geotagged Photos: Assessing the Spatial Dimension of the Phenomenon. *Geomatica (Special issue on VGI)*, 64(1), pp. 99-110.

Aronoff, S., 1989. *GIS: A Management Perspective*. Ottawa:WDL Publications.

Association for Geographic Information (AGI) and PricewaterhouseCoopers (PwC), 2010. *Opportunities in a changing world*. [Online] http://www.agi.org.uk/storage/AGI_PwC%20Survey.pdf [Accessed 10 December 2010].

Ather, A., 2009. *A Quality Analysis of OpenStreetMap Data*. Unpublished M.Eng. dissertation, Department of Civil, Environmental and Geomatic Engineering, University College of London.

Basiouka, S., 2009. *Evaluation of the OpenStreetMap quality*. Unpublished MSc Thesis, Department of Civil, Environmental and Geomatic Engineering, University College of London.

Bearden, M. J., 2007. *The National Map Corps*. [Online] http://www.ncgia.ucsb.edu/projects/vgi/participants.html [Accessed 27 May 2008].

Berners-Lee, T., 2009. *Linked data design issues*. [Online] http://www.w3.org/DesignIssues/LinkedData.html [Accessed 15 December 2010].

Bertolotto, M, 2007. Progressive Techniques for Efficient Vector Map Data Transmission: An Overview. In: Belussi A., Catania, B., Clementini, E. and Ferrari, E. (eds.) *Spatial Data on the Web: Modeling and Management*. Springer, New York , pp. 65-84.

Bertolotto, M., Egenhofer, M.J., 2001. Progressive Transmission of Vector Map Data over the World Wide Web. *GeoInformatica*, 5(4), pp. 345-373.

Birdsall, F.W., 2007. Web 2.0 as a Social Movement. *Webology*, 4(2) [Online] http://www.webology.ir/2007/v4n2/a40.html [Accessed 14 July 2010].

Bishr, M., Mantelas, L., 2008. A trust and reputation model for filtering and classification of knowledge about urban growth. *GeoJournal*, 72(3-4), pp.229-237.

Brando, C., Bucher, B., 2010. Quality in User Generated Spatial Content: A Matter of Specifications. *Proceedings of 13th AGILE International Conference on Geographic Information Science*. Guimarães, Portugal.

Bruns, A., 2008b. The future is user-led: the path towards widespread produsage. *FiberCulture Journal* (11). [Online] http://eprints.qut.edu.au/12902/1/12902.pdf [Accessed 18 August 2010].

Budhathoki, N., Bruce, B. and Nedovic-Budic, Z., 2008. Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal*, 72(3), 149-160.

Buttenfield, B.P., 2002. Transmitting vector geospatial data across the Internet. In M. Egenhofer and D. Mark (eds.), *Lecture Notes in Computer Science* vol. 2478, Berlin: Springer, pp. 51–64.

Cartwright, E.W., 2008. Mapping in a digital age. In Wilson P.J. and Fotheringham A.S. (eds) *The handbook of geographic information science*. Malden: Blackwell Publishing.

Cecconi, A., Galanda, M., 2002. Adaptive Zooming in Web Cartography. *Computer Graphics Forum*, vol. 21, pp. 787-799.

CEN/TC 287, 1998. *Geographic Information - Data description –Quality*. ENV 12656:1998.

Chrisman., N., 2006. Development in the Treatment of Spatial Data Quality. In Devillers R. and Jeansoulin R, (eds) *Fundamentals of spatial data quality elements*. ISTE:London, pp.21-30.

Clinton., W., 1994. Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure. *Executive Order 12906*, 59(71).

Clinton, W., 2000. *Improving the civilian global positioning system (GPS)*. Office of Science and Technology Policy, Executive Office of the President. [Online] http://clinton4.nara.gov/WH/EOP/OSTP/html/0053_4.html [Accessed 10 January 2009]

Coleman, D.J., Georgiadou, Y. and Labonte, J., 2009. Volunteered Geographic Information : The Nature and Motivation of Produsers. *International Journal of Spatial Data Infrastructures Research*, vol. 4, p.332-358.

Cook, D., Symanzik, J., Majure, J. J. and Cressie, N., 1997. Dynamic graphics in a GIS: more examples using linked software. *Computers and Geosciences*, vol. 23, pp. 371-385.

Cover., R., 2006. Audience inter/active: Interactive media, narrative control and reconceiving audience history. *New media & society*, 8(1), pp. 139-158.

Cox, M.A., 2008. Flickr: a case study of Web2.0. *Aslib Proceedings: NewInformation Perspectives*, 60(5), pp. 493-516.

Craglia, M., 2007. Volunteered Geographic Information and Spatial Data Infrastructures: When Do Parallel Lines Converge? [Online] http://www.ncgia.ucsb.edu/projects/vgi/participants.html [Accessed 23 May 2008].

Craglia, M., et al. 2008. Next-Generation Digital Earth. *International Journal of Spatial Data Infrastructures Research*, Vol. 3, 146-167.

Cross, R., Smith, J., 1996. Consumer-focused strategies and tactics. In Forrest, E., Mizerski, R. (eds.) *Interactive Marketing: The Future Present*. Business Books, pp. 5-27.

Davis, F.D., Bagozzi, R.P. and Warshaw, P.R., 1989. User acceptance of computer technology: a comparison of two theoretical models. *Management Science*, vol. 35, 982–1003.

Department of Commerce, 1992. *Spatial Data Transfer Standard (SDTS)*. Federal Information Processing Standard 173: Washington, Department of Commerce, National Institute of Standards and Technology.

Devillers, R., R. Jeansoulin. 2006b. *Fundamentals of Spatial Data Quality*. London: ISTE.

Duce, D., Herman I., and Hopgood, B., 2002. Web 2D Graphics File Formats. *Computer Graphics Forum*, 21(1), pp.43-64.

Dunfey, I. R., Gittings, M. B. and Batcheller, K. J., 2006. Towards an open architecture for vector GIS. *Computers & Geosciences*, vol. 32, pp.1720–1732.

Dunn, C.E., 2007. Participatory GIS a people's GIS? *Progress in Human Geography*, 31(5), 616-637.

Dykes A. J., 1997. Exploring spatial data representation with dynamic graphics. *Computers and Geosciences*, 23(4), pp. 345-370.

Dykes, J. and Wood, J., 2008. *Mashup Visualization with Google Earth and GIS*. [Online] http://www.gicentre.org/infovis/ [Accessed March 15, 2009].

Dykes, J., Purves, S.R., Edwardes, J.A. and Wood, J., 2008. Exploring Volunteered Geographic Information to describe Place: Visualization of the 'Geograph British Isles' Collection. *Proceedings of GIS Research UK 16th Annual Conference*. Manchester, pp. 256-267.

Elwood S., 2002. GIS use in community planning: a multidimensional analysis of empowerment. *Environment and Planning A*, 34(5), pp. 905-922.

Elwood, S., 2008a. Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. *GeoJournal*, 72(3-4), pp.133-135.

Elwood, S., 2008b. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72(3-4), pp.173-183

Elwood, S., 2008c. Grassroots groups as stakeholders in spatial data infrastructures: challenges and opportunities for local data development and sharing. *International Journal of Geographical Information Science*, 22(1), pp.71-90.

ESRI, 2006. *Comparing Vector and Raster Mapping for Internet Applications*. [Online] www.esri.com/library/whitepapers/pdfs/vector-raster-mapping.pdf [Accessed 02 March 2008].

Estes, J.E., Mooneyhan W., 1994. Of maps and myths. *Photogrammetric Engineering and Remote Sensing*, 60(5), pp. 517-524.

European Union (EU) 2007. *Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. [Online] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:en:PDF [Accessed 01 June 2010]

Flanagin A., Metzer M., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4), pp.137–148.

Flickr 2010. About Flickr. [Online] http://www.flickr.com/about/ [Accessed 18 December 2010]

Friedman, T. L., 2006. *The world is flat: A brief history of the twenty-first century*, updated and expanded edition. New York: Farrar, Straus and Giroux.

Geograph 2006. *Ordnance Survey to sponsor the Geograph Project*. [Online] http://www.geograph.org.uk/help/press_release_001 [Accessed 17 December 2010].

Ghose, S., Dou, W.Y., 1998. Interactive functions and their impacts on the appeal of internet presences sites. *Journal of Advertising Research,* vol. 38, pp. 29-43.

Giles, J. 2005. Special report: Internet encyclopaedias go head to head. *Nature*, vol. 438, pp. 900-901.

Goodchild, M. F., 2007a. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211–221.

Goodchild, M. F., 2007b. Citizens as voluntary sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial data Infrastructure Research*, vol. 2, pp. 23-32.

Goodchild, M. F., 2007c. *Beyond metadata: towards user-centric description of data quality*. [Online]www.geog.ucsb.edu/~good/papers/435.pdf [Accessed 17 August 2009].

Goodchild, M. F., 2008a. *Assertion and Authority: The Science of User-Generated Geographic Content*. [Online] http://www.geog.ucsb.edu/~good/papers/454.pdf [Accessed 23 July 2009].

Goodchild, M. F., 2008b. Commentary : Wither VGI?. *GeoJournal*, 72(3-4), pp. 239-244.

Goodchild, M. F., 2008c. Spatial Accuracy 2.0. *Proceeding of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences*. Shanghai, June 25-27.

Goodchild, M.F., 2009. NeoGeography and the Nature of Geographic Expertise. *Journal of Location Based Services*, 3(2), pp. 82-96.

Goodchild, M. F. and Hunter, G. J., 1997. A simple positional accuracy measure for linear features, *International Journal of Geographical Information Science*, 11(3), pp. 299-306.

Goodchild, M. F., Fu, P. and Rich, P., 2007. Sharing geographic information: An assessment of the geospatial onestop. *Annals of the Association of American Geographers*, 97(2), pp. 250-266.

Grira, J., Bédard, Y. and Roche, S., 2010. Spatial data uncertainty in the VGI world: going from consumer to producer. *Geomatica* , 64(1), pp. 61-71.

Guélat, J., C., 2009. Integration of user generated content into national databases - Revision workflow at swisstopo. *1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases*, Bern, 20-21 August.

Haklay, M. and Tobon, C., 2002, Usability Engineering and PPGIS: Towards a Learning-improving Cycle. *1st Annual Public Participation GIS Conference*, Rutgers University, New Brunswick, New Jersey, 21-23 July.

Haklay, M., Tobón, C., 2003. Usability Evaluation and PPGIS: Towards a User- Centred Design Approach. *International Journal of Geographical Information Science*, 17(6), pp. 577–592.

Haklay, M., 2006. Usability Dimensions in Collaborative GIS. In Balram S. and Dragicevic S., (eds.) *Collaborative Geographic Information Systems*, Idea Group Inc. pp. 24-42.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4), pp. 682-703.

Haklay, M., Basiouka, S., Antoniou, V. and Ather, A., 2010. How Many Volunteers Does It Take To Map An Area Well? The validity of Linus' law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4), pp. 315-322.

Haklay, M., Singleton, A. and Parker, C., 2008. Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass*, 2(6), pp. 2011-2039.

Hanke, J. 2007. The evolution of the GeoWeb. [Online] Where 2.0 conference video clip. http://conferences.oreillynet.com/cs/where2007/view /e_sess/12583. [Accessed May 05, 2008].

Havercroft, M., 2009. Crowdsourcing – some experiences and thoughts. *1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases*, Bern, 20-21 August.

Heipke, C., 2010. Crowdsourcing geospatial data. *Journal of Photogrammetry and Remote Sensing*, 65(6), pp. 550-557.

Held, G., Ullrich, T., Neumann, A. and Winter, A. M. 2004. *Comparing .SWF (Shockwave Flash) and .svg (Scalable Vector Graphics) file format specifications*. [Online] http://www.carto.net/papers/svg/comparison_flash_svg/ [Accessed 30 Jan 2008]

Hoffman, D.L., Novak, T.P., 1996b. New metrics for new media: toward the development of web measurement standards. [Online] http://www2000.ogsm.vanderbilt.edu/novak/Web.standards/Webstand.html [Accessed 21 May 2009].

Hudson-Smith, A., Crooks, A., 2008. *The Renaissance of Geographic Information:*

Neogeography, Gaming and Second Life. [Online]
http://www.casa.ucl.ac.uk/working_papers/paper142.pdf [Accessed 3 Februrary 2009].

International Organisation for Standardisation, 1988. *8601 Data elements and interchange formats -- Information interchange -- Representation of dates and times*, Geneva: ISO.

International Organisation for Standardisation, 2003. *19115 Geographic information - Metadata*, Geneva: ISO.

International Organisation for Standardisation, 2005a. *19101 Geographic information - Reference Model*, Geneva: ISO.

International Organisation for Standardisation, 2005b. *19113 Geographic information - Quality principles*, Geneva: ISO.

International Organisation for Standardisation, 2005c. *19114 Geographic information - Quality evaluation procedures*, Geneva: ISO.

International Organisation for Standardisation 2005d. *9000 Quality management systems - Fundamentals and Vocabulary*, Geneva: ISO.

International Organisation for Standardisation, 2006. *19138 Geographic information — Data quality measures*, Geneva: ISO.

International Organisation for Standardisation, 2007. *19139 Geographic information - Metadata – XML schema implementation*, Geneva: ISO.

Jarrett., K., 2008.Interactivity is Evil! A critical investigation of the Web 2.0. *First Monday*, 13(3). [Online]
http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2140/1947 [Accessed 9 June 2010].

Jones, C.B., Ware, J.M.: Map generalization in the Web age. *International Journal of Geographical Information Science,* 19(8), pp. 859-870.

Keen, A. 2007. *The Cult of the Amateur*. London: Nicholas Brearley.

Kiousis, S., 2002. Interactivity: A Concept Explication. *New Media & Society*, 4(2), pp. 271–291.

Kraak, M-J., 2001. Settings and needs for Web cartography. In Kraak M-J., and Brown A., (eds) *Web Cartography*. Taylor and Francis, London.

Kuhn, W., 2007. Volunteered Geographic Information and GIScience. *Position Paper for the NCGIA and Vespucci Workshop on Volunteered Geographic Information*, Santa Barbara, California. [Online] http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Kuhn_paper.pdf [Accessed 1 Oct 2008]

Lanier, J., 2006. Digital Maoism: The hazards of the new online collectivism. *Edge*, No. 183. [Online] http://www.edge.org/3rd_culture/lanier06/lanier06_index.html. [Accessed 25 October 2008].

Lehto, L., Sarjakoski, L.T., 2005. Real-time generalization of XML-encoded spatial data for the Web and mobile devices. *International Journal of Geographical Information Science,* 19(8), pp. 957-973.

Liu S.B., Palen, L., Sutton, J., Hughes, L.A.and Vieweg, S., 2008. In Search of the Bigger Picture: The Emergent Role of On-Line Photo Sharing in Times of Disaster. *In Proceedings of the 5th International ISCRAM Conference*. Washington, DC.

MacEachren, A. M., 1995. *How maps work. Representation, visualization, and design*. London: The Guilford Press.

MacEachren, A. M., 2000. Cartography and GIS: facilitating collaboration. *Progress in Human Geography*, 24(3), pp. 445-456.

MacEachren, A., Kraak, M., 1997. Exploratory cartographic visualization: advancing the agenda. *Computers and Geosciences*, 23(4), pp. 335-344.

MacEachren A.M. and J. Kraak. 2001. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28 (1), pp.3-12.

Mackaness, A.W., 2008. Generalization in spatial databases. In Wilson P.J. and Fotheringham A.S. (eds) *The handbook of geographic information science*. Malden:Blackwell Publishing,.

Malone, W.T., Laubacher, R., and Dellarocas, N.C., 2009. *Harnessing Crowds: Mapping the Genome of Collective Intelligence*, MIT Sloan Research Paper No. 4732-09. [Online] http://ssrn.com/abstract=1381502 [Accessed 15 April 2010]

Martin, D., 2007. SurfaceBuilder. [Online] http://www.public.geog.soton.ac.uk/users/martind/davehome/software.htm [Accessed 01 July 2008].

Maue, P., Schade, S., 2008. Quality of geographic information patchworks. *Proceedings of 11th AGILE International Conference on Geographic Information Science*, University of Girona, Spain.

McDougall, K., 2009. The potential of citizen volunteered spatial information for building SDI. *GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges*, Rotterdam 15-19 June. [Online] http://eprints.usq.edu.au/5406/1/McDougall_GSDI_11_AV.pdf [Accessed 10 January 2010].

McDougall, K., 2010. From Silos to Networks – Will Users Drive Spatial Data Infrastructures in the Future? *FIG Congress 2010. Facing the Challenges – Building the*

Capacity, Sydney, Australia, 11-16 April. [Online] http://www.fig.net/pub/fig2010/papers/ts08b/ts08b_mcdougall_4137.pdf [Accessed 18 September 2010].

McMillan, S., 2002. A Four-Part Model of Cyber-Interactivity: Some Cyber-Places are More Interactive than Others. *New Media & Society*, 4(2), pp. 271-91.

Miller, C., M., 2006. A beast in the field: the Google Maps mashup as GIS/2. *Cartographica*. 41 (3), pp. 187-199.

Morrison, J.L., 1995. Spatial data quality. In: Guptill, S.C. and Morrison, J.L. (eds.), *Elements of spatial data quality*. International Cartographic Association (ICA), Tokyo: Elsevier Science, pp. 1-12.

Navteq 2010. NAVTEQ Map Reporter. [Online] http://mapreporter.navteq.com/ [Accessed 13 June 2010].

Neumann, A., 2002. *Comparing .SWF (Shockwave Flash) and .svg (Scalable Vector Graphics) file format specifications*. [Online] http://www.carto.net/papers/svg/comparison_flash_svg/index07.shtml [Accessed 30 Jan 2008].

Neumann, A., Winter, M.A., 2003. Webmapping with Scalable Vector Graphics (SVG): Delivering the promise of high quality and interactive Web maps. In Peterson M.P., (ed.) *Maps and the Internet*, New York:Elsevier.

Neumann, A., Winter, A.M., 2004. *Vector-based Web Cartography: Enabler SVG*. [Online] http://www.carto.net/papers/svg/index_e.shtml [Accessed 01 Feb 2008].

Newhagen, J.E., 2004. Interactivity, Dynamic Symbol Processing, and the Emergence of Content in Human Communication. *The Information Society*, 20(5), pp. 397-402.

O'Reilly, T., 2005. *What is Web 2.0: design patterns and business models for the next generation of software*. [Online] http://www.oreillynet.com/lpt/a/6228 [Accessed 02 May 2008].

O'Reilly. T., 2007. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *MPRA Paper*, No. 4578. [Online] http://mpra.ub.uni-muenchen.de/4578/ [Accessed 12 January 2008]

Obermeyer, N., 2007. *Thoughts on volunteered (geo) slavery*. [Online] http://www.ncgia.ucsb.edu/projects/vgi/participants.html [Accessed 13 March 2008].

Organisation for Economic Co-operation and Development, 2007. *Participative Web: user-created content*. [Online] www.oecd.org/dataoecd/57/14/38393115.pdf [Accessed 15 January 2008]

Ordnance Survey, 2009. *Meridian 2. User Guide and Technical Specification*. [Online] http://www.ordnancesurvey.co.uk/products/meridian2/pdf/meridian2userguide.pdf [Accessed 11 January 2010].

Ordnance Survey, 2010a. *Information on the National Grid*. [Online] http://www.ordnancesurvey.co.uk/oswebsite/freefun/geofacts/geo0667.html [Accessed 15 July 2010].

Ordnance Survey, 2010b. *Corporate governance: agency performance monitors*. [Online] http://www.ordnancesurvey.co.uk/oswebsite/aboutus/corpgov/apm.html [Accessed 15 December 2010].

OpenStreetMap, 2009. *Any tags you like*. [Online] http://wiki.openstreetmap.org/wiki/Any_tags_you_like [Accessed 11 January 2010].

OpenStreetMap, 2010a. *Potential Datasources*. [Online] http://wiki.openstreetmap.org/wiki/Potential_Datasources [Accessed 11 July 2010].

OpenStreetMap, 2010b. *Main Page*. [Online]
http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 11 July 2010]

OpenStreetMap, 2010c. *OpenStreetMap – Fast Facts*. [Online]
http://www.cloudmade.com/press/wp-content/uploads/2010/08/100813-OSM_Facts.pdf [Accessed 11 July 2010].

Peng, Z.R., Tsou, M.H., 2003. *Internet GIS: Distributed geographic information services for the Internet and wireless networks*. Hoboken, NJ:Wiley.

Peng, Z.R., Zhang, C., 2004. The roles of geography markup language (GML), scalable vector graphics (SVG), and Web feature service (WFS) specifications in the development of Internet geographic information systems (GIS). *Journal of Geographical Systems*, 6(2), pp.95-116.

Peterson, M.P., 2003. Maps and the Internet: An Introduction. In Peterson M.P. (Ed.) *Maps and the Internet*, New York: Elsevier.

Peterson, M. P., 2008. *International Perspectives on Maps and the Internet*. New York: Springer.

Phelan, S., 2008. Opening statement of AGI 2008. *AGI GeoCommunity '08*, Stratford-upon-Avon, UK, 24-25 September.

Plewe, B., 1997. *GIS Online: Information Retrieval, Mapping, and the Internet*. Santa Fe, NM: OnWord Press.

Plewe, B., 2007. Web Cartography in the United States. *Cartography and Geographic Information Science*, 34(2), pp. 133-136.

Popescu, A., Grefenstette, G. and Moëllic., A.P., 2008. Gazetiki: automatic creation of a geographical gazetteer. *In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. Pittsburgh PA, PA, USA: ACM, pp. 85-93.

Preece, J., et al., 1994. *Human-Computer Interaction*. Wokingham: Addison-Wesley.

Purves S.R., Edwardes J.A., 2008. Exploiting Volunteered Geographic Information to describe Place. *In Proceedings of GIS Research UK 16th Annual Conference*. Manchester, pp.252-255.

Putz, S., 1994. Interactive Information Services Using World-Wide Web Hypertext. *Proceedings of the First International Conference on World-Wide Web*, Geneva, 25-27 May 1994.

Rafaeli, S., 1988. Interactivity: From New Media to Communication. In Hawkins, H., Wiemann, J. and Pingree, S., (eds) *Advancing Communication Science: Merging Mass and Interpersonal Processes*, London: Sage, pp. 110–34.

Raymond, E.S., 1999. The Cathedral and the Bazaar. *Knowledge, Technology, and Policy*, 12(3), pp. 23-49.

Richards., R., 2006. Users, interactivity and generation. *New Media Society*, 8(4), pp. 531-550.

Rideout, T., 2007. Scotland & The people's Map. *In preceding of AGI - Inspiring Scotland*, Edinburgh, 15th November 2007.

Rigaux , P., Scholl , M. and Voisard, A., 2002. *Spatial Databases: With Application to GIS*. San Francisco: Morgan Kaufmann Publishers.

Robinson, H. A., et al., 1995. *Elements of Cartography*. 6th ed. New York: Wiley.

Schroeder, P. 1996. Criteria for the design of a GIS/2. *Specialists' meeting for NCGIA Initiative 19: GIS and society*. [Online] http://www.spatial.maine.edu/_schroedr/ppgis/criteria.html [Accessed 29 January 2009].

Servigne., S., Lesage., N. and Libourel., T., 2006. Quality Components, Standards and Metadata. In Devillers R. and Jeansoulin R, (eds) *Fundamentals of spatial data quality elements*. ISTE:London, pp. 179-201.

Shi, W., Fisher., F. P. and Goodchild, F. M., 2002. *Spatial Data Quality*. New York: Taylor and Francis.

Shirky, C., 2005. Institutions vs. collaboration. [Online] http://www.ted.com/index.php/talks/clay_shirky_on_institutions_versus_collaboration.html [Accessed 02 June 2010].

Sieber, R. 2006. Public Participation and Geographic Information Systems: A Literature Review and Framework. *Annals of the American Association of Geographers*, 96(3), pp. 491-507.

Sieber, R., 2007. *Geoweb for social change*. [Online] http://www.ncgia.ucsb.edu/projects/vgi/supp.html [Accessed 13 March 2008].

Skupin, A., Fabrikant, S. I., 2007. Spatialization. In Wilson, J. and Fotheringham, J. (eds) *Handbook of geographic information science*. Malden, MA: Blackwell Publishers.

Spence, M., 1973. Job Market Signaling. *The Quarterly Journal of Economics*, 87(3), pp. 355-374.

Steiner, E.B., MacEachren , A.M. and Guo, D., 2001. Developing and assessing light-weight data-driven exploratory geovisualization tools for the Web. *Proceedings of the 20th International CartographicConference*, 6-10 August, Beijing, China.

Stiglitz, J.E., Rothschild, M.E, 1976. Equilibrium in competitive insurance markets. *Quarterly Journal of Economics*, 90(4), pp. 629-649.

Sui, Z., 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), pp. 1-5.

Sundar, S.S., 2004. Theorizing Interactivity's Effects. *The Information Society*, 20(5), pp. 385-389.

Svensson, P., 2008. *Creator of Web spots a flaw in Internet Explorer*. Associated Press. [Online] http://www.msnbc.msn.com/id/26646919 [Accessed 25 Feb 2010]

Talen E, 1999. Constructing neighborhoods from the bottom up: the case for resident-generated GIS. *Environment and Planning B: Planning and Design*, 26(4), pp. 533-554.

Talen, E., 2000. Bottom-Up GIS: A New Tool for Individual and Group Expression in Participatory Planning. *Journal of the American Planning Association*, 66(3), pp. 279-294.

Tapscott, D., 2009. *Grown Up digital*. London: McGraw-Hill.

Tapscott, D., Williams D. A., 2008. Wikinomic. London: Atlantic Books.

Teoa, H.H., Oha, L.N., Liua, C. and Weib, K.K., 2003. An empirical study of the effects of interactivity on web user attitude. *International Journal of Human-Computer Studies*, 58 (3), pp. 281-305.

Thomas J.J., Cook K.A., 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press.

Tran, T., 2007. *Google Maps Mashups 2.0*. [Online] http://google-latlong.blogspot.com/ [Accessed 2 November 2007].

Tulloch, L. D., 2008. Is VGI participation? From vernal pools to video games. *GeoJournal*, 72(3-4), 161–171.

Turner, A. J., 2006. *Introduction to neogeography*. Sebastopol, CA: O'Reilly Media Inc.

Van Oort, P.A.J, 2006. *Spatial Data Quality: From Description to Application*. PhD Thesis,Wageningen University. [Online] http://www.ncg.knaw.nl/Publicaties/Geodesy/pdf/60Oort.pdf [Accessed 12 December 2008].

Viglino, J.M., 2009. Handling partner's feedbacks through the web. *1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases,* Bern, 20-21 August.

Word Wide Web Concortium, 2001. *Scalable Vector Graphics (SVG) 1.0 Specification*. [Online] http://www.w3.org/TR/2001/REC-SVG-20010904/ [Accessed 01 Feb 2008].

Weibel, R., Dutton, G., 1999. Generalizing spatial data and dealing with multiple representations. In: Longley, P., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications*, 2nd Edition. Cambridge: GeoInformation International, pp. 125-155.

Wikipedia, 2010. *World Wide Web*. [Online] http://en.wikipedia.org/wiki/Www [Accessed 13 August 2010].

Williams, S., 2007. *Application for GIS specialist meeting*. [Online] http://www.ncgia.ucsb.edu/projects/vgi/participants.html [Accessed 15 April 2008].

Williams, N. J., Neumann, A., 2006 a. *Interactive hiking map of Yosemite national park.* [Online] http://www.mountaincartography.org/publications/papers/papers_bohinj_06/26_Williams_Neumann.pdf [Accessed 01 Feb 2008].

Wood, M., 1994. Visualization in historical context. In MacEachren , A. M., and Taylor, D. R. F. (eds), *Visualization in Modem Cartography*. Oxford:Pergamon.

Woolford., T., 2008. Using Geo-Tagged Images in GIS. *AGI GeoCommunity '08*, Stratford-upon-Avon, UK, 24-25 September.

Yahoo! 2009. *4,000,000,000*. [Online] http://blog.flickr.net/en/2009/10/12/4000000000/ [Accessed 13 July 2010].

Yahoo!, 2010. *Flickr: Explore everyone's photos on a Map*. [Online] http://www.flickr.com/map/ [Accessed 13 July 2010].

Yang, B., 2005. A multi-resolution model of vector map data for rapid transmission over the Internet. *Computers and Geosciences*, 31(5), pp. 569-578.

Yang, C., et al., 2005. Performance-improving techniques in web-based GIS. *International Journal of Geographical Information Science*, 19(3), pp. 319-342.

Yang, B., Purves, R. and Weibel, R., 2007. Efficient transmission of vector data over the Internet. *International Journal of Geographical Information Science*, 21(2), pp. 215-237.

YouTube, 2010. *YouTube Fact Sheet*. [Online] http://www.youtube.com/t/fact_sheet [Accessed 14 July 2010].

Zaslavsky, I., 2003. Online Cartography with XML. In Peterson M.P., (ed.) *Maps and the Internet*, New York:Elsevier.

Zhao, H., Shneiderman, B., 2005. Colour-coded pixel-based highly interactive Web mapping for georeferenced data exploration. *International Journal of Geographical Information Science*, 19(4), pp. 413-428.

Zhou, M., Bertolotto M., 2004. A Data Structure for Efficient Transmission of Generalised Vector Maps. In M. Bubak et al. (eds.), *Lecture Notes in Computer Science*, Berlin/Heidelberg:Springer-Verlag, pp. 948–955.

Zielstra, D., Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *13th AGILE International Conference on Geographic Information Science*. Guimaraes, Portugal.

Appendix A

Flickr[14]:

'…You agree not to reproduce, duplicate, copy, sell, trade, resell or exploit for any commercial purposes, any portion of the Service (including your Yahoo! ID), use of the Service, or access to the Service…'.

Picasa Web[15]:

'…Unless you have been specifically permitted to do so in a separate agreement with Google, you agree that you will not reproduce, duplicate, copy, sell, trade or resell the Services for any purpose…'.

Panoramio[16]:

'…The photographic imagery made available for display through Panoramio is provided under a nonexclusive, non-transferable license for use only by you. You may not use the photographic imagery in any commercial or business environment or for any commercial or business purpose for yourself or any third parties…'.

Geograph[17]:

'…The foregoing license rights are intended to provide to Geograph Project Ltd all rights under existing and future laws, including without limitation all rights under copyright and any other rights personal to You to publish the Submission on the Site, use the Submission in publicity and promotional materials for the Site and to create new Sites or derivative works (including without limitation by combining the Submission with other content) for public display or performance via any and all media or technology now known or later developed. The foregoing rights may be licensed and sublicensed through unlimited tiers of third parties…'.

OSM[18]:

Creative Common Licence (Attribution-Share Alike 3)[19]:

[14] http://info.yahoo.com/legal/us/yahoo/utos/utos-173.html
[15] https://www.google.com/accounts/TOS?hl=en
[16] http://www.panoramio.com/terms/
[17] http://www.geograph.org.uk/help/terms
[18] http://www.openstreetmap.org/
[19] http://creativecommons.org/licenses/by-sa/2.0/

'…If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one…'.

EveryTrail[20]:

'…You may only download or print a copy of any portion of the Content solely for your own personal, non-commercial use…'.

[20] http://www.everytrail.com/tos.php

Appendix B

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All motorway features with a 'node' tag	All motorway features with a 'FIXME' tag	All motorway features with a 'source' tag	All motorway features with a 'bridge' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - Domain Consistency	2 - Domain Consistency	2 - Domain Consistency	2 - Domain Consistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "note tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "FIXME tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "source tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "bridge tags should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0%**	**0%**	**0%**	**6.64%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All motorway features with a 'layer' tag	All motorway features with a 'maxweight' tag	All motorway features with a 'maxspeed' tag	All motorway features with a 'maxheight' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - Domain Consistency	2 - Domain Consistency	2 - Domain Consistency	2 - Domain Consistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "layer tags should follow the domain enumeration (-5 to 5)" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "for weight use metric tons" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "Values are assumed to be in kilometres per hour (km/h) unless units are explicit" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "maxheight expresses a height limit for using the way to which the tag is added. If no unit is included, the value is assumed to be in metres" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0,54%**	**100%**	**57.31%**	**100%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All motorway features with a 'lit' tag	All motorway features with a ‘oneway’ tag	All motorway features with a 'lanes' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - Domain Consistency	2 - Domain Consistency	2 - Domain Consistency
Data quality measure			
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID			
Data quality evaluation method			
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "lit tag should follow the domain enumeration (yes/no)" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "oneway tag should follow the domain enumeration (yes/no/-1)" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "Total number of physical travel lanes making up the way" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result			
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0%**	**23,18%**	**0.33%**
Data quality value unit	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified

Table 26. Motorway attribures domain evaluation.

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items classified as Highways	All items classified as Highways	All items classified as Highways	All items classified as Highways
Data quality element	1 – completeness	2 – logical consistency	2 – logical consistency	2 – logical consistency
Data quality sub-element	2 – omission	2 – domain consistency	2 – domain consistency	2 – domain consistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID	N/A	N/A	N/A	N/A
Data quality evaluation method				
Data quality evaluation method type	1 – internal (direct)	1 – internal (direct)	1 – internal (direct)	1 – internal (direct)
Data quality evaluation method description	Divide count of features which have null value to the "osmuser" attribute by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the domain of the "rec_time" attribute by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the domain of the "way_id" attribute by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the domain of the "changeset" attribute by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – percentage	4 – percentage	4 – percentage	4 – percentage
Data quality value	**0.55%**	**0.0%**	**0.0%**	**0.0%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All items classified as Highways	All items classified as Highways	All items classified as Highways
Data quality element	2 – logical consistency	2 – logical consistency	2 – logical consistency
Data quality sub-element	2 – domain consistency	2 – domain consistency	2 – domain consistency
Data quality measure			
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID	N/A	N/A	N/A
Data quality evaluation method			
Data quality evaluation method type	1 – internal (direct)	1 – internal (direct)	1 – internal (direct)
Data quality evaluation method description	Divide count of features which violate the domain of the "user_id" attribute by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the domain of the "version" attribute by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the domain of the "visible" attribute by the number of features in the data quality scope. Multiply the result by 100.
Data quality result			
Data quality value type	4 – percentage	4 – percentage	4 – percentage
Data quality value	**0.0%**	**0.0%**	**0.0%**
Data quality value unit	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified

Table 27. OSM Highways data quality. Domain consistency evaluation for system-generated attributes.

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All unclassified features with a 'note' tag	All unclassified features with a 'FIXME' tag	All unclassified features with a 'description' tag' tag	All unclassified features with a 'image' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "note tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "FIXME tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "description tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "image tags should have an image url" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0,59%**	**0,81%**	**0,00%**	**94,22%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All unclassified features with a 'source' tag	All unclassified features with a 'source_ref' tag	All unclassified features with a 'abutters' tag	All unclassified features with a 'oneway' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "source tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "source_ref tags should have a string value" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "abbuters tag should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "oneway tag should follow the domain enumeration (yes/no/-1)" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0,06%**	**4,37%**	**2,21%**	**19.32%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All unclassified features with a 'lit' tag	All unclassified features with a 'name' tag	All unclassified features with a 'smoothness' tag	All unclassified features with a 'surface' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "lit tag should follow the domain enumeration (yes/no)" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "name tag should be a valid road name" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "smoothness tag should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "surface tag should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**6.37%**	**0.07%**	**0,00%**	**1,41%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All unclassified features with a 'access' tag	All unclassified features with a 'traffic_calming' tag	All unclassified features with a 'tunnel' tag	All unclassified features with a 'bridge' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "access tag should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "traffic_calming tag should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "tunnel tag should follow the domain enumeration (yes)" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "bridge tag should follow the domain enumeration" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**1,88%**	**8,33%**	**11,29%**	**8,02%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Data quality component	**Component domain**	**Component domain**	**Component domain**	**Component domain**
Data quality scope	All unclassified features with a 'layer' tag	All unclassified features with a 'maxweight tag	All unclassified features with a 'maxspeed' tag	All unclassified features with a 'maxheight' tag
Data quality element	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency	2 - Logical Consistency
Data quality sub-element	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency	2 - DomainConsistency
Data quality measure				
Data quality measure description	Percentage of violating items	Percentage of violating items	Percentage of violating items	Percentage of violating items
Data quality measure ID				
Data quality evaluation method				
Data quality evaluation method type	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)	1- Direct (internal)
Data quality evaluation method description	Divide count of features which violate the principle: "layer tags should follow the domain enumeration (-5 to 5)" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "for weight use metric tons" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "Values are assumed to be in kilometres per hour (km/h) unless units are explicit" by the number of features in the data quality scope. Multiply the result by 100.	Divide count of features which violate the principle: "maxheight expresses a height limit for using the way to which the tag is added. If no unit is included, the value is assumed to be in metres" by the number of features in the data quality scope. Multiply the result by 100.
Data quality result				
Data quality value type	4 – Percentage	4 – Percentage	4 – Percentage	4 – Percentage
Data quality value	**0.06%**	**32.48%**	**35,20%**	**34.21%**
Data quality value unit	Percent	Percent	Percent	Percent
Data quality date	2009-09-14	2009-09-14	2009-09-14	2009-09-14
Conformance quality level	Not specified	Not specified	Not specified	Not specified

Table 28. Unclassified attributes domain evaluation.

Appendix C

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns="http://openstreetmap/namespace" xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:gml="http://www.opengis.net/gml" targetNamespace="http://openstreetmap/namespace"
elementFormDefault="qualified" attributeFormDefault="unqualified">
  <xs:import namespace="http://www.opengis.net/gml"
schemaLocation="http://schemas.opengis.net/gml/3.1.1/base/basicTypes.xsd"/>
  <xs:import namespace="http://www.opengis.net/gml"
schemaLocation="http://schemas.opengis.net/gml/3.1.1/base/geometryBasic0d1d.xsd"/>
  <xs:import namespace="http://www.opengis.net/gml"
schemaLocation="http://schemas.opengis.net/gml/3.1.1/base/geometryBasic2d.xsd"/>
  <xs:element name="osm_world">
    <xs:annotation>
      <xs:documentation>root_element</xs:documentation>
    </xs:annotation>
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Physical">
          <xs:annotation>
            <xs:documentation>Only selected sub-categories are presented</xs:documentation>
          </xs:annotation>
          <xs:complexType>
            <xs:sequence>
              <xs:element name="Highway" minOccurs="0">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="Roads" minOccurs="0" maxOccurs="unbounded">
                      <xs:annotation>
                        <xs:documentation>Only selected sub-categories are presented</xs:documentation>
                      </xs:annotation>
                      <xs:complexType>
                        <xs:sequence>
                          <xs:element name="motorway" minOccurs="0" maxOccurs="unbounded">
                            <xs:complexType>
                              <xs:complexContent>
                                <xs:extension base="osm_objectType">
                                  <xs:sequence>
                                    <xs:element name="tags_Roads" type="Roads_objectType"/>
                                    <xs:element name="tags_motorway" type="tags_motorwayType"/>
                                    <xs:element name="geom">
                                      <xs:complexType>
                                        <xs:sequence>
                                          <xs:element ref="polyline"/>
                                        </xs:sequence>
                                      </xs:complexType>
                                    </xs:element>
                                  </xs:sequence>
                                </xs:extension>
                              </xs:complexContent>
                            </xs:complexType>
                          </xs:element>
                          <xs:element name="motorway_link" minOccurs="0" maxOccurs="unbounded">
                            <xs:complexType>
                              <xs:complexContent>
                                <xs:extension base="osm_objectType">
                                  <xs:sequence>
                                    <xs:element name="tags_Roads" type="Roads_objectType"/>
                                    <xs:element name="tags_motorway_link" type="tags_motorway_linkType"/>
                                    <xs:element name="geom">
                                      <xs:complexType>
                                        <xs:sequence>
                                          <xs:element ref="polyline"/>
                                        </xs:sequence>
                                      </xs:complexType>
                                    </xs:element>
                                  </xs:sequence>
                                </xs:extension>
                              </xs:complexContent>
                            </xs:complexType>
                          </xs:element>
                          <xs:element name="trunk" minOccurs="0" maxOccurs="unbounded">
                            <xs:complexType>
                              <xs:complexContent>
                                <xs:extension base="osm_objectType">
                                  <xs:sequence>
```

```
            <xs:element name="tags_Roads" type="Roads_objectType"/>
            <xs:element name="tags_trunk" type="tags_trunkType"/>
            <xs:element name="geom">
              <xs:complexType>
                <xs:sequence>
                  <xs:element ref="polyline"/>
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="unclassified" minOccurs="0" maxOccurs="unbounded">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="osm_objectType">
          <xs:sequence>
            <xs:element name="tags_Roads" type="Roads_objectType"/>
            <xs:element name="tags_unclassified" type="tags_unclassifiedType"/>
            <xs:element name="geom">
              <xs:complexType>
                <xs:sequence>
                  <xs:element ref="polyline"/>
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="track" minOccurs="0" maxOccurs="unbounded">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="osm_objectType">
          <xs:sequence>
            <xs:element name="tags_Roads" type="Roads_objectType"/>
            <xs:element name="tags_track" type="tags_trackType"/>
            <xs:element name="geom">
              <xs:complexType>
                <xs:sequence>
                  <xs:element ref="polyline"/>
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="services" minOccurs="0" maxOccurs="unbounded">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="osm_objectType">
          <xs:sequence>
            <xs:element name="tags_Roads" type="Roads_objectType"/>
            <xs:element name="tags_services" type="tags_servicesType"/>
            <xs:element name="geom">
              <xs:complexType>
                <xs:choice>
                  <xs:element ref="point"/>
                  <xs:element ref="polygon"/>
                </xs:choice>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="bus_guideway" minOccurs="0" maxOccurs="unbounded">
    <xs:complexType>
```

```
            <xs:complexContent>
              <xs:extension base="osm_objectType">
                <xs:sequence>
                  <xs:element name="tags_Roads" type="Roads_objectType"/>
                  <xs:element name="tags_bus_guideway" type="tags_bus_guidewayType"/>
                  <xs:element name="geom">
                    <xs:complexType>
                      <xs:sequence>
                        <xs:element ref="polyline"/>
                      </xs:sequence>
                    </xs:complexType>
                  </xs:element>
                </xs:sequence>
              </xs:extension>
            </xs:complexContent>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="Paths" minOccurs="0" maxOccurs="unbounded">
    <xs:annotation>
      <xs:documentation>Only selected sub-categories are presented</xs:documentation>
    </xs:annotation>
    <xs:complexType>
      <xs:sequence>
        <xs:element name="path" minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:complexContent>
              <xs:extension base="osm_objectType">
                <xs:sequence>
                  <xs:element name="tags_Paths" type="tags_PathsType"/>
                  <xs:element name="tags_path" type="tags_pathType"/>
                  <xs:element name="geom">
                    <xs:complexType>
                      <xs:sequence>
                        <xs:element ref="polyline"/>
                      </xs:sequence>
                    </xs:complexType>
                  </xs:element>
                </xs:sequence>
              </xs:extension>
            </xs:complexContent>
          </xs:complexType>
        </xs:element>
        <xs:element name="cycleway" minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:complexContent>
              <xs:extension base="osm_objectType">
                <xs:sequence>
                  <xs:element name="tags_Paths" type="tags_PathsType"/>
                  <xs:element name="tags_cycleway" type="tags_cyclewayType"/>
                  <xs:element name="geom">
                    <xs:complexType>
                      <xs:sequence>
                        <xs:element ref="polyline"/>
                      </xs:sequence>
                    </xs:complexType>
                  </xs:element>
                </xs:sequence>
              </xs:extension>
            </xs:complexContent>
          </xs:complexType>
        </xs:element>
        <xs:element name="bridleway" minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:complexContent>
              <xs:extension base="osm_objectType">
                <xs:sequence>
                  <xs:element name="tags_Paths" type="tags_PathsType"/>
                  <xs:element name="tags_bridleway" type="tags_bridlewayType"/>
                  <xs:element name="geom">
                    <xs:complexType>
                      <xs:sequence>
```

```
                                    <xs:element ref="polyline"/>
                                  </xs:sequence>
                                </xs:complexType>
                              </xs:element>
                            </xs:sequence>
                          </xs:extension>
                        </xs:complexContent>
                      </xs:complexType>
                    </xs:element>
                    <xs:element name="byway" minOccurs="0" maxOccurs="unbounded">
                      <xs:complexType>
                        <xs:complexContent>
                          <xs:extension base="osm_objectType">
                            <xs:sequence>
                              <xs:element name="tags_Paths" type="tags_PathsType"/>
                              <xs:element name="geom">
                                <xs:complexType>
                                  <xs:sequence>
                                    <xs:element ref="polyline"/>
                                  </xs:sequence>
                                </xs:complexType>
                              </xs:element>
                            </xs:sequence>
                          </xs:extension>
                        </xs:complexContent>
                      </xs:complexType>
                    </xs:element>
                    <xs:element name="steps" minOccurs="0" maxOccurs="unbounded">
                      <xs:complexType>
                        <xs:complexContent>
                          <xs:extension base="osm_objectType">
                            <xs:sequence>
                              <xs:element name="tags_Paths" type="tags_PathsType"/>
                              <xs:element name="tags_steps" type="tags_stepsType"/>
                              <xs:element name="geom">
                                <xs:complexType>
                                  <xs:sequence>
                                    <xs:element ref="polyline"/>
                                  </xs:sequence>
                                </xs:complexType>
                              </xs:element>
                            </xs:sequence>
                          </xs:extension>
                        </xs:complexContent>
                      </xs:complexType>
                    </xs:element>
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
              <xs:element name="Intersections" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType/>
              </xs:element>
              <xs:element name="Other_features" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="ford" type="osm_objectType" minOccurs="0" maxOccurs="unbounded"/>
                    <xs:element name="construction" type="osm_objectType" minOccurs="0"
maxOccurs="unbounded"/>
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element name="Barrier" minOccurs="0">
          <xs:annotation>
            <xs:documentation>any features that belong to the Barrier category</xs:documentation>
          </xs:annotation>
        </xs:element>
        <xs:element name="Cycleway" minOccurs="0">
          <xs:annotation>
            <xs:documentation>any features that belong to the Cycleway category</xs:documentation>
          </xs:annotation>
        </xs:element>
```

```
            <xs:element name="Tracktype" minOccurs="0">
              <xs:annotation>
                <xs:documentation>any features that belong to the Tracktype category</xs:documentation>
              </xs:annotation>
            </xs:element>
            <xs:element name="Waterway" minOccurs="0">
              <xs:annotation>
                <xs:documentation>any features that belong to the Waterway category</xs:documentation>
              </xs:annotation>
            </xs:element>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="NonPhysical">
        <xs:annotation>
          <xs:documentation>any features that belongs to the NonPhysical category</xs:documentation>
        </xs:annotation>
      </xs:element>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="name" type="nameType"/>
<xs:element name="ref" type="refTypeCN"/>
<xs:complexType name="osm_objectType">
  <xs:sequence>
    <xs:element name="annotation" type="annotationType" minOccurs="0"/>
  </xs:sequence>
  <xs:attribute name="osm_id" type="xs:integer" use="required"/>
  <xs:attribute name="changeset" type="xs:integer" use="required"/>
  <xs:attribute name="user" type="xs:string" use="required"/>
  <xs:attribute name="user_id" type="xs:integer" use="required"/>
  <xs:attribute name="version" type="xs:integer" use="required"/>
  <xs:attribute name="visible" type="xs:boolean" use="required"/>
  <xs:attribute name="rec_time" type="xs:dateTime" use="required"/>
  <xs:attribute name="comment" type="xs:string"/>
</xs:complexType>
<xs:element name="osm_object" type="osm_objectType"/>
<xs:element name="point">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="gml:doubleList"/>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:complexType name="osm_point">
  <xs:sequence>
    <xs:element ref="osm_object"/>
    <xs:element ref="point"/>
  </xs:sequence>
</xs:complexType>
<xs:element name="polyline" type="gml:CoordinatesType"/>
<xs:complexType name="osm_line">
  <xs:sequence>
    <xs:element ref="polyline"/>
  </xs:sequence>
</xs:complexType>
<xs:element name="polygon" type="polygonType"/>
<xs:complexType name="polygonType">
  <xs:sequence>
    <xs:element name="exterior" type="osm_line" minOccurs="0" maxOccurs="unbounded"/>
    <xs:element name="interior" type="osm_line" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:element name="minspeed" type="speedType"/>
<xs:element name="maxspeed" type="speedType"/>
<xs:simpleType name="oneway_enum">
  <xs:restriction base="xs:string">
    <xs:enumeration value="yes"/>
    <xs:enumeration value="no"/>
    <xs:enumeration value="-1"/>
  </xs:restriction>
</xs:simpleType>
<xs:element name="oneway">
  <xs:simpleType>
```

```
      <xs:restriction base="xs:string">
        <xs:enumeration value="yes"/>
        <xs:enumeration value="no"/>
        <xs:enumeration value="-1"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:complexType name="refType">
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="int_ref" type="xs:string"/>
        <xs:attribute name="nat_ref" type="xs:string"/>
        <xs:attribute name="reg_ref" type="xs:string"/>
        <xs:attribute name="loc_ref" type="xs:string"/>
        <xs:attribute name="old_ref" type="xs:string"/>
        <xs:attribute name="source_ref" type="xs:anyURI"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:complexType name="nameType">
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="name_de" type="xs:string"/>
        <xs:attribute name="name_es" type="xs:string"/>
        <xs:attribute name="name_el" type="xs:string"/>
        <xs:attribute name="int_name" type="xs:string"/>
        <xs:attribute name="nat_name" type="xs:string"/>
        <xs:attribute name="reg_name" type="xs:string"/>
        <xs:attribute name="loc_name" type="xs:string"/>
        <xs:attribute name="alt_name" type="xs:string"/>
        <xs:attribute name="official_name" type="xs:string"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:element name="speed" type="speedType"/>
  <xs:complexType name="speedType">
    <xs:simpleContent>
      <xs:extension base="xs:integer">
        <xs:attribute name="units">
          <xs:simpleType>
            <xs:restriction base="xs:string">
              <xs:enumeration value="kmp"/>
              <xs:enumeration value="mph"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:attribute>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:element name="lanes">
    <xs:simpleType>
      <xs:restriction base="xs:integer">
        <xs:minInclusive value="1"/>
        <xs:maxInclusive value="9"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:simpleType name="layer_enum">
    <xs:restriction base="xs:integer">
      <xs:minInclusive value="-5"/>
      <xs:maxInclusive value="5"/>
    </xs:restriction>
  </xs:simpleType>
  <xs:element name="bridge">
    <xs:complexType>
      <xs:simpleContent>
        <xs:extension base="yes">
          <xs:attribute name="maxweight" type="xs:string"/>
          <xs:attribute name="maxspeed" type="xs:string"/>
          <xs:attribute name="maxheight" type="xs:string"/>
          <xs:attribute name="height" type="xs:string"/>
          <xs:attribute name="length" type="xs:string"/>
          <xs:attribute name="layer" type="xs:nonNegativeInteger"/>
        </xs:extension>
```

```
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>
  <xs:simpleType name="yes">
    <xs:restriction base="xs:string">
      <xs:enumeration value="yes"/>
    </xs:restriction>
  </xs:simpleType>
  <xs:element name="tunnel">
    <xs:complexType>
      <xs:simpleContent>
        <xs:extension base="yes">
          <xs:attribute name="layer" type="layer_enum"/>
          <xs:attribute name="maxweight" type="xs:string"/>
        </xs:extension>
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>
  <xs:complexType name="tags_motorwayType">
    <xs:sequence>
      <xs:element ref="name"/>
      <xs:element ref="ref"/>
      <xs:element ref="oneway"/>
      <xs:element ref="lanes"/>
      <xs:element ref="minspeed" minOccurs="0"/>
      <xs:element ref="maxspeed"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="tags_motorway_linkType">
    <xs:sequence>
      <xs:element ref="name"/>
      <xs:element ref="ref"/>
      <xs:element ref="minspeed"/>
      <xs:element ref="maxspeed"/>
    </xs:sequence>
  </xs:complexType>
  <xs:element name="smoothness">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="excellent"/>
        <xs:enumeration value="good"/>
        <xs:enumeration value="intermediate"/>
        <xs:enumeration value="bad"/>
        <xs:enumeration value="very_bad"/>
        <xs:enumeration value="horrible"/>
        <xs:enumeration value="very_horrible"/>
        <xs:enumeration value="impassable"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="maxweight">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:pattern value="[1-9]?[0-9]\.[0-9]"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:complexType name="tags_secondaryType">
    <xs:sequence>
      <xs:element ref="name"/>
      <xs:element ref="ref"/>
      <xs:element ref="oneway"/>
      <xs:element ref="maxspeed"/>
      <xs:element ref="maxweight"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="tags_trunkType">
    <xs:sequence>
      <xs:element ref="name"/>
      <xs:element ref="ref"/>
      <xs:element ref="oneway"/>
      <xs:element ref="lanes"/>
      <xs:element ref="minspeed"/>
      <xs:element ref="maxspeed"/>
```

```
  </xs:sequence>
</xs:complexType>
<xs:element name="operator" type="xs:string"/>
<xs:complexType name="tags_bus_guidewayType">
  <xs:sequence>
    <xs:element ref="name"/>
    <xs:element ref="operator"/>
  </xs:sequence>
</xs:complexType>
<xs:element name="surface">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="paved"/>
      <xs:enumeration value="unpaved"/>
      <xs:enumeration value="compacted"/>
      <xs:enumeration value="gravel"/>
      <xs:enumeration value="pebblestone"/>
      <xs:enumeration value="cobblestone"/>
      <xs:enumeration value="cobblestone:flattened"/>
      <xs:enumeration value="paving_stones"/>
      <xs:enumeration value="paving_stones:30"/>
      <xs:enumeration value="paving_stones:20"/>
      <xs:enumeration value="grass_paver"/>
      <xs:enumeration value="ground"/>
      <xs:enumeration value="earth"/>
      <xs:enumeration value="mud"/>
      <xs:enumeration value="grass"/>
      <xs:enumeration value="sand"/>
      <xs:enumeration value="asphalt"/>
      <xs:enumeration value="concrete"/>
      <xs:enumeration value="metal"/>
      <xs:enumeration value="wood"/>
      <xs:enumeration value="dirt"/>
      <xs:enumeration value="ice_road"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="tracktype">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="grade1"/>
      <xs:enumeration value="grade2"/>
      <xs:enumeration value="grade3"/>
      <xs:enumeration value="grade4"/>
      <xs:enumeration value="grade5"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:complexType name="tags_trackType">
  <xs:sequence>
    <xs:element ref="name"/>
    <xs:element ref="tracktype"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="tags_unclassifiedType">
  <xs:sequence>
    <xs:element ref="name"/>
    <xs:element ref="oneway"/>
    <xs:element ref="footway"/>
    <xs:element ref="abutters"/>
    <xs:element ref="maxspeed"/>
  </xs:sequence>
</xs:complexType>
<xs:element name="footway">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="both"/>
      <xs:enumeration value="left"/>
      <xs:enumeration value="right"/>
      <xs:enumeration value="none"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="abutters">
```

```
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="residential"/>
        <xs:enumeration value="retail"/>
        <xs:enumeration value="commercial"/>
        <xs:enumeration value="industrial"/>
        <xs:enumeration value="mixed"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="access">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="unknown"/>
        <xs:enumeration value="yes"/>
        <xs:enumeration value="designated"/>
        <xs:enumeration value="official"/>
        <xs:enumeration value="destination"/>
        <xs:enumeration value="agricultural"/>
        <xs:enumeration value="forestry"/>
        <xs:enumeration value="delivery"/>
        <xs:enumeration value="permissive"/>
        <xs:enumeration value="private"/>
        <xs:enumeration value="no"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="sac_scale">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="hiking"/>
        <xs:enumeration value="mountain_hiking"/>
        <xs:enumeration value="demanding_mountain_hiking"/>
        <xs:enumeration value="alpine_hiking"/>
        <xs:enumeration value="demanding_alpine_hiking"/>
        <xs:enumeration value="difficult_alpine_hiking"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="width" type="xs:decimal"/>
  <xs:element name="bicycle">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="yes"/>
        <xs:enumeration value="designated"/>
        <xs:enumeration value="official"/>
        <xs:enumeration value="private"/>
        <xs:enumeration value="permissive"/>
        <xs:enumeration value="dismount"/>
        <xs:enumeration value="destination"/>
        <xs:enumeration value="delivery"/>
        <xs:enumeration value="agricultural"/>
        <xs:enumeration value="forestry"/>
        <xs:enumeration value="unknown"/>
        <xs:enumeration value="no"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="horse">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="designated"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="snowmobile">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="designated"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="ski">
    <xs:simpleType>
```

```
      <xs:restriction base="xs:string">
        <xs:enumeration value="designated"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:element name="foot">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="designated"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:complexType name="tags_pathType">
    <xs:sequence>
      <xs:element ref="sac_scale" minOccurs="0"/>
      <xs:element ref="bicycle" minOccurs="0"/>
      <xs:element ref="horse" minOccurs="0"/>
      <xs:element ref="snowmobile" minOccurs="0"/>
      <xs:element ref="ski" minOccurs="0"/>
      <xs:element ref="foot" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="annotationType">
    <xs:sequence>
      <xs:element name="note" type="xs:string" minOccurs="0"/>
      <xs:element name="fixme" type="xs:string" minOccurs="0"/>
      <xs:element name="description" type="xs:string" minOccurs="0"/>
      <xs:element name="image" type="xs:anyURI" minOccurs="0"/>
      <xs:element name="source" type="xs:string" minOccurs="0"/>
      <xs:element name="source_ref" type="xs:string" minOccurs="0"/>
      <xs:element name="source-name" type="xs:string" minOccurs="0"/>
      <xs:element name="source-ref" type="xs:string" minOccurs="0"/>
      <xs:element name="attribution" type="xs:anyURI" minOccurs="0"/>
      <xs:element name="URL" type="xs:anyURI" minOccurs="0"/>
      <xs:element name="website" type="xs:anyURI" minOccurs="0"/>
      <xs:element name="wikipedia" type="xs:anyURI" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="refTypeCN">
    <xs:simpleContent>
      <xs:extension base="refType">
        <xs:attribute name="ncn_ref" type="xs:nonNegativeInteger"/>
        <xs:attribute name="rcn_ref" type="xs:nonNegativeInteger"/>
        <xs:attribute name="lcn_ref" type="xs:nonNegativeInteger"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:complexType name="refTypeAir">
    <xs:simpleContent>
      <xs:extension base="refType">
        <xs:attribute name="icao" type="xs:string"/>
        <xs:attribute name="iata" type="xs:string"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:complexType name="addressType">
    <xs:sequence>
      <xs:element name="addr-housenumber" type="xs:string" minOccurs="0"/>
      <xs:element name="addr-housename" type="xs:string" minOccurs="0"/>
      <xs:element name="addr-street" type="xs:string" minOccurs="0"/>
      <xs:element name="addr-postcode" minOccurs="0">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:maxLength value="6"/>
            <xs:minLength value="3"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element name="addr-city" type="xs:string" minOccurs="0"/>
      <xs:element name="addr-country" minOccurs="0">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:enumeration value="GB"/>
          </xs:restriction>
```

```
        </xs:simpleType>
      </xs:element>
      <xs:element name="addr-interpolation" type="xs:string" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="tags_servicesType">
    <xs:sequence>
      <xs:element ref="name"/>
      <xs:element ref="operator"/>
    </xs:sequence>
  </xs:complexType>
  <xs:element name="traffic_calming">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="yes"/>
        <xs:enumeration value="bump"/>
        <xs:enumeration value="chicane"/>
        <xs:enumeration value="cushion"/>
        <xs:enumeration value="hump"/>
        <xs:enumeration value="rumble_strip"/>
        <xs:enumeration value="table"/>
        <xs:enumeration value="choker"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:complexType name="Roads_objectType">
    <xs:sequence>
      <xs:element ref="smoothness" minOccurs="0"/>
      <xs:element ref="surface" minOccurs="0"/>
      <xs:element ref="access" minOccurs="0"/>
      <xs:element ref="traffic_calming" minOccurs="0"/>
      <xs:choice>
        <xs:element ref="tunnel" minOccurs="0"/>
        <xs:element ref="bridge" minOccurs="0"/>
      </xs:choice>
      <xs:element ref="lit" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:element name="lit">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="yes"/>
        <xs:enumeration value="no"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <xs:complexType name="tags_PathsType">
    <xs:sequence>
      <xs:element ref="name" minOccurs="0"/>
      <xs:element ref="access" minOccurs="0"/>
      <xs:element ref="surface" minOccurs="0"/>
      <xs:element ref="lit" minOccurs="0"/>
      <xs:choice>
        <xs:element ref="bridge" minOccurs="0"/>
        <xs:element ref="tunnel" minOccurs="0"/>
      </xs:choice>
      <xs:element ref="width" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="tags_cyclewayType">
    <xs:sequence>
      <xs:element name="cycleway">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:enumeration value="lane"/>
            <xs:enumeration value="track"/>
            <xs:enumeration value="opposite_lane"/>
            <xs:enumeration value="opposite_track"/>
            <xs:enumeration value="oposite"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element ref="ref"/>
      <xs:element ref="foot"/>
```

```
      <xs:element ref="horse"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="tags_footwayType">
    <xs:sequence>
      <xs:element name="footway">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:enumeration value="both"/>
            <xs:enumeration value="left"/>
            <xs:enumeration value="right"/>
            <xs:enumeration value="none"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element ref="bicycle"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="tags_bridlewayType">
    <xs:sequence>
      <xs:element ref="ref"/>
      <xs:element ref="foot"/>
      <xs:element ref="bicycle"/>
    </xs:sequence>
  </xs:complexType>
  <xs:element name="escalator">
    <xs:complexType>
      <xs:simpleContent>
        <xs:restriction base="escalatorType">
          <xs:enumeration value="yes"/>
          <xs:enumeration value="parallel"/>
          <xs:enumeration value="no"/>
        </xs:restriction>
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>
  <xs:complexType name="escalatorType">
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="escalator_dir">
          <xs:simpleType>
            <xs:restriction base="xs:string">
              <xs:enumeration value="up"/>
              <xs:enumeration value="down"/>
              <xs:enumeration value="dynamic"/>
              <xs:enumeration value="skywalk"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:attribute>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:complexType name="handrailType">
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="handrail-left" type="yes"/>
        <xs:attribute name="handrail-right" type="yes"/>
        <xs:attribute name="handrail-center" type="yes"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
  <xs:complexType name="rampType">
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="ramp-stroller" type="yes"/>
        <xs:attribute name="ramp-bicycle" type="yes"/>
        <xs:attribute name="ramp-wheelchair" type="yes"/>
        <xs:attribute name="ramp-luggage">
          <xs:simpleType>
            <xs:restriction base="xs:string">
              <xs:enumeration value="yes"/>
              <xs:enumeration value="automatic"/>
              <xs:enumeration value="manual"/>
            </xs:restriction>
```

```
            </xs:simpleType>
          </xs:attribute>
        </xs:extension>
      </xs:simpleContent>
    </xs:complexType>
    <xs:complexType name="tags_stepsType">
      <xs:sequence>
        <xs:element name="step_count">
          <xs:simpleType>
            <xs:restriction base="xs:nonNegativeInteger">
              <xs:minInclusive value="1"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:element>
        <xs:element name="incline">
          <xs:simpleType>
            <xs:restriction base="xs:string">
              <xs:enumeration value="up"/>
              <xs:enumeration value="down"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:element>
        <xs:element ref="escalator"/>
        <xs:element name="tactile_paving">
          <xs:simpleType>
            <xs:restriction base="xs:string">
              <xs:enumeration value="yes"/>
              <xs:enumeration value="no"/>
              <xs:enumeration value="incorrect"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:element>
        <xs:element name="handrail">
          <xs:complexType>
            <xs:simpleContent>
              <xs:restriction base="handrailType">
                <xs:enumeration value="yes"/>
                <xs:enumeration value="no"/>
              </xs:restriction>
            </xs:simpleContent>
          </xs:complexType>
        </xs:element>
        <xs:element name="ramp">
          <xs:complexType>
            <xs:simpleContent>
              <xs:restriction base="rampType">
                <xs:enumeration value="yes"/>
                <xs:enumeration value="no"/>
              </xs:restriction>
            </xs:simpleContent>
          </xs:complexType>
        </xs:element>
        <xs:element name="wheelchair">
          <xs:simpleType>
            <xs:restriction base="xs:string">
              <xs:enumeration value="yes"/>
              <xs:enumeration value="no"/>
              <xs:enumeration value="limited"/>
              <xs:enumeration value="only"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
</xs:schema>
```