POLITECNICO DI MILANO

Corso di Laurea Magistrale in Ingegneria Informatica
Dipartimento di Elettronica, Informazione e Bioingegneria

OPEN LAND MAP

"VISUALIZATION, VALIDATION AND CROWD-SOURCING OF LAND COVERAGE DATA SETS"

SUPERVISOR: Prof. BROVELLI Maria
CO-SUPERVISOR: Dr. ZAMBONI Giorgio

Master Degree Thesis of:
JOLAK Rodi, mat.771380

Academic Year 2013/2014

To

My Father
Eng. JOLAK Kamal

&

My Country
Syria

"The Cosmos is also within us, we are made of stars stuff. We are a way for the Cosmos to know it self"

Carl Sagan

Contents

Introduction

Global land-cover data are key sources of information for understanding the complex interactions between human activities and global change. They are also some of the most critical variables for climate change studies. They play a critical role in improving performances of ecosystem, hydrologic, and atmospheric models. The knowledge about land cover has become increasingly important as the nation plans to overcome the problems of haphazard, uncontrolled development, deteriorating environmental quality, loss of prime agricultural lands, destruction of important wetlands, and loss of fish and wildlife habitat. Land cover data are needed in the analysis of environmental processes and problems that must be understood if living conditions and standards are to be improved or maintained at current levels. Effective management of this data will maximize our strategic response to these increasingly dynamic environmental conditions.

For better use of land, we need information on both existing land cover and land use patterns, together with their change through time. Knowledge of the present distribution of agricultural, recreational, and urban lands, as well as information on how their distribution has changed, is needed by planners and governmental entities. This will allow them to better determine land use policy, project transportation and utility needs, identify development pressure points, and strategically design effective plans for regional development.

Different organizations have developed land cover maps. The classification methods used by the many organizations looking at this data are not standardized; in addition, some land coverage areas have been miss-classified due to image acquisition during various seasons. These problems are exacerbated by difficulty in geometric and radiometric correction, and other issues related to non-standardized spectral interpretation of imagery. In order for the land coverage data sets be effective, their accuracy has to be quantitatively evaluated. To do that, high quality analyzing, classification and validation services are needed. We propose to help accomplish this with an open source geo-platform that allows sharing of these techniques and also comparing land cover classification results.

Chapter 1

General Overview

1.1 Project Description

This project implements an open Global Land Cover (GLC) information geo-platform allowing citizen science to improve land cover classification through geo-visualization and geo-crowdsourcing via internet and mobile platforms. This project will also contribute to build GLC web-based services that will allow sharing and comparing of land cover data sets to evaluate the coherency and highlight the differences between developed land cover maps. The goal is to produce more accurate and consistent land cover data for more accurate environmental change studies, greater geo-graphical understanding and improved Earth system modeling.

The GLC web-based services and the open GLC geo-platform enable easier, more efficient data sharing and information services, letting users, citizens and experts to find, evaluate, access, visualize and validate various land cover data sets. Land cover spatial-temporal information and geo-computing through web service are provided as they are important tools for supporting global change research, Earth system simulation and many other societal benefit areas. This requires an integrated knowledge representation and web implementation of land cover and how it changes over time, as well as the related operations for geo-computing. Expert users of the geo-platform will be able to upload and browse different land cover data on 3D viewers over a range of time, then classify and match the land cover categories of different

maps in order to unify the classes. To begin with, this process can be made using the main CORINE land cover categories and process the data sets having the same resolution, using rasterization or vectorization algorithms. They can also identify the discordant or miss-classified land cover areas by creating a difference map from two or more land cover maps. Moreover, they can calculate the accuracy indexes of the data using suitable algorithms, then validate the land cover maps selecting the correct land cover class and finally re-calculate the accuracy indexes after the validation phase.

Many land cover datasets have been created, but comparison studies have shown that there are large spatial discrepancies between the products. One of the reasons for these discrepancies is the lack of sufficient in-situ data for the development of these products. To address this issue, a crowdsourcing method is adopted since the exchange of geographic information has increased significantly and an enormous resource of volunteered geographic information (VGI) has become available. Volunteer citizens, very often laypersons or citizen scientists, are asked to contribute either by providing in-situ observations on the ground, adding comments, photos and videos of the land cover with geo-localized smartphones, or tracing data from other sources, such as aerial photographs or satellite. They are involved also within the validation phase via citizen science campaigns improving the accuracy of land cover datasets and creating a geospatial user-created content (crowdsourcing). Such an approach allows internet users from any region of the world to evaluate land cover data, identify the inaccuracies in that land cover data and be involved in the global validation task. The validation information is recorded and used iteratively to produce more accurate land covers. Statistics related to validation are presented to both experts and citizens.

The project provides methods to examine the crowd sourced data from the crowdsourcing services for land cover validation to determine, in the domain of remote sensing, whether there are significant differences in quality between the answers provided by experts and non-experts. Thus to identify the extent to which geo-crowd sourced data describing human impact and land cover can be used to benefit scientific research. If there are significant differences, it may be useful to create training materials with more examples in those areas where difficulties in classification are encountered. We could offer methods for contributors to reflect on the information they contribute, perhaps by providing evaluations of their contributed data along with making additional training material available.

The International Society for Photogrammetry and Remote Sensing (ISPRS) decided to establish an inter-commission working group on GLC Mapping and Services. One of the main objectives of this working group is to coordinate and plan international efforts on further development of GLC data production, validation and web services. This project contributes to build a GLC information portal to connect all major global, national and regional land cover services providing 'one stop' service with interoperable standardized service infrastructure. The GLC portal will help data users, developers and suppliers find, evaluate, access, visualize and publish land cover data and web services. It also guarantees access to a range of heterogeneous network services, local and remote, structured and unstructured, such as GLC resource discovery services, data processing and analysis services, online discussion, collaborative data editing and validation.

1.2 Land Cover and Land Use

Truth, as in a single, incontrovertible and correct fact, simply does not exist for much geographical information (GI); rather information is frequently interpreted from personal and group conceptualizations of the world and geographical data are mapped into those conceptualizations. Thus land cover information is inherently subject to indeterminacy and relativism. Herein as the number of non-specialist users of GI increases and spatial data is used to answer more questions about the environment, the need for users to understand the wider meaning of the data concepts becomes more urgent.

There are a number of current trends that contribute to the significance of this situation: First, initiatives, which originate from policy and computing developments, are promoting increased public access to spatial information with the aim of informing decision-making about the space in which people live. An example is the EU INSPIRE project which seeks to make available "relevant, harmonized and quality geographic information to support formulation monitoring and evaluation of community policies". More recently, the development of the computing grid is providing "pervasive, dependable, consistent and inexpensive access to advanced computational capabilities, databases, sensors and people". That is to say that in the area of databases, the rid has broadly the same objectives as the Spatial Data Infrastructure (SDI, INSPIRE in this case). Because of these initiatives, as well as the increasing ease of digital data transfer and a wider acknowledgement of the spatial component of much data, the number of GI users continues to

increase. Secondly, many users are interested only in the digital map. Fisher (2003) documents the shift away from extensive reports accompanying the mapped information as metadata and comments that "fewer than ever users are even aware of the existence of the survey report". Thirdly, potential users do not have to go through lengthy processes of data selection involving dialogue with the providers, nor do they often have to go through the time and expense of capturing the data through abstraction and digitizing. Rather they are able to transfer the data to their local system over the Internet or from local high capacity storage devices. Therefore strong financial incentives exist to use the readily available digital data in preference to any other source. If the data is shown to be completely unsuitable for a particular analysis, then the user can search for another source. Finally, current metadata standards are adequate to guide assessment of technical constraints on data integration caused by structure (raster to vector) or scale (generalizations to lower level classes); but, they convey nothing about the organizational cultural or epistemological context which gave rise to the data in the first place.

The net result of reducing the effort required to obtain the data, also reduces the incentive for users to understand that data in a wider sense. One of the consequences of this whole situation is that extensively manipulated information is treated as data by users who do not fully understand what it represents: its meaning or semantics. They assume that it fits their conceptualizations because of familiar class names and labels that apparently match their prototypical categories with those names. Unfortunately for almost all users the available information can only be a surrogate for the specific information they actually require a situation of which they may be unaware. The consequence of not fully understanding the conceptualizations and specifications hidden beneath familiar class labels are naive and flawed analyses, a situation that many users may not be prepared to acknowledge, and is hard to document.

1.3 Geographical Data versus Information

The geographical data and geographical information are regarded as separate and distinct phenomena. We define data as the result of measurement of some agreed phenomenon, while information is the result of interpretation, categorization, classification or some other form of processing. Measurements of grass height made in the field, records of the number and

distributions of plant species, and surveying of the elevation of the ground above a datum are all examples of geographical data. Whilst observer bias and value systems are embedded in the selection of what to measure and how to measure it, a shared conception usually exists such that if multiple observers visit a location at the same time to measure these properties within an agreed protocol, then the value which is reported has a reasonable chance of being the same. Fluctuations in that value are a matter of either the accuracy or precision of the measurement of the phenomenon.
Geographical information, on the other hand, is different. It involves processing, interpreting or transforming data to derive some sort of interpretation. For example, the identification of the cover of a parcel of land as 'Pasture' is information even if done directly, in the field. It is common for there to be some disagreement over the interpretation, for example the extent, attributes and position of the geographic phenomenon of interest such as Forest.

In providing interpretation the creation of information adds value to data. We can measure the height of a point above Ordnance Datum, but without any other "contextual" information the data is of limited value. A visit to the site would allow an observer to interpret the point in the context of its wider landscape and their conceptualization of it. The observer might identify a mountain, a hill, or a valley, all of which are information classes, and are, for most people, much richer concepts than the height data. In automated processing of geographical information parameterization is a major issue, and thus, for example, an area viewed as a channel at a detailed scale may be viewed as, a ridge, a slope or even a peak with changes in the parameterizing of the scale of measurement. Whilst the concept of channel is an unambiguous classification at a specific scale, it is not stable over changes of scale.

In the case of land cover information creation, differences between how land cover features are conceptualized have immediate implications. Harvey and Chrisman (1998) described how notions of wetlands were constructed by different environmental agencies in order to manage their policy objectives. Hoeschele (2000) documented the conflict between land cover and land use mapping for the Attappadi district of India. He revealed serious differences in how land is used and regarded by indigenous commercial and subsistence farmers, on the one hand, and by forestry technocrats, on the other. Similarly, working in Rajasthan, Robbins (2001) documented differences in the concept of forest between different users of the land, and actually implemented this

difference in a land cover classification of satellite imagery. Fisher *et al.* (in press) suggest that an origin of this problem may be in the confusion miss conceptualization of land cover as opposed to land use.

Many geographic conceptualizations can also change over time for a number of reasons. Bowker (2000) showed the influence of institutional politics on biodiversity data, and Bowker and Star (1996) noted that seemingly objective techniques for measuring nature depend on bureaucratic and institutional systems of categorization.

The implication of this situation is that one characteristic of geographical information as opposed to geographical data is that it is necessarily unreliable; there is no truth. However, accepting the absence of any single truth is not the same as saying that all interpretations are correct; for any given application many characterizations can be easily and unambiguously identified as being inappropriate. Thus a plot of land with a house on it in which a family lives would be correctly identified as residential, but it may or may not be urban, or agricultural land depending on the context; how urban and agriculture are defined and the spatial and thematic resolutions of the classification scheme in use.

Much work in GIS is conducted within an implicit conceptualization that geographical information and data are synonymous. For data it is possible to make a direct and incontrovertible measurement of the phenomenon or property of interest. The result is that different techniques, algorithms and individuals often derive equally correct but different information from the same data.

1.4 Categorization

There seem to be two ways to assign objects to categories: estimating closeness (bottom-up) or matching characteristics (top-down). In the first case experiments in cognition (Rosch 1978), show that in general people do not match characteristics but instead compare objects to “prototypes”, (“good” examples of a category), an object is assigned to the category which has the closest prototype. Unfortunately they may not be able to say how they estimate distance and what constitutes a good prototype depends on the background of the person.

In the second case, when an object has all the required characteristics then it belongs to that category, and therefore an object may belong to one, several or belong to no category. This is the more common situation in GI. Therefore, in generating geographical information using the top-down approach, we first need to agree that there is an objective reality that we wish to record, and, furthermore, that we can make precise reliable and accurate measurements of that reality or of properties of the reality (data). To generate geographical information we then need to conceptualize what it is we want to know about the reality; determine how we are going to divide the conceptual space to separate that concept into categories; decide how the properties may relate to that conceptualization and the categories; and make this relationship explicit in the form of some procedures or protocols.

Nominally this provides us with a formal ontology for mapping from the observable measurements (data) onto the concepts (information). The geographic world and geographical categorization, however, is not that simple. Varzi (2001) refers to the "double-barreled" nature of geographic entities as they are intimately connected to the space that they occupy and also infected by the manner of their human conceptualization. Furthermore, whilst many (non-geographic) objects have boundaries that correspond to physical discontinuities in the world, this is not the case for many geographic objects. Boundary placement is often problematic. Smith, in a series of excellent papers has recognized this phenomenon and developed the concept of *fiat* and *bone fide* boundaries, corresponding to *fiat* and *bone fide* geographic objects (Smith, 1995; 2001; Smith and Mark, 2001). Briefly, *fiat* boundaries are boundaries that exist only by virtue of the different sorts of demarcations effected cognitively by human beings: they owe their existence to acts of human decision. Fiat boundaries are ontologically dependent upon human fiat. *Bone fide* boundaries are all other boundaries. They are those boundaries which are independent of human fiat. So whilst ordinary (nongeographic) objects may be closed, having bone fide boundaries corresponding to physical discontinuities in the world, geographic objects may overlap. But even this is not enough, because it still assumes that the definition of object whether fiat or bona fide is relatively uncontroversial. Geographical categories, however, exist in space and react to scale (Fisher et al., 2004) and to the interface between human conceptualizations and the physical environment. Thus categories can depend on the interaction amongst human perception, spatial arrangement and properties or characteristics and can vary fundamentally with scale.

Smith and Mark (1998) commented that geographic categorization is a matter of linguistic and cultural factors. This is because defining many geographical objects necessarily involves an arbitrary drawing of boundaries in a continuum. These boundaries will differ from culture to culture, often in ways that result in conflict between groups. Therefore the boundaries contribute as much to geographic categorical definitions as the elements that they contain in their interiors. Thus we see that the two concepts of a boundary are crucial to our understanding of the world of geography. This conceptual vagueness not only affects the categorical apparatus with which geographers articulate the world; it also seems to affect the vast majority of the individual objects that geographers talk about.

1.5 Origins of different meanings of land cover classes

Land cover information derived from satellite imagery provides a convenient illustration of the way information is subsequently treated by users and how the meaning of much geographic information can be ignored. There is confusion in the way that different users treat land cover information, which originates in part from how land cover information is generated. Most users assume that land cover information can be treated as land cover data.

I. Technical, Epistemic, Physics-based:

The sensor specification, its resolving power and any image preprocessing executed influence the land cover information that can be derived from remotely sensed data. Thus the nature of the land cover features that can be identified from image data is influenced by the scale of the imagery (Woodcock and Strahler, 1987), the sampling grid (Chavez, 1992), the data captured by the pixel (Fisher, 1997), the sensor's Instantaneous Field of View (IFOV) and in which parts of the electro-magnetic spectrum it records. Commonly scale in remote sensing is a function of the sensor's IFOV which represents the ground area covered by the sensor (Forshaw et al., 1983) and the sensor's spatial resolution (Woodcock and Strahler, 1987). These determine the granularity of the data; the level of detail of the processes or objects of interest that can be extracted at that spatial resolution. Changing the scale alters the granularity of patterns of recorded reality.

Spatial resolution is commonly expressed in terms of pixel size. The pixel may correspond to a mixture of several surface types, and an area weighted average of land surface properties (Fisher, 1997). The precision of pixel values are affected by the interaction of the point spread function (PSF) which may degrade (smoothing and widening the image of sharp features), with the sensor IFOV. These factors can result in blurring of detail and reduction of the dynamic range of the measured values. Raw satellite data is subject to extensive pre-processing prior to being used for applications. Standard remote sensing textbooks describe the techniques by which remotely sensed data is corrected for geometric and radiometric errors (Lillesand and Keifer, 1987; Richards and Jia, 1993). Both types of error change the relative distribution of brightness over an image or the values of a single pixel (Richards and Jia, 1993). Corrections are made to image brightness and image geometry.

Underpinning pre-processing corrections are assumptions that surface features of interest, such as land cover, directly affect the transfer of radiation within the constraints of the sensors' IFOV and pixel size. Verstraete et al. (1996) note that the formal relations between sensor data, the properties of the classes and the effects of the state variables of radiative transfer (atmosphere, vegetation, soil, and position, size, shape, orientation or density of the objects) are rarely established. They are assumed and it is unusual to see these assumptions reported: the choice of pre-processing algorithms and control points to correct for haze and geometric distortion are not included in land cover metadata. Land cover information is not only influenced by assumptions about radiative transfer, resolution and scale factors, but also by data pre-processing prior to classification.

These issues involved in data preprocessing can be characterized as being technical and addressed by a part of the remote sensing discipline that is grounded in physics or statistics. The physicists can be caricatured as ignorant about how data and information are combined to make measurements of the biophysical world; how measurements or data are transformed into information. Yet many data pre-processing factors contribute towards the meaning of the land cover data in terms of the features on the ground that can be identified. They influence the nature of data collection or the epistemology of land cover.

II. Semantic, Ontological, Biology-based

The classification of the pre-processed (corrected) data into land cover classes also influences the nature of the thematic land cover information. Statistical classification of the pre-processed remotely sensed data identify clusters (classes) by their spectral similarity (unsupervised classification) or allocate class labels to pixels on the basis of their similarity to a set of predefined spectral classes (supervised classification). There are different statistical similarities and clustering techniques. These can be broken down into approaches where an object can belong to only one class (hard) and those where an object has a membership, however small, to every class (fuzzy). In addition, most approaches treat each pixel as the object to be classified, while few use additional information from some sort of "neighborhood" or patch.

The classification process is dependent on a number of factors and assumptions. First and primarily, that the features of interest on the ground are spectrally similar and can be separated in spectral space. This is not necessarily the case and many workers have reported problems in differentiating between different classes. Second, the process requires some biological knowledge to relate the specifications of the image data to the process of interest. For instance pixel size influences information extraction; woodland is inherently a number of trees interspersed with an understory which itself may be a mixture of bare ground, shrubs, herbaceous vegetation and grass. When the pixel size is small compared to the crown of a tree the spectral response has a bi-modal distribution (tree or understory). If the pixel is a similar size as a tree crown then pixels are tree, understory or both and considerable spectral overlap might be expected with open classes such as grasses. Third, there is an implicit assumption that the different land cover classes can be clustered in spectral space, and the *N* classes desired will be identified by *N* separable clusters (unsupervised classification). Whilst supervised classifications assume that the data on which the classifier is trained adequately characterizes the target classes. Yet land cover class definitions may be determined outside of the laboratory for instance by field survey, and they may not relate to spectral classes (Cherrill and McClean, 1995). Further, a minimum mapping unit (MMU) is often applied to classified data. It defines the lower areal limit for representing homogenous land cover regions. Although the application of a MMU is an additional legacy of cartographic map production to those identified by Fisher (1998), the choice of the MMU will influence the representational detail and spatial pattern of the land cover map (Saura, 2002).

The issues in classification that can be characterized as semantic are addressed by a part of the remote sensing community that is grounded in statistics, geography, biology and ecology. Their activity can be caricatured as applying knowledge of how features on the ground relate to the image specifications and the objectives of the study which are often grounded in policy (Comber et al., 2003). Many aspects such as pixel size, supervised classification training data, and image temporal attributes influence the land cover information that can be derived. The work of the biologist determines how abstract conceptualizations of land cover are specified within classified image data: the ontology of land cover.

The process of statistical land cover classification from remotely sensed imagery as practiced by geographers is parallel to prototypic classification as described by Rosch (1978). Clusters are identified in a reflectance feature space composed of the different image reflectance bands. Typically vegetation categories are defined by their positions in a feature space of bands. Supervised classification proceeds by allocating each pixel to the class to which it is closest in this feature-space. Effectively the distance between the pixel digital numbers and the typical values for each category in each of the selected bands are combined to generate a set of category membership probabilities for each pixel. This is a probabilistic variant of the prototypic approach to categorization that treats each category as a summary description.

1.6 Land cover information treated as data

Digital land cover information is transferred from producers to users. For many disciplines the concept of “Land Cover” provides a useful surrogate with which to describe the landscape. Land cover has been transformed into a universal panacea for land inventory due to the ease of data transfer and the increased use of spatial data in a range of different disciplines. The land cover information becomes a boundary object in the sense of Harvey and Chrisman (1998); at the boundary between responsibilities - land cover information is produced by one group (the producers), and then adopted by a variety of users. The underlying perceptions of the information differ, however, among the various actors according to their disciplinary perceptions. As Hunter (2002) points out, although we may transfer data between databases, “we may find that data in one database does not necessarily have the same

meaning as data carrying the same name in another database, or that data by different names in the two databases actually mean the same thing".

Remote sensing views land cover in terms of spectral properties of objects. Areas of spectral homogeneity are identified and the influence of scale, resolution and classification are generally acknowledged. Analyses of land cover, however, are commonly reported with neither reference to ecological process (Smith *et al.* 2003) nor the ontological meaning of the land cover features, as defined by the epistemology of data processing (Griffiths *et al.*, 2000). In ecology, land cover is defined by the botany of different classes. On the other hand, soil surveys use the presence of different land covers as an indication of the underlying soil type, while landscape ecology is concerned with relating spatial pattern to ecological process (Forman, 1995). Landscape analyses therefore are concerned with how changes in landscape scale, resolution, and classification can have complex consequences for landscape pattern, analysis, and interpretation. However, they are not concerned with the origins of land cover. In GIS land cover is treated as another analytical layer. A false perception of accuracy may be produced as the precision of coordinates in GIS is greater than the accuracy of the spatial data. Computer Scientists, brought into this arena by the advent of GIS and digital mapping, can be caricatured as considering only an object (pixel or vector) with some attributes that may have a class hierarchy (matching their experience from other applications of computer science). In both cases only the class identity is of interest.

As users, all of the above disciplines can be characterized as not understanding the precise meaning of the data in the same way as the data producers nor being able to interpret heuristically commonly found artifacts such as spectral confusions, or boundary issues. Because very few of the stages of land cover information production, and because for land cover there is no agreed data primitive or natural kind, the following scenarios occur:

- Users assume it represents measurement of some agreed phenomenon that is independent of the mapping process;
- Users accept that the land cover information presented is appropriate for their analysis;
- The implicit conceptualizations in land cover datasets are not always understood by the users; they may use them without fully understanding (or even considering) what the land cover information means in terms of the assumptions that underpin it;

- Users treat the derived information as data.

In treating the land cover information as data users are implicitly ascribing different meanings to the information according to their disciplinary constraints, focus, or objectives. That is they impose their own interpretations of what land cover should encapsulate relative to the objects of interest. For instance, landscape ecologists are concerned with the impacts of changes in spatial configuration of the landscape and they use the information as if it were data to support this endeavor. They rarely think in terms of land cover primitives and the nature of the data they are using. In computing science data is commonly considered to represent only data primitives, blocks of which can be aggregated according to need. In short, different users have different conceptualizations of the land cover. In their applications either they assume their disciplinary primitives are recorded by or nest into land cover information or they ignore the problem.

1.7 The social construction of land cover

Geographic data necessarily abstracts from a reality or perception of the reality on the ground, through a social and policy process interfacing between the data, the information and its use. The abstraction process is deeply entrenched in the social and political context of the operatives, indeed some work has described the extent to which land cover information is overtly politically and socially constructed (Hoeschele, 2000; Robbins, 2001; Comber et al. 2003), and results in relativist measures of reality. Relativism is multilayered. Some relativism originates in raw data pre-processing for geometric and radiometric correction, processes so common, universal and uncontested that they are not even reported in the derived thematic products, and often poorly reported even for the image products. A further layer originates from partitioning the data into land cover classes.

The implication of the social construction of land cover data is that different agencies will have their own view of the world due to different social contexts. Social constructionism rejects the notion that knowledge can be divorced from social experience in order to access objectively an external reality. Instead it is necessary to understand the constructions (interests, power relations, etc.) rather than trying to determine 'objective conditions' through more data and better science. If this view is accepted then the question is how real environmental problems are when a plurality of

perspectives exists. Jones (2002) suggests a middle ground in the realism relativism debate: to accept epistemological relativism (we can never know reality exactly as it is), while rejecting ontological relativism (that our accounts of the world are not constrained by nature). This position accepts diverse interpretations of a common reality as meanings rather than truths and sees the real world as being culturally filtered as meanings are constructed, thus avoiding both the naivety of pure realism and the impracticality of pure relativism.

Whilst social construction introduces the question of the relativism of the land cover, the lack of primitives is in parallel with social scientists, which are much more open about the need to discuss "what we are talking about". Perhaps with land cover we need to be more open about the assumptions and underlying meanings of the information we record and classify.

The process of land cover feature identification from remotely sensed data is a series of complex processes. Users may be unaware of the influence that each stage has on how data becomes information. Some may be closer to the caricatured physicist others to the biologist. Decisions about whether to use the information ought to be based on the interaction between the epistemology of the imagery and the ontology of the derived land cover information, in light of the external influences such as policy and the implied uncertainty and risk assessment for their application.

1.8 The CORINE programme (land cover project)

If our environment and natural heritage are to be properly managed, decision-makers need to be provided with both an overview of existing knowledge, and information which is as complete and up-to-date as possible on changes in certain features of the biosphere. To this end, the three aims of the CORINE (Coordination of information on the environment) programme of the Commission European are:

- to compile information on the state of the environment with regard to certain topics which have priority for all the Member States of the Community;
- to coordinate the compilation of data and the organization of information within the Member States or at international level;
- to ensure that information is consistent and that data are compatible.

On 27 June 1985 the Council, on a proposal from the Commission, adopted a decision on the CORINE programme. This Commission work programme concerns 'an experimental project for gathering, coordinating and ensuring the consistency of information on the state of the environment and natural resources in the Community. In order to determine the Community's environment policy, assess the effects of this policy correctly and incorporate the environmental dimension into other policies, we must have a proper understanding of the different features of the environment:

- the state of individual environments,
- the geographical distribution and state of natural areas,
- the geographical distribution and abundance of wild fauna and flora,
- the quality and abundance of water resources,
- land cover structure and the state of the soil,
- the quantities of toxic substances discharged into environments,
- lists of natural hazards, etc.

The land cover project is part of the CORINE programme and is intended to provide consistent localized geographical information on the land cover of the 12 Member States of the European Community. The project is necessary for the following reasons:

- preliminary work on the CORINE information system showed that information on land cover, together with information on relief, drainage systems etc., was essential for the management of the environment and natural resources; information on land cover therefore provides a reference source for various CORINE database projects;

- in all the countries of the Community, the information on land cover available at national level is heterogeneous, fragmented and difficult to obtain.

At Community level, in the CORINE system, information on land cover and changing land cover is directly useful for determining and implementing environment policy and can be used with other data (on climate, inclines, soil, etc.) to make complex assessments (e.g. mapping erosion risks). The benefits of using a single joint project to meet both community and national (or even regional) needs considerably influenced the general features of the land cover project: scale, area of the smallest mapping unit and nomenclature.

Until recently, it was generally assumed that in the long term human activity had little lasting effect on the land thanks to nature's ability to restore itself. This view remained prevalent for a long time despite the fact that farming practices have been causing irreversible damage in certain areas for centuries. Over the last few decades, the effects of certain phenomena have shown that we do need to look after land cover and all its various components. These include: the gradual desertification of certain regions, the rapid disappearance of vast areas of forest, the wholesale of poor farmland, the gradual drying-up of wetlands, and continuous urban development along coastlines, etc. If the aim is to do more than resort to basic emergency action in the face of disaster and instead manage vast areas of land rationally, information on land cover is essential. Yet, despite the urgency and the scale of the problem, confirmed by all the studies, progress in this area is limited and often disappointing.

Those industrialized countries which have devoted considerable resources to producing large-scale maps of national territories and keeping up-to-date inventories and maps of land ownership have never seriously considered the problem of making and updating land cover inventories. This may be because the serious nature of the effects of some of man's actions on the biosphere has only recently been fully understood, or because data compilation and management techniques did not previously lend themselves to this type of operation. As a result, information on land cover has been available only for small areas affected by urban development, agricultural development, major infrastructure projects, etc. Against this background, the CORINE land cover project launched by the Commission of the European Communities sets out to meet a new need and provides support for the Commission in its efforts to use and develop advanced data compilation and management techniques in carrying out its policies. The main classes which are used for the classification of the lands are: artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands and water bodies. The lands are also classified according to other two levels of classification details.

1.9 China 30m-resolution Global Land Cover

On 23rd July 1972, the Earth resource technology satellite, ERTS-1, later renamed as Landsat-1, was launched by the United States. Ten years later, the US launched Landsat-4 carrying onboard the new sensor, Thematic Mapper (TM), with a ground resolution of 30m. The Landsat series of

satellites have acquired approximately 2.9 million scenes of images across the globe. Although successful applications of Landsat data have been reported in tens of thousands of papers and various governmental reports, so far not a single map has been produced on a specific theme for the whole world using Landsat data.
Led by Professor Chen Jun, ISPRS secretary general and chief scientist of the National Geomatics Centre of China, China is mapping the global land cover based primarily on Landsat TM data. The plan is to use Landsat images for circa 2000 and to use data from China's HJ-1 and Beijing-1 satellites in addition to Landsat images for circa 2010. The map product is expected to better serve the needs in Earth system modeling efforts.

The project will produce a series of land cover maps with specific themes - water bodies, wetlands and human settlements - as well as a general land cover map for the world. Over 90,000 training samples and nearly 40,000 test samples have been collected so far in support of the global land cover mapping project. Five algorithms are being assessed to map the world, continent by continent. The five algorithms include the conventional maximum likelihood classifier, the J4.8 classification tree algorithm, as well as random forest, AdaBoost based on J4.8, and support vector machine (SVM). In addition, a classification system that is flexible for expansion, cross-workable to existing ones, has been developed. The current Level I classes contain 10 categories with over 30 Level II subcategories. Initial experiments only involving the use of the six bands of spectral data of TM have been conducted for China, Europe and Africa. SVM, the best performer, achieved an overall accuracy of 66.4%, 60.8% and 68.5%, for China, Europe and Africa respectively, for Level I categories. The image above shows the land cover classification results for Europe. The massive classification for the whole world except Greenland and Antarctica is being implemented on Tsinghua's 120 Tflops supercomputer. This product improves the current global land cover products by one order of precision from 300m to 30m.

Chapter 2

GeoWeb Services

GeoWeb 2.0 has rapidly changed the way in which information, and particularly geographic information, is produced, shared and consumed. The arena of geospatial applications, whose access was formerly restricted to highly-trained experts of mapping agencies, governmental institutions and universities, has been suddenly entered by the large and heterogeneous community of neogeographers. The considerable flourishing of mobile sensors, including human sensors volunteering geographic information, has turned the GeoWeb into a much more complex framework featuring new actors and new contents. At all levels, ranging from the administrative up to the social and academic, a strong need emerged to integrate Web Mapping into almost any spatially-related application. It was in this context that Spatial Data Infrastructures (SDIs) started to play a crucial role in providing geospatial data maintenance, sharing, access and usage. This chapter presents a structured overview of the available technologies to perform such operations, which enable current SDIs to fit the intricate nature of GeoWeb 2.0. The following discussion is thus focused on GeoWeb applications, i.e. software tools allowing accessing geographic data and functionalities over the Internet, and GeoWeb services, i.e. programs able to serve those data and functionalities. According to the purpose of the present work the attention is specially placed on Free and Open Source Software (FOSS), whose nature allows creating fully-customized products according to the needs. Therefore, after an introduction on the most relevant GeoWeb services standards delivered by the Open Geospatial Consortium (OGC), an overview of the main FOSS tools is offered.

2.1 Web services for geospatial interoperability

Geographic data have become a vital source of information for decision-makers in a number of applications at the local, regional and global levels, e.g. crime management, business development, flood mitigation, community land use and disaster recovery. However the potential of geographic data cannot be fully exploited together with the associated infrastructures, the so-called Spatial Data Infrastructures or SDIs (Nebert, 2004). They denote "a coordinated series of agreements on technology standards, institutional arrangements, and policies that enable the discovery and use of geospatial information by users and for purposes other than those it was created for". The term infrastructure highlights the concept of a reliable, supporting environment providing a basis for geographic data access, evaluation and application within all the levels in society (government, commercial sector, non-profit sector, citizen community and academia). The massive literature on SDIs allows to distinguish five main components: spatial information, technologies (i.e. software and hardware), laws and policies, people (i.e. data providers, service providers and users), and standards for data acquisition, representation and transfer.

Initiatives aimed at increasing the availability and accessibility of geographic information through the development of SDIs have been common since the last decade of the 20th century, by the mid of which Masser (1999) identified at least 11 available SDIs at different stages of development. The establishment in 1996 of the Global Spatial Data Infrastructure (GSDI) Association pushed the worldwide diffusion of SDIs, both at national and international levels, by promoting best practices and sharing experiences (Craglia et al., 2008). Major examples of SDIs include the US National Spatial Data Infrastructure (NSDI) established in 1994 and the Canadian Geospatial Data Infrastructure (CGDI) born in 2001. In Europe a legal framework adopted in 2007 established an Infrastructure for Spatial Information in Europe (INSPIRE) which was built on the SDIs of the 27 Member States of the European Union (European Parliament and Council, 2007). Another international example that is worth mentioning is the United Nations Spatial Data Infrastructure (UNSDI) initiative, whose vision, strategy and institutional governance framework were developed in 2006.

The practical implementation of SDIs, which has been among others recognized as a key element for achieving the new vision of Digital Earth (Craglia et al., 2008), requires a specific range of software. In few words, this software must enable the discovery and delivery of geospatial data from a repository (i.e. a collection of spatial datasets stored on one or multiple servers) via one or moreWeb services. Therefore, according to Steiniger and Hunter (2012) the basic software components of an SDI include (see also Figure 3.1):

- a software client, which can display, query and analyze geospatial data;
- a catalogue service for the discovery, browsing and querying of metadata or spatial services, spatial datasets and other resources;
- a spatial data service, which enables the delivery of data and/or processing services (e.g. datum and projection transformations) via the Internet;
- a spatial data repository;
- GIS software (desktop or client) that allows data creation and maintenance.

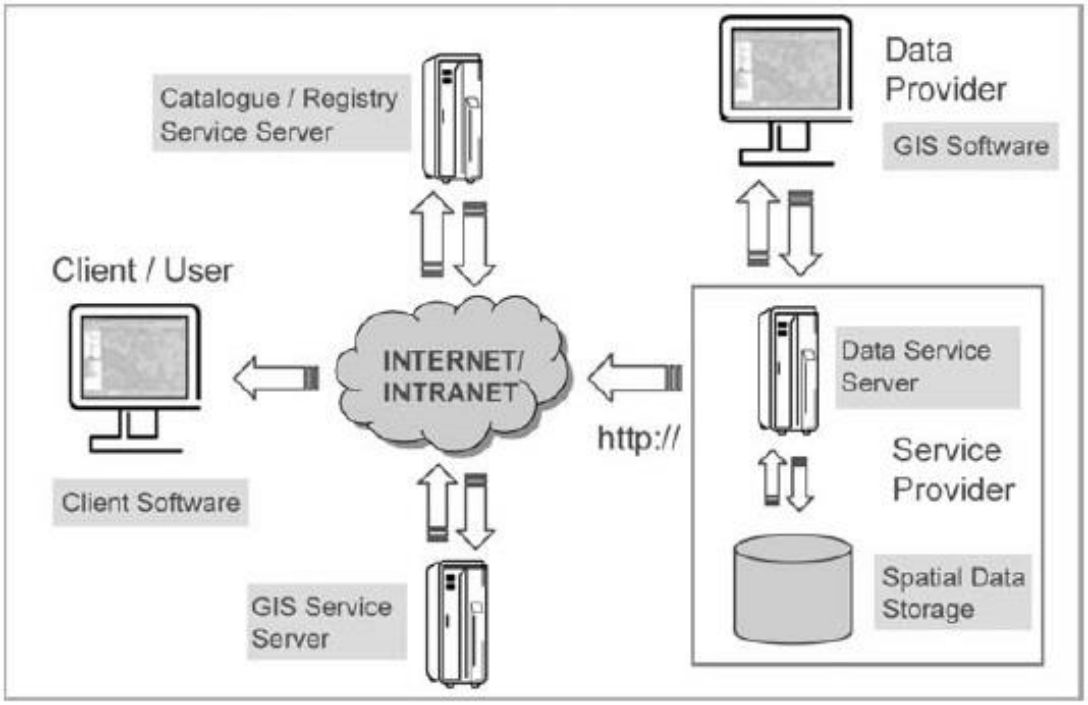


Figure 2.1: Software components needed for implementing an SDI.

Allowing these software components to properly interact each other, i.e. exchange data via a common set of formats, read and write the same file formats, and use the same protocols, means making the whole system interoperable. Concerning software, the generally-accepted technical definition of interoperability was provided by the Institute of Electrical and Electronic Engineers (IEEE) as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged". It goes without saying that the interoperability of services to discover, view, access and integrate geospatial information represents a key point of SDIs and requires a well-defined standardization frame.

The cornerstones of most of the current SDIs are constituted by several technical standards delivered by two organizations: the International Organization for Standardization (ISO), particularly its Technical Committee and the Open Geospatial Consortium (OGC). In general these standards describe communication protocols between data servers, servers which provide spatial services, and client software that request and display spatial data (Steiniger and Hunter, 2012). ISO and OGC standards are in turn dependent on other industry standards, especially those developed by theWorldWide Web Consortium (W3C) for data dissemination (e.g. HTML, XML and SOAP), which therefore should also be considered.

Established in 1994, ISO/TC 211 covers the areas of digital geographic information and geomatics by defining a structured set of standards concerning georeferenceable objects and phenomena. The work of ISO/TC 211 is strongly coordinated with the action of national standardization committees and many other international entities, including the OGC, UN agencies, professional bodies (e.g. the International Federation of Surveyors) and sectorial bodies (e.g. the Digital Geographic Information Working Group). ISO/TC 211 standards, numbered starting from 19101, specify methods, tools and services for geospatial data acquisition, access, management, presentation, processing, analysis and transfer. Among them it is worth mentioning the ISO Standard 19115 Geographic Information – Metadata (International Organization for Standardization, 2003) which defines a schema for describing geographic data (including contents, spatio-temporal purchases, data quality, access and right to use), and the ISO Standard 19119 Geographic Information – Services (International Organization for Standardization, 2005) which identifies and describes the architecture patterns for services interfaces used for geographic information.

2.2 OGC standards

The interoperability of systems through services has been also the major focus of the Open Geospatial Consortium, an international industry consortium established in 1994 which is currently composed of 474 worldwide organizations including commercial companies, government agencies, non-profit corporations and universities. All of them participate in a consensus process aimed at developing publicly-available, interoperable interface standards which make geospatial information and services accessible and useful within all kinds of Web applications. Named also Open GIS Consortium till 2004, OGC serves as a global forum for the collaboration of both users and developers of spatial data products and services, pursuing its mission to advance the development of international standards for geospatial interoperability. OGC's strategic goals include to lead worldwide in the creation of geospatial standards; to provide standards to the market and accelerate its assimilation of interoperability; to facilitate the adoption of open, spatially-enabled reference architectures in worldwide enterprise environments; and to advance standards with the purpose of favoring new markets and applications for geospatial technologies. It is worth mentioning that, together with a Standards Program and an Interoperability Program, which are focused on standards development, approval and acceptance, OGC features also a Compliance Program with the goal of providing resources, tools and policies (e.g. an online free testing facility and a process for certification and branding) for improving software implementation's compliance with the developed standards.

OGC standards, which represent the main product of the consortium, consist of technical documents detailing interfaces or encodings. They have been created by the OGC members to address specific interoperability issues, and they are used by developers to build open interfaces and encodings into their products and services. Ideally, the components of products or online services implementing OGC standards should properly work together without further debugging. A full and updated list of the more than 30 currently-existing OGC standards is available at http://www.opengeospatial.org/standards. All the standards are available to the public at no cost together with their supporting documentation. For the sake of the present discussion it is not useful to describe all OGC standards, many of which are relevant for specific applications but feature no interest within this work.

2.2.1 Web Map Service (WMS)

The Web Map Service (WMS) is an OGC data delivery standard, i.e. it specifies the interaction between a software client requesting geospatial data and a data service providing those data via the Internet. Together with the WFS (subsection 2.1.1.2) and the WCS (subsection 2.1.1.3), WMS constitutes the so-called OGC Web Services (OWS), i.e. the set of OGC standards created for use in World Wide Web applications. Developed and first published by the Open Geospatial Consortium in 1999 (Scharl and Tochtermann, 2007), the WMS Interface Standard provides a simple HyperText Transfer Protocol (HTTP) interface for requesting georeferenced map images from one or more distributed geospatial databases. Therefore the WMS specifications do not concern the real geospatial data, but rather the portrayals of those data in the form of digital image files suitable for display on computer screens. WMS operations can be invoked using a standard Web browser by submitting ad hoc requests in the form of HTTP Uniform Resource Locators (URLs). The standard defines three main operations, namely (Open Geospatial Consortium, 2006):

- GetCapabilities (mandatory): returns service-level metadata;
- GetMap (mandatory): returns a map image with well-defined parameters;
- GetFeatureInfo (optional): returns information on specific map features.

The purpose of the GetCapabilities operation (mandatory for whatever WMS provider) is to obtain service metadata, which is a machine-readable (and humanreadable) description of the server's information content and acceptable request parameter values. The response to a GetCapabilities request shall be a well formatted XML document which provides indication of e.g. the available geographic information contents (called layers), their description, representation style, reference system and geographic bounding box. A basic WMS shall also support the GetMap operation, whose response consists of a map image. The Uniform Resource Locator (URL) of a GetMap request specifies the geographic bounding box, size (i.e. width and height) and format of the desired output map, the geographic information (i.e. the layers) to be served, its reference system and representation style. WMS-produced maps are generally rendered in a pictorial format such as PNG, GIF or JPEG, and only occasionally as vector-based graphical elements in Scalable Vector Graphics (SVG) or Web Computer Graphics Metafile (WebCGM)

formats. The use of image formats supporting transparent background (e.g. PNG and GIF) makes the underlying maps generated from a multiple-layer WMS GetMap request visible. Layer styles can be defined through the OGC SLD specification (Subsection 2.1.1.4) if the specific WMS serving the data is also SLD-enabled.

Unlike GetCapabilities and GetMap, GetFeatureInfo is an optional WMS request and it is only supported for the layers defined as queryable within the service. This request is designed to provide the software client with additional information about the features of the map returned from a GetMap request. More in detail, this information is returned to the user when clicking on a point of the map which corresponds to a particular layer. Within SLD-enabled WMSs other optional operations, which are specifically related to the layers representation styles, are also available (Open Geospatial Consortium, 2007a). These operations are described in Subsection 2.1.1.4.

2.2.2 Web Feature Service (WFS)

The Web Feature Service (WFS) is an OGC data delivery standard which takes the next logical step from the simple WMS by defining interfaces for operations of data access and manipulation. In other words, WFS interfaces (which use again HTTP as the distributed computed platform) enable Web users and services to query, style, edit and download the real geospatial data, i.e. the feature information which stays behind a simple WMS map image. Within WFSs geospatial data features must be encoded in the Geography Markup Language (GML), an OGC XML-based specification that enables the storage, transport, processing and transformation of geographic data. Geography Markup Language (GML) encoding was designed to facilitate the implementation of the WFS standard as well as to increase interoperability between WFS servers. WFS data manipulation functionalities include the ability not only to get and query features based on spatial and non-spatial constraints, but also to create, delete and update feature instances. More in detail, the main operations defined by the OGC standards are the following:

- GetCapabilities (mandatory): returns service-level metadata;
- DescribeFeatureType (mandatory): returns feature types description;
- GetFeature (mandatory): returns requested features;
- Transaction (optional): edits features (i.e. creates, updates and deletes);

- LockFeature (optional): prevents feature editing through a long-term lock.

A basic WFS implements only the GetCapabilities, DescribeFeatureType and GetFeature operations. Conversely, a transaction WFS supports also the Transaction operation and, optionally, even the LockFeature operation. Similarly to a WMS, also a WFS must be able to describe its capabilities. The GetCapabilities operation generates an XML metadata document specifying which feature types the service can provide, and which operations are supported on each feature type. The function of the DescribeFeatureType operation is instead to generate a schema description of feature types served by the WFS. This description defines how the WFS expects feature instances to be encoded in input, and how feature instances will be generated in output (e.g. in response to a GetFeature request). The purpose of the WFS GetFeature operation is precisely to service requests to retrieve feature instances using GML. The client should also be able to specify which feature properties to fetch, and to constrain the query spatially and non-spatially.

Transaction WFS interfaces enable also client applications to alter the state of Web-accessible feature instances by means of data transformation operations, i.e. insert, update and delete. When a transactional request has been completed, the WFS generates an XML response document indicating the completion status of the transaction. Finally the LockFeature operation (which, if available, must be advertised in the capabilities document) allows preventing a feature from being edited through a persistent feature lock. This is particularly useful during transaction requests for feature updates, as in principle there is no guarantee that, while a feature is being modified by a client, another client does not come along and update the same feature. Therefore the LockFeature operation forces a mutually exclusive data access, i.e. no transaction can act on a data feature if a transaction on that feature is already in progress. Consistency is assured by a long-term feature locking, because network latency makes locks last relatively longer than the native database locks.

2.2.3 Web Coverage Service (WCS)

The last OGC data delivery standard belonging to the OWS family is Web Coverage Service (WCS). The WCS allows electronic retrieval of geospatial information as coverages, i.e. raster data representing space&time varying

phenomena which are accessed in forms that are directly useful for client-side rendering (e.g. as input into scientific models).

As WMS and WFS, also WCS service instances allow clients to discovery and interrogate data according to spatial constraints and other query criteria. However a clear difference exists compared to both WMS and WFS. Whilst the output of WMS is a portrayal of geospatial data in the form of a static map image, the WCS provides the real data together with their detailed descriptions and original semantics, which can be interpreted and extrapolated and not just portrayed. With respect to WFS, which returns the "source code" of the map as vector data, one can think to WCS as its analogue for the raster case. WCS coverages represent phenomena which relate a spatio-temporal domain to a range of properties. The WCS suite is organized as a Core, which any WCS implementation must support, and a set of possible extensions which define additional functionalities. Neglecting in this discussion the extensions, the WCS Core interface (Open Geospatial Consortium, 2010b) specifies the following operations:

- GetCapabilities (mandatory): returns service-level metadata;
- DescribeCoverage (mandatory): returns a full coverage description;
- GetCoverage (mandatory): returns requested coverage.

As already seen for WMS and WFS, the GetCapabilities operation allows a WCS client to retrieve an XML-encoded description of the service metadata and the coverages offered by a WCS server. In the same way, the DescribeCoverage operation allows a WCS server, which receives a request with a list of coverage identifiers, to return an XML document containing the description of the requested coverages (e.g. their space and time domain, reference system, metadata and available formats). Finally the GetCoverage operation delivers a requested coverage (or a part of it, identified through a subset space and time domain) in one of multiple formats, both image formats (e.g. JPEG, PNG, GIF and TIFF) and georeferenced formats (e.g. GeoTIFF and ArcGrid).

2.2.4 Styled Layer Descriptor (SLD)

The Styled Layer Descriptor (SLD) defines an encoding language which extends the WMS standard to allow user defined symbolization of geospatial features. SLD is therefore an OGC data format standard (like the already

mentioned KML and GML) and it addresses the need for users and software to be able to control the visual portrayal of geospatial data. As a matter of fact, standard WMSs are able to provide users with a predefined choice of layer styles, but: a) they can tell the users only the name of each style, thus preventing them to know in advance what the layer portrayal will look like on the map; and b) users have no way of defining their own style. The SLD is therefore the styling language, based on a structured XML encoding, which can be used to portray the output of WMS, WFS and WCS servers. By way of example, the SLD allows to style data features differently depending on the visualization scale or on the value of some attribute (e.g. roads can be styled as lines with different colors and width according to their typology, e.g. highways, four-lane roads and two-lane roads). The OGC SLD specifications, which define how a WMS can be extended to allow user-defined styling, are described in two documents. An SLD-enabled WMS shall first of all provide the two mandatory WMS operations described in Subsection 2.1.1.1, i.e. GetCapabilities and GetMap. The response of a GetCapabilities request is now extended by an element defining the SLD capabilities (i.e. which styles are available for each served layer), while a GetMap request allow clients to specify the style to be used for portraying layers. Two other operations are defined:

- DescribeLayer (optional): indicates the WFS or WCS to retrieve additional information about the layer;
- GetLegendGraphic (optional): returns an image depicting the map's legend.

The DescribeLayer operation bridges the gap between the WMS concepts of layers and styles and the WFS/WCS concepts of feature (subsection 2.1.1.2) and coverage (subsection 2.1.1.3). In fact, to define an SLD styling it is required to know the structure of the feature or coverage data to be styled. Therefore the DescribeLayer operation allows clients to obtain the feature/coverage-type information (given by the DescribeFeatureType and DescribeCoverage operations, respectively) for a named layer, by routing the clients to the appropriate service (WFS or WCS). Finally the GetLegendGraphic operation provides a mechanism for generating images of legend graphics based on the layers' user-defined SLD styles. A GetLegendGraphic request should thus indicate the layer and the style for which to produce the legend graphic and the size and format of the legend image to be generated.

2.2.5 Catalog Service for the Web (CSW)

The last OGC standard presented in this overview is the CSW (Catalog Service for the Web), which is used for exposing a catalogue of geospatial records over the Internet. CSW is the profile of the OGC Catalog Service (Open Geospatial Consortium, 2007b), which defines common interfaces between clients and catalogue services for the discovery and retrieval of spatial data and services metadata over HTTP. More in detail, catalogue services can publish and search metadata (i.e. series of descriptive information) about geospatial data, services (e.g. WMS) and other related resources. Catalogue services shall also support the query and discovery of metadata, and in many cases also the invocation or retrieval of the metadata referenced resource, for further use or processing by both humans and software.

The CSW standard defines the metadata format only as an XML-based encoding, specifying that whatever data profile is used (e.g. the FGDC or the Dublin Core) it must be consistent with the core metadata elements defined by ISO 19115 (see Section 2.1) and its XML implementation given by ISO/TC 19139 Geographic information – Metadata – XML schema implementation. Service metadata elements should instead be consistent with ISO 19119 (see Section 2.1). The CSW operations are the following (Open Geospatial Consortium, 2007b):

- GetCapabilities (mandatory): returns service-level metadata;
- DescribeRecord (mandatory): returns some info about the model of records;
- GetDomain (optional): returns the range of values for a given record;
- GetRecords (mandatory): search for records and returns record IDs;
- GetRecordById (mandatory): returns records specified by their IDs;
- Transaction (optional): create/edit/delete metadata records by "pushing" them to the server;
- Harvest (optional): create/update metadata by asking the catalogue server to "pull" them from somewhere.

The operations can be classified in three classes. The first one includes the so called *service operation*, i.e. the usual GetCapabilities request that a CSW client may use to query the service and determine its capabilities. The response is again an XML document containing service metadata about the CSW server. To the second class belong the so called *discovery operations*

that a client may use to determine the information model of the catalogue and to query catalogue metadata records. The mandatory DescribeRecord operation allows a CSW client to discover elements of the information model supported by the catalogue service. Through the optional GetDomain operation, a client can obtain runtime information about the range of values of a metadata record element. Finally, the mandatory GetRecords and GetRecordsById operations allow a client to search and retrieve the representation of catalogue metadata records.

At last, the class of the so called *discovery operations* allows a CSW client to create or change metadata records in the catalogue. The *Transaction* operation defines an interface enabling CSW clients to create, modify and delete catalogue metadata records. A locking interface is not defined by the standard, thus requiring that concurrent accesses to the catalogue records are managed by the underlying repository. While *Transaction* "pushes" data into the catalogue, the *Harvest* operation "pulls" data into the catalogue. In other words it only references the data to be inserted or updated in the catalogues, and then it is a job of the CSW to resolve the reference, fetch data and process it into the catalogue.

2.2.6 Web Processing Service (WPS)

The Open GIS Web Processing Service (WPS) Interface Standard provides rules for standardizing how inputs and outputs (requests and responses) for geospatial processing services, such as polygon overlay. The standard also defines how a client can request the execution of a process, and how the output from the process is handled. It defines an interface that facilitates the publishing of geospatial processes and clients' discovery of and binding to those processes. The data required by the WPS can be delivered across a network or they can be available at the server.

WPS has the following properties:

- Inputs can be web-accessible URLs or embedded in the request
- Outputs can be stored as web-accessible URLs or embedded in the response
- It supports multiple input and output formats.
- It supports long-running processes
- It supports SOAP and WSDL

WPS defines three operations: GetCapabilities which returns service-level metadata, service description, access description, brief process descriptions, DescribeProcess which returns a description of a "process" including its inputs and outputs, and Execute which returns the output of a "process".

Chapter 3

Geospatial Clients

Geospatial web mapping clients play a significant role in Geoportals of Spatial Data Infrastructures (SDI) allowing the visualization of spatial data from several sources. Likewise, these clients may be part of web-based Geographic Information Systems (GIS) applications, in which users can directly interact with SDI services, visualize, query and integrate them with local data and GIS tools. There exists a wide variety of free and open source software (FOSS) projects that make the creation and configuration of Web mapping clients easier. There is a wide collection of Geospatial web mapping clients capable to access Open Geospatial Consortium (OGC) web services, and examines some of their most relevant properties. A comparison of free and open source geospatial web clients addressed more than 40 products (Carrillo, 2012), including those which have been abandoned or without recent releases. In this chapter, the most popular clients and those which are used in the Open Land Map project are discussed.

The graph (see Figure 3) shows that most projects revolve around two paradigms: UMN MapServer and OpenLayers. Clients using UMN MapServer as a basis were created years ago taking advantage of the features that this client provides: map scale, map reference, basic navigation tools, identification of geographic objects and its Application Programming Interface (API) called MapScript, which has been implemented in different programming languages such as PHP, Python, Java, Perl and Ruby. On the other hand, a more recent generation of clients uses OpenLayers due to its optimal performance in rendering tasks on the web and to the wide variety of

data sources it supports. Several companies contribute to its development and projects like MapBuilder have come to an end to accelerate its progress, which makes it the state-of-the-art library for building web mapping applications. Nowadays, even projects with their own rendering component are adopting or at least supporting OpenLayers in order to avoid duplicating efforts in an area where there is already a dominant one. It should be noted that some projects use Flash/Flex for building Rich Internet Applications (RIAs) such as Flamingo, worldKit, OpenScales and Geoide, providing a pleasanter experience for users intending to interact with on-line maps. Finally, the latest generation of clients are built upon HTML5, taking advantage of significant improvements in interaction with multimedia and vector content, this time native (for web browsers) rather than through third party plug-ins. Leaflet and ReadyMap web SDK are examples of projects using HTML5-related technologies, the latter one, based on WebGL, even allowing 3D globes to be rendered with pure JavaScript.

Figure 3: Overview of free and open source Geospatial Web clients (Carrillo, 2012).

3.1 OpenLayers 2D

The most well-known and worldwide used geospatial Web client is OpenLayers, a JavaScript library providing a pool of functions to easily insert interactive maps in any web page. The development of OpenLayers was started in 2005 by the US Company MetaCarta with the purpose of building an open source equivalent of Google Maps. In 2007 it became an official OSGeo project and is currently supported by a team of developers from around the world. OpenLayers provides an API for building rich Web Mapping applications running on most of modern Web browsers without server-side dependencies. OpenLayers is released under the 2-Clause BSD License. OpenLayers implements the OGC industry-standard methods for geographic data access, e.g. WMS, WFS (including WFS-T) and WCS and it also supports data formats such as GeoRSS, KML, GML and GeoJSON. OpenLayers philosophy is to separate map tools from map data, so that all the tools can operate on all the data sources. Among its many functionalities, it is worthwhile to note the integration of OpenStreetMap, Google Maps, Yahoo! Maps and Bing Maps, whose availability allows to greatly enhance map mash-ups. In addition OpenLayers includes the capability to manage touch-screen commands, thus providing broad support also for mobile devices.

The popularity of OpenLayers is demonstrated by the huge number of related online tutorials (e.g. http://workshops.boundlessgeo.com/openlayers-intro) and books (Hazzard, 2011). Conversely it is usually combined with or integrated into other libraries or frameworks in order to create richer and advanced geospatial Web clients.

3.2 MapServer

MapServer is an Open Source geographic data rendering engine written in C. Beyond browsing GIS data, MapServer allows you create “geographic image maps”, that is, maps that can direct users to content. For example, the Minnesota DNR Recreation Compass provides users with more than 10,000 web pages, reports and maps via a single application. The same application serves as a “map engine” for other portions of the site, providing spatial context where needed. MapServer was originally developed by the University of Minnesota (UMN) ForNet project in cooperation with NASA, and the Minnesota Department of Natural Resources (MNDNR). Later it was hosted by

the TerraSIP project, a NASA sponsored project between the UMN and a consortium of land management interests. MapServer is now a project of OSGeo, and is maintained by a growing number of developers from around the world. It is supported by a diverse group of organizations that fund enhancements and maintenance, and administered within OSGeo by the MapServer Project Steering Committee made up of developers and other contributors.

It supports popular scripting and development environments like PHP, Python, Perl, Ruby, Java, and .NET. It also support many Open Geospatial Consortium (OGC) standards e.g. WMS (client/server), non-transactional WFS (client/server), WMC, WCS, Filter Encoding, SLD, GML, SOS, OM.

3.3 MapFish

MapFish is a flexible and complete framework for building rich web-mapping applications. It emphasizes high productivity, and high-quality development. MapFish is based on the Pylons Python web framework. MapFish extends Pylons with geospatial-specific functionality. For example MapFish provides specific tools for creating web services that allows querying and editing geographic objects. MapFish also provides a complete RIA-oriented JavaScript toolbox, a JavaScript testing environment, and tools for compressing JavaScript code. The JavaScript toolbox is composed of the ExtJS, OpenLayers and GeoExt JavaScript toolkits. MapFish is compliant with the Open Geospatial Consortium (OGC) standards. This is achieved through OpenLayers or GeoExt supporting several OGC norms, like WMS, WFS, WMC, KML, GML etc. MapFish is open source, and distributed under the BSD license.

MapFish is a project of the Open Source Geospatial Foundation (OSGeo Foundation), OSGeo‘s mission is to support and build the highest-quality open source geospatial software. The MapFish framework is built around an open HTTP-based protocol, allowing various interoperable implementations. In addition to the reference implementation provided by the Python/Pylons-based framework, two other implementations are currently available: a Ruby/Rails plugin (GPLv3), and a PHP/Symfony plugin (BSD).

3.4 ArcGIS

ArcGIS is a geographic information system (GIS) for working with maps and geographic information. It is used for: creating and using maps; compiling geographic data; analyzing mapped information; sharing and discovering geographic information; using maps and geographic information in a range of applications; and managing geographic information in a database.

ArcGIS desktop conducts spatial analysis and offers hundreds of tools for performing spatial analysis. These tools allow turning data into actionable information and automating many of GIS tasks. It helps to manage the data more efficiently with support for more than 70 data formats, easily integrate all types of data for visualization and analysis making available extensive set of geographic, tabular, and metadata management, creation, and organization tools.

By using the desktop application it is easy to create maps producing high-quality maps without the hassles associated with complex design software and taking advantage of a large library of symbols and Simple wizards and predefined map templates. It also provides tools to manipulate data with a minimum number of clicks and automate editing workflow.

3.5 QuantumGIS

QGIS (as known as Quantum GIS) is a cross-platform free and open source desktop geographic information systems (GIS) application that provides data viewing, editing, and analysis capabilities. QGIS allows users to create maps with many layers using different map projections. Maps can be assembled in different formats and for different uses. QGIS allows maps to be composed of raster or vector layers. Typical for this kind of software the vector data is stored as point, line, or polygon-feature. Different kinds of raster images are supported and the software can perform geo-referencing of images.

QGIS provides integration with other open source GIS packages, including PostGIS, GRASS, and MapServer to give users extensive functionality. Plugins, written in Python or C++, extend the capabilities of QGIS. There are plugins to geocode using the Google Geocoding API, perform geoprocessing (fTools) similar to the standard tools found in ArcGIS, interface with

PostgreSQL/PostGIS, SpatiaLite and MySQL databases, and use Mapnik as a map renderer. QGIS is maintained by an active group of volunteer developers who regularly release updates and bug fixes. As of 2012 developers have translated QGIS into 48 languages and the application is used internationally in academic and professional environments.

3.6 Leaflet

A rather new FOSS4G product which has gained much interest and success is Leaflet (http://leafletjs.com), a lightweight JavaScript library (all the code weights about 33 KB) for building mobile-friendly interactive maps. It has been developed since 2010 by Vladimir Agafonkin with a team of dedicated contributors. Current stable version is 0.7.3, released in November 2013 and updated in May 2014. Leaflet is distributed under a custom open source license (https://github.com/Leaflet/Leaflet/blob/master/LICENSE).

What make Leaflet different from traditional geospatial Web clients is its simplicity, performance and usability. Unlike OpenLayers and QGIS Web Client, it runs with the same version on both desktop and mobile platforms, by taking advantage of HTML5 and CSS3 on modern browsers while still being accessible on older ones. The purpose of Leaflet (at least up to the present day) is not to provide all the possible client-side functionalities, but rather to satisfy the basic needs of the vast majority of Web Mapping creators. Besides advanced interaction and visual features, it provides support for WMS and GeoJSON layers, vector layers, tile layers, markers and popups. Leaflet is currently used by numerous organizations and projects.

3.7 MapGuide

MapGuide Open Source is a web-based platform that enables users to develop and deploy web mapping applications and geospatial web services. MapGuide features an interactive viewer that includes support for feature selection, property inspection, map tips, and operations such as buffer, select within, and measure. MapGuide includes an XML database for managing content, and supports most popular geospatial file formats, databases, and standards. MapGuide can be deployed on Linux or Windows, supports Apache and IIS web servers, and offers extensive PHP, .NET, Java, and JavaScript APIs

for application development. MapGuide Open Source is licensed under the LGPL license.

MapGuide was first introduced as Argus MapGuide in 1995 by Argus Technologies in Calgary, Alberta. Autodesk acquired Argus Technologies in the fall of 1996 and within a few months the first release under the Autodesk brand was introduced, Autodesk MapGuide 2.0. The software progressed through a number of releases leading up to the current Autodesk MapGuide 6.5. To this day MapGuide 6.5 and previous releases are known for ease of deployment, rapid application development, data connectivity, scalability, and overall performance.
Despite its success the MapGuide 6.5 architecture has some inherent limitations. To this day most MapGuide applications rely upon a client Plug-in, ActiveX Control, or Java Applet with much of the application logic written in JavaScript using the APIs offered by the client-side plug-in. All spatial analysis is performed client-side on rendered graphics rather than on the underlying spatial data. And finally the server platform is very Windows centric. In the spring of 2004 a dedicated team of developers began work on what is now MapGuide Open Source. The goals were simple, retain all of the best aspects of MapGuide 6.5 while also meeting the goals set out above. The result is MapGuide Open Source. Autodesk released MapGuide Open Source under the LGPL in November 2005, and contributed the code to the Open Source Geospatial Foundation in March 2006. The mission of MapGuide is to create, as a community, the leading international web-based platform for developing and deploying web mapping applications and geospatial web services. The goals for the platform are as follows:

- use the service-oriented architecture pattern
- fast, scalable, and cross platform
- make use of open source components
- support rich access to spatial data both vector and raster
- provide a full suite of spatial analysis
- produce visually stunning cartographic maps
- include viewers that work within any browser on any platform
- provide the highest degree of map interactivity possible
- conform with open standards
- offer a single API that works with both vector and raster based client-side viewers

3.8 Geomajas

Geomajas is a free and open source GIS framework which seamlessly integrates powerful server side algorithms into the web browser. The focus of Geomajas is to provide a platform for server-side integration of geospatial data (be it through GeoTools or Hibernate), allowing multiple users to control and manage the data from within their own browsers. In essence, Geomajas provides a set of powerful building blocks, from which the most advanced GIS application can easily be built. What makes Geomajas unique is its strong server side focus. The processing, styling, filtering, caching, etc. of geospatial data always happens within a secured context. All this makes Geomajas applications incredibly scalable and performing, keeping the client a real thin client.

3.8.1 Core features and project roadmap

Geomajas is developed under the GNU Affero general public license (APGL), and the core features are:

- Integrated client-server architecture
- Geometry and attribute editing
- Custom attribute definitions
- Advanced querying capabilities (CQL)
- Out-of-the-box security
- Extensible plug-in mechanism
- Multiple front-end technologies
- Cross browser support, without the need for browser plug-ins

A configuration GUI will be delivered which will further improve the usability of the system. With this, users will have access to personalized and dynamic configuration capabilities. Serious effort will go into supporting more OGC and INSPIRE standards, such as WCS, CS-W, WPS, and so on. In addition to the security features, extra encryption will be added for transfer between client and server on one hand, and data transfer from the Geomajas server to its original data source. A data versioning will provide a general way to add data versioning to the vector layer model. The idea is to have the ability to not only apply versions to changes, but to also provide a way of retracing a past condition. Users want to know what their maps looked like a year or two

years ago, in order to compare with the current state. Geomajas project also will move in the direction of 3D support at some point in the future.

3.9 NASA World Wind 3D

Released under the NASA Open Source Agreement, World Wind is written in multi-platform Java language and thus runs on Linux, Windows and Mac OS X. First formal release was 1.2 in July 2011; World Wind is available as a highly extensible SDK which allows a full customization of the developed applications. It can be run either as a desktop Java application, or into a Web browser as a Java applet or a Java Web Start application. It integrates both Swing and Abstract Window Toolkit (AWT) Java toolkits and the Java Open Graphics Library (JOGL) for maximizing graphics capabilities. World Wind accommodates any desired data format and provides open-standard interfaces to GIS services and databases. It can be deployed as a WMS server and enables to locate on or above the globe both 2D objects (e.g. lines, polygons, markers, callouts, and multimedia viewers) and 3D objects built up from geometric primitives (e.g. parallelepipeds, spheres, and extruded polygons).

A rich collection of spatial datasets is natively provided by World Wind. This includes both satellite imagery with multiple resolutions (e.g. BlueMarble, SGS Orthophoto/Urban Area Orthophoto, and Microsoft Virtual Earth Imagery) and standard Digital Elevation Models such as SRTM, ASTER, and USGS National Elevation Dataset (NED). Both imagery and DEMs are dynamically served by NASA and USGS WMS servers. However World Wind allows users to access any other OGC-compliant WMS, serving both georeferenced images or data to be projected on the globe, and also DEMs to be superimposed on the geoid model implemented within the platform. Full control of the terrain model strongly distinguishes World Wind from the majority of virtual globes.

Chapter 4

User Participation in Geographic Data Production

Fostered by the burgeoning development of Web 2.0 and GeoWeb 2.0 technologies, the phenomenon of Volunteered Geographic Information (VGI) has marked a profound transformation on how geographic knowledge is produced and circulated. Citizens have become a new important player in the mapping scene by contributing georeferenced information about the Earth's surface and near-surface. At first, this chapter outlines the nature and the multiple characteristics of crowd sourced geographic information. An explanation of the basic related terminology, a comparison with traditional mapping, a review of the most well-known VGI applications and a glance at the quality of crowd sourced data are then offered.

4.1 Crowdsourcing geographic information

The production and dissemination of geographic information has known an increasing growth over the last decades, with an unprecedented boost brought by the popularizing of the Internet and the advent of the Web. The crucial technological developments occurred since the early 1990s, when geospatial information started to be delivered on the Web, have ushered in what (Goodchild et al. 2007) have named the post-modern era of geographic information production. This modern epoch was characterized by the worldwide birth of Spatial Data Infrastructures (SDIs), defined by the Mapping

Science Committee of the US National Research Council as the aggregation of agencies, technologies, people and data, that together constitute a nation's mapping enterprise.

Despite SDIs have addressed a number of issues such as semantic interoperability, spatial data sharing and legal issues (Onsrud, 2007), a concern for the basic supply of geographic information, and the processes by which it is acquired and compiled, was largely missing.
Furthermore, (Estes and Mooneyhan 1994) have called attention to what they termed the *mapping myth*, i.e. the mistaken belief that the world was well mapped and that maps were constantly being updated and becoming more and more accurate over time. On the contrary they argued that mapping, which is a governmentally sponsored activity, had been in decline in many countries since the middle of the 20th century, and that few efforts existed to improve the available maps.

However, the general failure of an accurate mapping was only partially due to the costly and labour intensive actions that governments had to sustain to update (or replace) their geospatial datasets. As a matter of fact, an essential limitation intrinsic to the traditional mapping practices was the inability to extract meaningful kinds of geographic information not only from the massive collection of available maps, but also from the constant flow of Earth imagery provided by remote sensing (Goodchild, 2007). These data include for instance gazetteers (i.e. the names humans attribute to places, also known as geonames), cultural information (e.g. information on land and building use), environmental information (e.g. measures of air quality) and population information (e.g. population density and socio-economic information).

A radical change in mapping perspectives, which made it possible to fill the gap in the acquisition of geographic information, thus supplementing the traditional efforts of mapping agencies and the power of remote sensing, emerged after the boom of GeoWeb 2.0. Initially there was no consensus on how to call this new trend, with different appeared lexicons such as user-generated content, collective intelligence, neogeography, crowdsourcing, citizen science and eScience. All of them blended into the general idea of exploiting Web 2.0 to create, share and analyze geographic information via multiple computing devices and platforms. However the most successful definition of this dramatic innovation in the history of geography was introduced by Goodchild (2007), who coined the term Volunteered

Geographic Information (VGI) as a special case of Web user-generated-content.

Goodchild himself compared humanity to "a large collection of intelligent, mobile sensors" able to register an incredibly rich amount of geographic information (Goodchild, 2007). Since their childhood and through all the five senses, augmented then by books, magazines, television and the Internet, human beings acquire precious geospatial knowledge related to the areas where they live and work and consisting e.g. of place names, topographic features and transport networks. The ability of capturing, integrating and interpreting this knowledge can be enriched by a tremendous number of available sensor-equipped devices, for instance GPS-enabled cellphones (evolved into the modern smart phones), digital cameras, video monitors, devices tracking vehicles positions and portable sensors for atmospheric pollution. Together with the traditional static sensors, usually focused on environmental purposes (e.g. monitoring the seabed or the city traffic), portable sensors and humans (conceived as sensors themselves) form the three sensor networks able to acquire and synthesize geographic information (Goodchild, 2007).

Only a very small proportion of the human-acquired knowledge had been previously exploited to assemble and disseminate geographic information. Residents were sometimes interviewed by professionals working for mapping agencies for verification purposes (particularly about place names) and by statistical agencies interested in socio-economic variables. This underutilization of the potentially-valuable human knowledge had several reasons: the belief that acquiring some types of geographic information required training, being thus beyond the abilities of amateurs; the lack of mechanisms for communicating and assembling user-acquired contents; and the general lack of trust in people whose work is voluntary and unpaid.

The technological and social context in which all these conditions could be accomplished was clearly the one brought by Web 2.0 and its GeoWeb 2.0 extension. Enabled by broadband Internet communication, users could exploit GPS technology and the free maps from commercial providers (mostly Google) to create and disseminate their VGI, usually in the form of map mash-ups, according to the philosophy of neogeography. Moreover, VGI paradigm fits well the notion of *patchwork*, introduced in the SDI context in relation to the need for national mapping agencies to provide, instead of a uniform coverage of the entire extent of the country, the standards and protocols

under which different groups and individuals could create a composite coverage. Focused on the joint mapping action of multiple users, VGI could exactly create a patchwork coverage helpful for SDI creation and maintenance.

4.2 VGI vs. Crowdsourcing

Before going ahead with the discussion, it is useful to point out a reflection about the term which best describes the phenomenon of geographic information production by the general public. In fact, besides the definition of Volunteered Geographic Information introduced by Goodchild (2007b), another successful, widely-used term in GIS literature is crowdsourcing. The concept was coined by (Howe 2006a) as related to the practice of outsourcing, i.e. the process of transferring business operations to remote cheaper locations. Similarly, crowdsourcing identifies a work performed by an undefined public rather than an organization to which it has been commissioned. In other words, being a form of outsourcing addressing the crowd, the phenomenon was termed crowdsourcing. According to a more formalized definition by Howe (2006b):

Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers.

On the other side geography is far more than the simple data acquisition, as it also includes data modeling prior to acquisition, data integration and also interpretation. For this reason, and in line with some authors, the term crowdsourcing should be preferred to VGI in describing the process of data acquisition using Web technology by large and diverse groups of people, who often are not trained surveyors and do not possess special computer knowledge. Therefore, the most correct term to use to fully outline the practice under examination should be “crowd sourced geographic information” instead of solely crowdsourcing (which does not necessarily imply a geographic component) or VGI (which in itself does not strictly imply the crowdsourcing nature). However, assuming that the distinction between them is clear, and following some notable literature references, hereafter VGI

and crowdsourcing will be used as synonymous of crowd sourced geographic information.

4.3 Volunteered vs. Contributed geographic information

A second reflection about the correctness of the terminology for describing the trend under analysis focuses on whether or not the VGI concept introduced by Goodchild is representative of the whole set of crowd sourced geographic data. A recent review by Harvey (2013) examines the question by making a distinction along ethical lines between Volunteered Geographic Information (VGI) and Contributed Geographic Information (CGI). People would agree that, according to the common meaning of the term *volunteered*, it refers to data that users freely choose to collect. However, being crowd sourced geographic information an ubiquitous element of the current information society, there are plenty of cases in which location information provided by users is anything but volunteered. An example is the detail and amount of data which is daily collected by smart phone users without their knowledge and without any ability of control. This information, which is constantly registered unless users turn off their smart phones or disable location services, is able to surprisingly reveal where users were at a given time instant. Similar examples have recently proliferated in literature (Acohido, 2011) and have shown the important role played by crowd sourced geographic information in the ability of commercial companies and government agencies to know and predict people's activities.

Thus, the first key distinction between Volunteered and Contributed Geographic Information is that the former is collected with user control, and the latter with no or limited user control. In other words, VGI refers to geographic information collected with the knowledge and explicit decision of a person; CGI refers to geographic information collected without the knowledge and explicit decision of a person using mobile technology which records location. A further example of CGI consists of data collected by a car navigation system.

Another element of distinction between VGI and CGI is connected with geographic information collection and reuse. A geotagged picture consciously uploaded by a person on a sharing or a social networking website, which

provides him/her with control over access, is a straightforward example of volunteered geographic data. Nevertheless, if the website later uses the same image for advertisement purposes, or if the geographic location of the picture is used by the website to profile the user and sell aggregated data to mobile advertisers, the originally uploaded picture turns out to be a contributed geographic data. Therefore, crowd sourced geographic data can be defined as VGI if also clarity about purposes and ability to control collection and reuse are guaranteed; if this is not the case, the information is said to be contributed.
The difference between VGI and CGI can be further understood by analyzing the nature of the opt-in and opt-out principles (opt stands for option) in agreeing to use mobile devices and applications. Opt-in provisions allow users the explicit choice of joining or permitting something, thus affording more flexibility and control over the service, e.g. the possibility of using some location service functions while disabling others. On the contrary, under opt-out provisions users face the choice between completely using a service or a device and entirely rejecting the service or the device. In line with the principles of volunteering, opting-in agreements make it clear to users the specifics of how the data they agree to provide is collected and may be reused. On the contrary opt-out provisions may be clear but they are often totalizing, as their acceptance involves the loss of control and influence over the collection and usage of information. This is in clear accordance with crowd sourced CGI.

4.4 Citizen Science

A class of VGI activities which require special attention and analysis is the so called citizen science. Being probably the longest running type of VGI, with projects showing a continuous effort over a century, citizen science is defined as the set of scientific activities in which non-professional scientists voluntarily participate in data collection, analysis and dissemination of a scientific project. Among the wide range of citizen science practices, the real subset of VGI is the one embracing projects where the collection of location information is an integral part of the activity. This intersection between VGI and citizen science is accordingly named geographical citizen science and represents the focus of interest in the present discussion.

Defining as scientists all the active participants in a scientific project, it can be argued that until the late 19th century almost all science was citizen science. In

fact, in that era science was mainly developed by people having additional sources of employment that allowed them to spend time on data collection and analysis. However, still within the phenomenon of professionalization of sciences (which started in the late 19th and went ahead throughout the 20th century), the activity of volunteers has remained constant and productive. Typical disciplines of citizen science projects include archaeology, ecology, zoology, ornithology, astronomy and meteorology.

Haklay (2013) provides a useful classification of both citizen science and geographical citizen science activities. The latter can be differentiated into active and passive according to the role of the volunteers, or into explicit and implicit according to the aim of the activity itself. A geographical citizen science project is active when participants consciously contribute to the observation or the analysis (e.g. taking a picture of an observed animal species and geotagging it); it is instead passive when data are gathered without an active user engagement (e.g. when users are tagged by GPS receivers which monitor their walking activity). A geographical citizen science project is explicit when the activity is aimed at collecting geographic information (e.g. expressly asking to record specific locations of animal observations); it is instead implicit when the aim of the activity is not to collect geographic information (e.g. when a project asks only to take pictures of an observed animal species without requiring to geotag them).

Conversely, citizen science can be distinguished into 'classic' citizen science, community science and citizen cyberscience. The 'classic' expression of citizen science is the one described above in which amateurs are engaged into traditionally scientific activities requiring expertise in a specific field. In community science, scientific measurements and analysis are carried out by members of local communities in order to develop an evidence base and subsequently set action plans to deal with local (typically environmental) problems. Finally, the emergence of the Web as a new global infrastructure has enabled a new dimension of citizen science, termed cyberscience by Grey (2009) and focusing on the use of personal computers, GPS receivers and mobile phones as scientific instruments.

In turn, citizen cyberscience can be classified into volunteered computing, volunteered thinking and participatory sensing. A volunteered computing project requires participants to locally install some software, and use the Internet to receive and send back 'working packages' that are automatically analyzed and then sent back to the main server. Conversely, volunteered

thinking engages participants at a more active and cognitive level (Grey, 2009) asking them to access a website where information or images are presented to them. After the training phase, they are exposed to new (i.e. not previously accessed) information and are asked to carry out classification work.

4.5 Collaborative Mapping

The nature of map production and the dissemination of spatially referenced information have changed radically over the last decade. This change has been marked by an explosion of user generated spatial content via Web 2.0, access to a rising tide of big data streams from remotely-sensed and public data archives, and the use of mobile phones and other sensors as mapping devices. All of these developments have facilitated a much wider use of geodata, transforming ordinary citizens into neogeographers. This increase in user-generated content has resulted in a blurring of the boundaries between the traditional map producer, *i.e.*, national mapping agencies and local authorities, and citizens as consumers of this information. Citizens now take an active role in mapping different types of features on the Earth's surface as volunteers, either by providing observations on the ground or tracing data from other sources, such as aerial photographs or satellite imagery.

Volunteered geographic information (VGI) is potential as it is a low cost and effective way of collecting comprehensive amounts of spatial data to augment more authoritative sources. This innovative technology comes at the right time because maps are outdated in many parts of the world. This situation is unlikely to be resolved by traditional mapping agencies, many of which have been unable, for a number of reasons, to regularly update topographic and other maps, and is further exacerbated by the current financial climate of budget cuts. The lack of up-to-date information is undesirable and hindering development, particularly in areas of rapid change such as expanding cities and the developing world. Collaborative mapping or VGI might offer a solution for obtaining more up-to-date spatial data, or in some situations, it may form the only source of information available. However, the provision of up-to-date geo-information in itself does not mean that collaborative mapping will replace the products of traditional mapmaking organizations because maps need to be accurate and authoritative, aspects for which traditional mapping organizations have the capacity and reputation.

Thus, a big challenge for VGI lies in assessing data quality and in developing procedures to ensure that volunteers produce high quality data, usable in an authoritative context. In addition to volunteered information, citizens also provide other sources of spatially relevant data but in a more indirect manner. For example, spatial information can be harvested from blogs, forums, twitter and other web-based media that could benefit the research and public sector communities.

The collaborative mapping is considered as a mean for citizen engagement in decision-making and public participation. Collaborative mapping and VGI are clearly on the rise, and in the future they will address types of themes *i.e.*, data quality, integration of VGI and incidental data with authoritative data, and enhancement of public participation in decision-making processes through the power of online mapping and social media. Giving ordinary citizens the tools to map and document their environment will lead not only to an unprecedented amount of valuable geodata in the future, but will also produce a new generation of geo-empowered citizens. For us, it is clear that collaborative mapping will become a key component of this future world.

4.6 Crowd sourced data contributed by Expert and Non-Experts

The proliferation of Web2.0 technology over the last decade has resulted in changes in the way that data are created. Individual citizens now provide vast amounts of information to websites and online databases, much of which is spatially referenced. The analysis and exploitation of this georeferenced subset of crowd sourced data, or what is more commonly referred to as volunteered geographic information (VGI), has the potential to fundamentally change the nature of scientific investigation. Citizens have a long history of being involved in scientific research or the more recently coined 'citizen science'. There are many successful examples of citizen science that have led to new scientific discoveries, including unraveling protein structures and discovering new galaxies, as well as websites for public reporting of illegal logging/deforestation and waste dumping, which have demonstrated how citizens can have a visible impact upon the environment and local governance. Analysis of more passive sources of geo-tagged data from the crowd from search engines such as Google has also revealed interesting scientific trends, e.g. the relationship between GDP and searches about the

future, trends in influenza and the ability to characterize crop planting dates. One of the critical advantages of VGI is the potential increase in the volumes of data about all kinds of spatially referenced phenomena. Such data can be collated and used for many different scientific activities: from the calibration of scientific models (e.g. economic prediction models that require information about land use) to the validation of existent data (e.g. maps derived through Earth Observation).

With improved connectivity via mobile phones and the use of low cost, ubiquitous sensors (e.g. those which directly and instantaneously capture data about their immediate environment), the opportunities to exploit such rich veins of VGI are many and varied. However, whilst one of the pressing challenges concerns how to manage large data volumes in terms of processing and storage, a number of yet unaddressed issues persist. These include how to handle data privacy, how to ensure adequate security, and critically, how to assess VGI data quality. Data quality is an area that has attracted increasing attention in the literature: quantifying VGI data quality underpins its usefulness (that is, its reliability and credibility) and potential for incorporation into scientific analyses. The critical issue is whether ordinary citizens can provide information that is of high enough quality to be used in formal scientific investigations.

With open access to high resolution satellite imagery through providers such as Google Earth and Bing Maps, it is possible to collect vast amounts of volunteered information about the Earth's surface such as land cover and land use. The collection of crowd sourced land cover data is the main aim of the Geo-Wiki project in what is currently a contributory approach to citizen science. Geo-Wiki is a web-based geospatial portal (http://www.geo-wiki.org) with an interface linked to Google Earth. It can be used to visualize and validate global land cover datasets such as GLC-2000, MODIS and GlobCover which frequently disagree over the land cover they record at any given location. Since its inception, a number of Geo-Wiki branches have been initiated, each one specifically devoted to gathering different types of information such as agriculture (agriculture.geo-wiki.org), urban areas (cities.geo-wiki.org), biomass (biomass.geo-wiki.org) and more recently human impact (humanimpact.geo-wiki.org).

4.6.1 Data from human impact competition

Crowd sourced data on land cover were collected using a branch of Geo-Wiki called Human Impact (http://humanimpact.geowiki.org) and the data were subsequently used to validate a map of land availability for biofuel production. The volunteers were presented with pixel outlines of 1 km resolution (at the equator) projected onto Google Earth (where pixels in this context refer to the smallest area for which information is collected) and were then asked to determine the percentage of human impact and the land cover type at each location from the following list: (1) Tree cover, (2) Shrub cover, (3) Herbaceous vegetation/Grassland, (4) Cultivated and managed, (5) Mosaic of cultivated and managed/natural vegetation, (6) Flooded/wetland, (7) Urban, (8) Snow and ice, (9) Barren and (10) Open Water. The concept of 'human impact' was defined as the amount of evidence of human activity visible in the Google Earth images. A spectrum of these intensities is shown in *Table 4.6.1.1*, which is loosely based on the ideas of Theobald. Volunteers were also asked to indicate their confidence in the class type and the impact score, whether they had used high resolution imagery and the date of the image.

Volunteers were recruited by emails sent to registered Geo-Wiki volunteers, relevant mailing lists and contacts, in particular those with students, and through social media. Background information on the competitors was collected through the registration procedure. The competition ran for just under 2 months in the autumn of 2011. The top ten volunteers were offered coauthor ship on a paper resulting from the competition as well as Amazon vouchers as an incentive. Other incentives included inviting friends, which resulted in extra points, a leader board so that competitors could gauge the competition, and appealing to the environmental motivation of individuals through the biofuel theme.

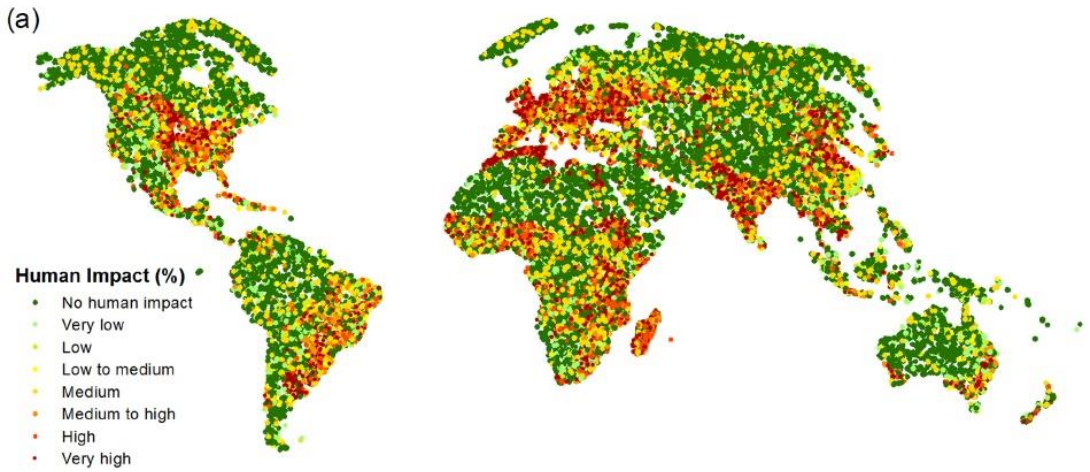
Human Impact	Description
0%	No evidence of any human activity visible
1 to 50%	Some visible evidence of human activities such as tracks/roads; evidence of managed forests; some evidence of deforestation; some scattered human dwellings, some scattered agricultural fields; some evidence of grazing
51% to 80%	Increasing density of agriculture from subsistence on the lower end to intensive, commercial agriculture with large field sizes on the upper end

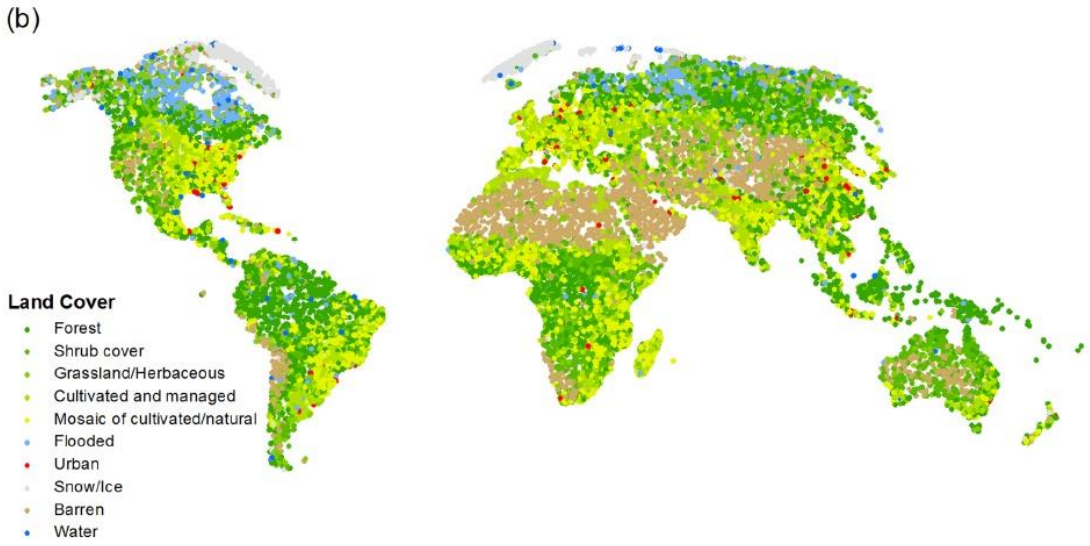
81% to 99%	Urban areas with decreasing amounts of green space and increasing density of housing
100%	A built up urban area with no green space, typically the business district of a city

Table 4.6.1.1 The spectrum of human impact.

A set of 299 'control' points was used to determine quality where three experts with backgrounds in physical geography, geospatial sciences, remote sensing and image classification agreed upon the land cover at each location. The first 99 control points were provided to the volunteers at the start of the competition, the next 100 were provided three-quarters of the way through and the final 100 were provided at the end, where the latter were drawn from higher resolution imagery. The volunteers were then ranked by an index that combined quality and quantity through equal weighting, and the top ten were declared the winners. Interestingly, there were some minor changes in the top ten once quality was considered.

A total of 53,000 locations were validated by more than 60 individuals and *Figure 4.6.1.2* shows the rapid increase in contributions in the last 20 days of the competition, with a particularly large spike at the end. Figure 2 illustrates the spatial distribution of the 53,000 points collected expressed as measures of human impact and land cover. Note that the crowd sourced data can be freely downloaded from http://www.geo-wiki.org.





*Figures 4.6.1.2 Global distribution of pixels collected by the volunteers. The distribution is shown by (**a**) human impact and (**b**) land cover type.*

Of these 53,000 validations, 7657 were at the control locations, which were then used to assess quality. The data were then filtered for 'unknown' expertise resulting in 4020 control data points scored by 29 Expert volunteers and 3548 control data points scored by 33 Non-expert volunteers. Experts were considered to be individuals with a background in remote sensing/spatial sciences versus non-experts who were new to this discipline or had some self-declared limited background. The control data, whose analysis forms the basis of the paper, have the following characteristics. Experts evaluated an average of 64.8 control data points each (s.d. 108.1) and non-experts 57.2 (s.d. 95.1). Although there is the potential for a few individuals to have a disproportionately large impact on data quality and composition, in this case, of the 29 experts, 18 contributed more than 50 evaluations, and of the 33 non-experts, 19 evaluated more than 50 data points. The volunteers' demographics (age, gender, socioeconomic status etc.) were not captured as part of the contributor registration. This is unfortunate, because although a proxy for previous experience is evaluated in this paper, it is well recognized that such factors can influence contributor responses. Such data will be collected in future campaigns.

4.6.2 Analysis of human impact

To determine how well the answers provided by the volunteers matched the control data in terms of the degree of human impact, a linear regression was fit as follows:

$$Y_i = a + bX_i + \varepsilon_i \qquad (1)$$

Where Y_i is the degree of human impact from the control data, X_i is the degree of human impact from the volunteers, a and b are coefficients of the linear regression equation and ε_i is a normally distributed random error term for each observation i.

Each volunteer provided information on expertise during registration. Equation (1) was extended to include an indicator of respondent expertise in the regression model:

$$Y_i = a + b_x X_i + b_E E_i + \varepsilon_i \qquad (2)$$

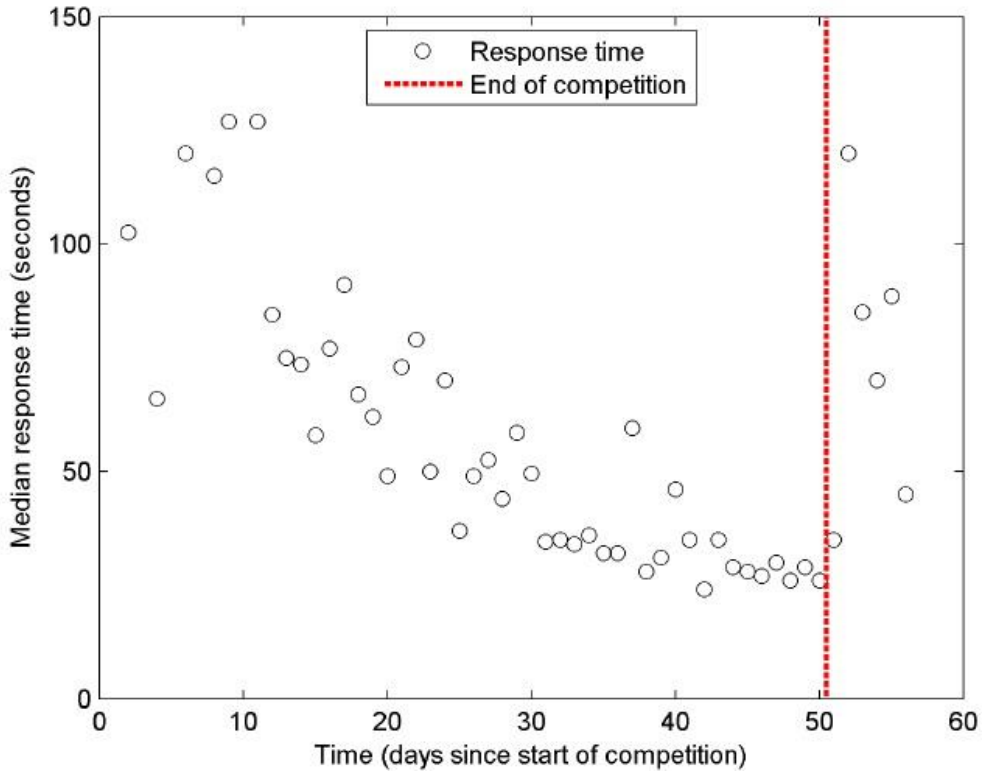


Figure 4.6.2 Median response time of the volunteers. The response time is in seconds measured from the start of the competition until the end at just over 50 days.

Where, in addition to the previously defined variables, b_x is the regression coefficient for volunteer human impact, E_i is the expertise indicator variable for observation i (0 for Non-Expert, 1 for Expert), and b_E is the regression coefficient for this variable. Thus, this coefficient is a measure of the difference in human impact (on aggregate) between the Non-Expert and Expert contributions. This model implicitly assumes human impact is equally predicted by experts and non-experts (i.e. is uniform), and assumes a uniformity of the intercept term within each expert group, if the intercept is considered to be *a* for the non-expert group, and $a+b_E$ for the expert one. The data provided by the volunteers were then analyzed for consistency, which is a known issue in ground validation. After every 50 points, the volunteers were provided with a point they had previously validated. The average, median and standard deviation of the maximum difference between the volunteers and the controls were calculated for all control points, by expertise, by volunteer consistency in the land cover they recorded, and by confidence.

Finally, the response times of the volunteers were calculated between each successive data point they scored. The median response time was 55 secs with a first and third quartile of 32 and 100 secs respectively. The average response time was 5,226 secs, indicating a highly skewed distribution, which reflects large pauses in contributions, e.g. at the end of a validation session. *Figure 4.6.2* shows the median response time per day over the course of the competition. There is a general trend towards shorter response times as the competition unfolded with the shortest response times between successive validations occurring at the end of the competition. Thus, we were interested in understanding the relationship between response time and quality of the human impact responses overall and whether there was any difference in quality towards the end of the competition.

The response time data were first pre-processed in two ways. First, all response times greater than 5 minutes were removed as these were deemed unrepresentative of typical behavior. This was based on visual inspection of the distribution. However, 5 minutes also represents the 92.5th percentile and therefore includes the majority of the data. Second, response times were log transformed due to the skewness of the distribution. A linear regression equation of the form given in (1) was fit to the entire dataset where the dependent variable, Y_i, was the absolute difference in the answers for human impact between the control data and the volunteers' scores, and the independent variable, X_i, was the log of the response times, with a and b

representing coefficients of the linear regression, and ε_i the error term for each observation i.

The last 100 control points provided to the volunteers at the end of the competition were locations of cropland or agricultural land covers (the classes of Cultivated and managed and Mosaic of cultivated and managed/natural vegetation) and where high resolution images existed. In order to evaluate how volunteer performance changed with experience, only control points with agricultural land cover and where high resolution images were available were selected from the first 199 control points. The average accuracy in human impact across the first two control sets was then compared to the average accuracy of the third set using a t-test to determine whether there were any significant differences.

4.6.3 Analysis of land cover

As in the analysis of human impact scores above, control points were used to evaluate volunteer accuracy in terms of the land cover they indicated. An error or confusion matrix was populated for all contributors *(Table 4.5.3)* and the overall accuracy was calculated as follows:

$$Accuracy = \frac{\sum_{i,j=1}^{n} xij}{\sum_{i=1}^{n}\sum_{j=1}^{n} xij} * 100 \qquad (3)$$

where i is the volunteer class, j is the control class and n is the total number of classes.

	Class 1 (control j)	Class 2 (control j)	…	Class n (control j)
Class 1 (volunteer i)	x1,1	x1,2	...	xn,1
Class 2 (volunteer i)	x2,1	x2,2	…	xn,2
…	…	…	…	…
Class n (volunteer i)	xn,1	xn,2	…	

Table 4.6.3 A confusion matrix for the comparison of controls with responses from the crowd.

In addition, two other measures of accuracy were calculated, specific to each land cover class: user's and producer's accuracies. User's accuracy describes errors of commission or Type I errors. For example, the user's accuracy for the forest class indicates the likelihood that what was labeled as forest by the volunteers really is forest. Producer's accuracy reflects errors of omission or

Type II errors. Using the forest example again, this measure reflects how well the forest cover control pixels were classified by the volunteers. These two measures are calculated as follows:

$$User's\ Accuracy\ (by\ class\ i) = \frac{x_{i,i}}{\sum_{j=1}^{n} xij} * 100 \qquad (4)$$

$$Producer's\ Accuracy\ (by\ class\ j) = \frac{x_{j,j}}{\sum_{i=1}^{n} xij} * 100 \quad (5)$$

Where i is the volunteer class, j is the control class, and n is the total number of classes. Separate accuracy measures were calculated for the three sets of control pixels (to determine whether accuracies change over time) for locations where the volunteers were the most confident and to compare experts and non-experts. Contributor consistency in land cover labeling was then analysed by determining the proportion of times when the same land cover type was chosen when presented with the same data point. This was calculated for all points, by expertise, and by various degrees of confidence.

Finally, the impact of response time on the quality of land cover validations was analyzed using logistic regression of the following form:

$$logit(P_i) = a + bX_i \qquad (6)$$

Where the probability P_i that the land cover is correctly identified is expressed as a function of response time, X_i.

The effect of response time on accuracy in the final set of controls was compared with the first and second set to determine whether contributors were more interested in scoring a greater number of points and spent less time on each data point towards the end of the competition. A two-tailed binomial test was used to test whether the number of correct classifications at the end of the competition was greater than expected based on the total number of classifications performed and the probability of correct classification in the earlier part of the competition.

4.6.4 Results of human impact

The result of the regression described in Equation (1) to determine how well the degree of human impact can be predicted by the contributors based on the control points is provided in *Table 4.5.4.I* This shows that b differs significantly from zero and is positive but less than 1 suggesting that there is evidence that the users underestimated the degree of human impact by roughly 30 percent. The results of including an indicator variable describing respondent expertise (Equation 2) are shown in *Table 4.5.4.II* The slopes are still positive and suggest that allowing for expertise even in a simple way changes the results of relating to the slope term. To investigate this further, Equation (1) was extended to include variables describing expertise. Although computed together, this effectively splits the regression into two models - one for each of the expert groups - and the results are shown in *Table 4.5.4.III* These results indicate that there is little variation in the degree to which the expert and non-expert group underestimated the degree of human impact.

	Estimate	Std. Error	t value	Pr(>|t|)
a	11.300	0.363	31.16	0.000
b	0.699	0.006	122.43	0.000

Table 4.6.4.I. Regression analysis for the model $Y_i = a + bX_i + \varepsilon_i$ where Y_i is the degree of human impact from the control data, X_i is the degree of human impact from the participants.

	Estimate	Std. Error	t value	Pr(>|t|)
a	9.009	0.432	20.85	0.000
b_x	0.705	0.006	123.49	0.000
b_E	4.251	0.442	9.62	0.000

Table 4.6.4.II. Extending the regression to include an indicator of expertise, where b_E is the regression coefficient for this indicator and b_x is the regression coefficient for participant human impact scores.

	Estimate	Std. Error	t value	Pr(>|t|)
a (Expert)	7.960	0.527	15.12	0.000
a (Non-Expert)	14.200	0.494	28.74	0.000
b (Expert)	0.725	0.008	91.06	0.000
b (Non-Expert)	0.685	0.008	83.61	0.000

Table 4.6.4.III. The regression analysis of predicting the degree of human impact by expert and non-expert groups, when the regression is split into 2 simultaneous models.

The distribution of human impact scores for the control pixels and the contributor data by land cover class. It shows a general trend for contributors to underestimate the degree of human impact across the different land cover types with the exception of (5) Mosaic of cultivated and managed/natural vegetation. A further analysis explored how human impact scores varied with land cover class. The standard regression described in Equation 1 was extended to include indicators for the land cover classes. Since there was only a small number of data points classified as Open water, Barren or Urban, these classes were excluded from the regression analysis.

The results show that the prediction of the degree of human impact varies with land cover classes. The coefficients for the herbaceous vegetation/Grassland class most strongly predict human impact; the coefficients for the Shrub cover class are the weakest predictors and all classes underestimate human impact. This indicates that the conceptualizations of these classes may need to be more clearly defined and perhaps more training examples used to illustrate the different degrees of human impact by land cover type.

Overall the contributors were consistent in their answers regarding the degree of human impact, with an average deviation of less than 10% (i.e. 9.6%) although the spread of answers was higher at 17.4%. When expertise was considered, non-experts had a lower average deviation than the experts by just fewer than 3%. When the consistency was extended to land cover, those pixels which showed consistent choices in land cover had a lower average deviation in human impact by 8.3% compared to those which showed inconsistency in land cover choice. This reflects pixels that were clearly more difficult to identify. Finally, when contributors were the most confident in their choice of human impact, they were also more consistent (average deviation of 7.9%), with consistency decreasing as confidence decreased

resulting in an average deviation of as much as 25.9% for the least confident category. This analysis of consistency serves to highlight the need to examine those pixels which were not consistently labeled and which are probably more difficult to judge in terms of both human impact and land cover, which can then be used to help train the volunteers.

The results of the regression analyzing the effect of response times indicate that the agreement between the volunteers and the control pixels increased significantly with a faster response time for human impact, although the effects were small. For each increase in magnitude in response time, the agreement between the crowd and the control pixels increased in accuracy by 1.4%. The average deviation in human impact for pixels of (4) Cultivated and managed and (5) Mosaic of cultivated and managed/natural vegetation and high resolution imagery from the first two control sets was 17.1%. This was compared to the third set of control data points (consisting of only these pixel types) and the average deviation in human impact was lower, decreasing to 14.7%. A t-test confirmed that the means are significantly different from one another (p,0.0001; t =24.8533; degrees of freedom = 3326.222) and showed that accuracy in human impact actually increased at the end of the competition. Thus, these analyses indicate that there are no particular concerns over quality in relation to response time.

4.6.5 Results of Land Cover

The overall accuracies for the three sets of control points labeled C1, C2 and C3 are presented in *Table 4.5.5* for the full dataset, considering only those contributions where confidence was high (i.e. 'sure' on the slider bar) and then disaggregated by expertise (i.e. experts or non-experts). Considering all three sets of control data, accuracy varies between 66 and 76%. There is little difference between the first and second set of controls but there is a marked increase in accuracy for the final set (C3) with 76%. This is unsurprising since the final control sample was drawn from high resolution imagery. When taking only those answers where the volunteers indicated high confidence (or 'sure' on the slider bar), there was around a 3% increase in the accuracy to 69%. Unlike with human impact, experts were more accurate than non-experts, e.g. 62% for no experts and 69% for experts for C1 with even larger differences observed for C2 and C3. This suggests that extra training should be provided to those individuals with a non-expert background. As training manuals are often unread or rarely consulted, a more

interactive approach could be introduced such that the volunteers are made aware of their errors as they progress through a competition. In addition, a forum could be set up to discuss pixels that present difficulties in identification, particularly for non-experts.

Dataset used	No allowance for confusion between classes		
	C1	C2	C3
Full dataset	66.4	66.5	76.2
Confidence rating of sure	69.4	69.3	78.9
Experts	69.2	72.3	84.6
Non-experts	62.4	61.9	65.9

Table 4.6.5 Accuracy of land cover (in %) based on comparison of volunteer response with three sets of controls.

There is generally an increase in the accuracy across control sets although C3 should only really be considered for cropland and mosaic classes. The lowest accuracies are in shrub cover, grassland/herbaceous and the mosaic cropland class, which indicates the need to provide more examples of how these classes appear on Google Earth within the training materials as the volunteers are confusing these classes more often than others. When considering points where the volunteer had a high confidence, the patterns are similar and there is generally an increase in accuracy although the mosaic cropland class continues to be more problematic, with a decrease in the user's accuracy across control sets. Finally, the effect of expertise on land cover classification accuracy produced variable results depending upon the land cover type and the control set considered. For the forest class, the non-experts improved in their ability to correctly identify forest by the second set of controls, while the non-experts actually showed a decrease in the producer's accuracy. Similarly, for the shrub class, the non-expert showed a greater level of improvement in the second set of controls compared to the expert and outperformed them in terms of both user's and producer's accuracy in C2. The experts were better than non-experts at identifying herbaceous, cropland and mosaic but once again there were differences in the user's and producer's accuracies. By building up a picture of where experts and non-experts have differing performance by land cover class, we can tailor the kinds of training materials provided to the volunteers, focusing on areas where greater problems in identification lie.

The volunteers were consistent in their response just over 76.1% of the time where this was slightly lower for experts (75.7%) and slightly higher for non-experts (76.7%). A very minor increase to 77.6% was observed when considering only those pixels where the volunteer was sure but when the volunteers were less sure or unsure about their responses, their consistency in response decreased to 66.7%.

4.6.6 Discussing the obtained results

The results showed that there is little difference between experts and non-experts in identifying human impact while experts were better than no experts in identifying land cover. However, the results for both varied by land cover type and through the competition. For example, experts were better than non-experts at identifying shrub land cover at the start of the competition but non-experts improved more than experts and then outperformed them in shrub cover identification by the middle of the competition, indicating that volunteers were learning over time. The volunteers were shown to be reasonably consistent in their characterizations of human impact and land cover with non-experts outperforming the experts in terms of human impact and vice versa for land cover. Moreover, when contributors were confident in their choice of human impact, they were also more consistent, and unsurprisingly, consistency decreased as confidence decreased.

Finally, increased response times (as observed towards the end of the competition) did not have a negative impact on quality, and volunteers were therefore not sacrificing quality for the desire to complete more locations and thereby win the Geo-Wiki competition. Thus overall, the non-experts were as reliable in what they identified as the experts were for certain, identifiable situations, and the reliability of the information provided by non-experts improved faster and to a greater degree than experts. Thus, better, targeted training materials and a continual learning process built into the competition might help address these issues. Also, allowing volunteers to reflect on the information they contribute, for example by regularly feeding back evaluations of their data through the use of control points or by making additional material available to them, would also potentially decrease differences between experts and non-experts, particularly in the classification of land cover.

Chapter 5

Open Land Map (The Software)

Different organizations have developed land cover maps. The classification methods used by the many organizations looking at this data are not standardized; in addition, some land coverage areas have been miss-classified due to image acquisition during various seasons. These problems are exacerbated by difficulty in geometric and radiometric correction, and other issues related to non-standardized spectral interpretation of imagery. In order for the land coverage data sets be effective, their accuracy has to be quantitatively evaluated. To do that, high quality analyzing, classification and validation services are needed. We propose to help accomplish this with an open source geo-platform that allows sharing of these techniques and also comparing land cover classification results.

Open Land Map aims primarily to explore citizen science for improving the land cover classification through geo-visualization and geo-crowdsourcing on internet platforms. The purpose of the application is to let the users analyze two different land coverage maps - at the same time - on two NASA World Wind virtual globes (see section 3.9), to evaluate the coherency and highlight the differences between them. It helps to validate the land coverage maps in order to produce more accurate land coverage data sets. The users can upload, process, and publish layers via GeoServer API. They also can classify land coverage classes of the two maps according to CORINE (see CORINE programme in section 1.6) categories, execute some operations to process the maps using rasterization or vectorization algorithms, identify the discordant or miss-classified land cover areas by creating a difference map

from two land cover maps, calculate the accuracy indexes of the data using suitable algorithms, validate the land cover data selecting the correct land cover class, upload cartographic data and orthophotos which are useful to validate the classified land cover maps, and finally re-calculate the accuracy indexes after the validation phase.

The system also makes available a tool to manage the descriptive information (Metadata) for the available data resources. The Metadata are considered as general proprieties of the data source, it includes information about the identification, constraint, extent, quality, spatial and temporal reference, distribution, lineage, and maintenance of the digital geographic dataset.

Since the exchange of geographic information has increased significantly and an enormous resource of volunteered geographic information (VGI) has become available, Open Land Map provides crowdsourcing methods to consolidate the validation tasks. Volunteer citizens will be involved within the validation phase via citizen science campaigns improving the accuracy of land cover datasets. In particular, they will be asked to contribute by providing *in-situ* observations on the ground, adding photos and videos of the land cover. Such an approach, allows internet users from any region of the world to evaluate land cover data, identify the inaccuracies in land cover data, and get themselves involved in the global validation task. The validation information will be recorded in a database, and used iteratively to produce more accurate land covers.

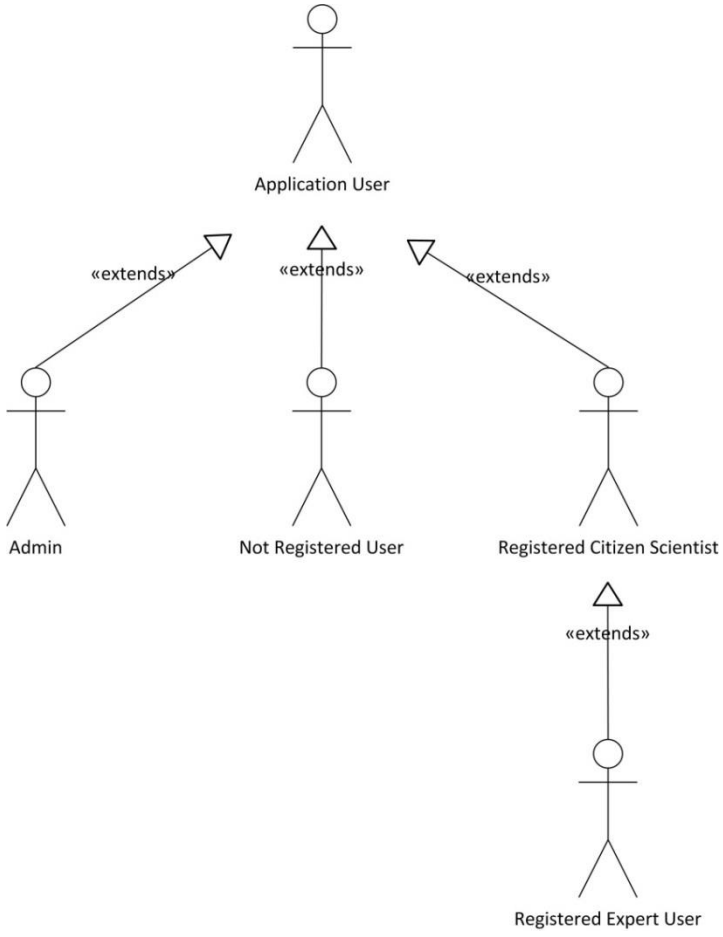
5.1 Requirements Analysis: Users, Goals and Functionalities

5.1.1 User Group Specification

A. Registered Expert User

Description: a person who is familiar with GIS domain, able to process and interpret the results which he/she obtain using the application.
Profile data: first name, last name, country, e-mail and password.
Super-group: registered citizen scientists.
Sub-groups: none.

Relevant use cases: in addition to the use cases of the registered citizen scientists, the registered expert users can “manage layers”, “upload maps”, “publish maps via GeoServer”, “match classes”, “process maps”, “calculate accuracy indexes”.
Access rights*:* none.



5.1.1 User Groups Hierarchy

B. Registered Citizen Scientists

Description: amateurs, volunteers and nonprofessionals, they contributes to the project with their intellectual effort, surrounding knowledge or with their tools and resources.
Profile data: first name, last name, country, e-mail and password.
Super-group: none.

Sub-groups: registered expert users.
Relevant use cases: “login”, “add multimedia files”, “validate pixels” and “vote pixels”.
Access rights*:* none.

C. Not registered User

Description: a person who wants to use the application, he/she can register from the main page of the application.
Profile data: first name, last name, country, e-mail and password.
Super-group: none.
Sub-groups: none.
Relevant use cases: “register”.
Access rights*:* none.

D. Admin

Description: a person who can manage the list of registered users by viewing the users’ list, removing user accounts, and sending notifications to the users of the application.
Profile data: first name, last name, country, e-mail and password.
Super-group: none.
Sub-groups: none.
Relevant use cases: “view users” and “manage users”.
Access rights*:* registered users’ information (full name, e-mail, date of register, etc.).

5.1.2 Use Case Specification

A. Login

Purpose: To express how a registered user can access the functions of the application.
Pre-Condition: The user is registered: has valid e-mail and password.
Post-Condition: The user successfully logs into the application.
Workflow:
1. The user receives an input form asking for e-mail and password.
2. The user inputs his credentials.
3. If the credentials are correct, the user is authenticated and the application functionalities are activated. Otherwise, the user is requested to reinsert his/her credentials.

B. Register

Purpose: To express how an unregistered user can ask for the login Credentials.
Pre-Condition: None.
Post-Condition: The system activates the main menu for the Login.
Workflow:
1. The user clicks on registration button.
2. System displays the registration screen.
3. User enters biographical data and chooses username and password.
4. System validates the information.
5. System stores the data.
6. System activates the main menu for the login.

C. Add multimedia files

Purpose: To express how a registered user can contribute by in-situ observations to consolidate the validation process.
Pre-Condition: The user is registered and has some photos or videos of the zone which is going to be validated.
Post-Condition: The system saves the files added by the user in the DB and publishes them online.
Workflow:
1. The user clicks on add multimedia files button.
2. System displays the uploading screen.
3. User adds the file and confirms the operation.
4. System validates the uploaded file and save it into the database.
6. System adds the file in the multimedia section.

D. Validate pixels

Purpose: To express how a registered user can collaborate to correct the misclassified zones on the maps.
Pre-Condition: The user is registered.
Post-Condition: The system registers both the position and the class (the category) given by the user for a specific zone or pixel.
Workflow:
1. The user clicks on a specific position of the NASA World Wind globe to add a marker.
2. System displays a marker (pin).
3. User clicks on the marker.
4. System displays the validation screen.
6. User chooses the right class and click on confirm.

7. System saves information about the validation operation.

E. Vote pixels

Purpose: To express how a registered user can vote a validated position.
Pre-Condition: The user is registered and a validated position is available on the globe.
Post-Condition: The system registers the vote of the user.
Workflow:
1. The user clicks on a validated marker added to the map.
2. System displays the voting screen.
3. User gives his/her vote of the right class and click on confirm.
4. System saves information about the voting operation.

F. Manage layers

Purpose: To express how a registered user can add WMS layers, show them and hide them on the map.
Pre-Condition: The user is registered and a connection to the layer server is available.
Post-Condition: none.
Workflow:
1. The user clicks on manage layers button.
2. System displays the layers panel.
3. User requires WMS layers.
4. System provides the needed layers to user.
5. User manages (turn on, turn off) the layers according to his/her needs.
6. System performs the management action requested by the user.

G. Upload maps

Purpose: To express how a registered user can upload land cover maps to the system.
Pre-Condition: The user is registered.
Post-Condition: The system saves land coverage maps.
Workflow:
1. The user clicks on upload layers button.
2. System displays the uploading screen.
3. User selects the map and click on confirm.
4. System saves the uploaded maps.

H. Upload Cartographic Data

Purpose: To express how a registered user can upload cartographic documentation or some photos of the land skin like orthophotos into the system. These data are helpful during the validation tasks.
Pre-Condition: The user is registered.
Post-Condition: The system saves cartographic data correctly.
Workflow:
1. The user clicks on upload cartographic data button.
2. System displays the uploading screen.
3. User selects the data and click on confirm.
4. System makes the uploaded data available to users.

I. Publish maps

Purpose: To express how a registered user can make land cover maps available as WMS layers to the system.
Pre-Condition: The user is registered and a connection to GeoServer REST API is available.
Post-Condition: The system publishes the maps on GeoServer.
Workflow:
1. The user clicks on publish layers button.
2. System displays the publishing screen.
3. User selects the uploaded land coverage map and click on publish button.
4. System communicate with the REST API of GeoServer which publishes the maps like WMS layers.

J. Manage Metadata

Purpose: To express how a registered user can manage geospatial metadata in order to manipulate the descriptive information for data.
Pre-Condition: The user is registered.
Post-Condition: none.
Workflow:
1. The user clicks on manage metadata button.
2. System displays metadata managing screen.
3. User selects the data resource which he/she wants to handle its metadata.
4. System allows the user to manage the descriptive information.
5. User manipulates the metadata and confirms.

K. Match classes

Purpose: To express how a registered user can unify the classes of the two maps by making them correspondent CORINE land cover categories.

Pre-Condition: The user is registered.
Post-Condition: none.
Workflow:
1. The user clicks on match classes button.
2. System displays the matching screen.
3. User selects the two maps which he/she wants to analyze their classes.
4. System shows the classes of each map.
5. User reclassifies the maps according to CORINE land cover categories.
6. System records the new classified maps.

L. Process maps

Purpose: To express how a registered user can rasterize the vector data, and produce a difference map between the two maps.
Pre-Condition: The user is registered and a connection to a geospatial processing services (GRASS) is available.
Post-Condition: The system returns the processed map as a downloadable file.
Workflow:
1. The user clicks on process maps button.
2. System displays the processing option screen.
3. User selects the needed processing operation.
4. System performs the requested operation and make the resultant map ready for downloading.

M. Calculate accuracy indexes

Purpose: To express how a registered user can obtain some results which show the level of precision of the land cover maps.
Pre-Condition: The user is registered, and a land cover map is available.
Post-Condition: none.
Workflow:
1. The user clicks on calculate accuracy indexes button.
2. System displays the related screen.
3. User asks for the accuracy indexes for a specific map.
4. System returns the requested results.

N. View users

Purpose: To express how the admin can view the list of registered users.
Pre-Condition: At least one registered user exists.
Post-Condition: the system visualizes the list of registered users.
Workflow:

1. The admin picks the view list button.
2. The system visualizes the list of registered users.

O. View users

Purpose: To express how the admin can manage the list of the registered users (remove user account and send notifications).
Pre-Condition: none.
Post-Condition: The system executes the managing actions.
Workflow:

- *Remove User Account*

1. The admin selects a user account and clicks on the delete button.
2. The system asks for a confirmation.
3. The admin confirms the action.
4. The system deletes the user's account from the database.

- *Send Notification*

1. The admin selects a specific user and clicks on the send notification button.
2. The system opens a massage body.
3. The admin writes the notification in the massage's body.
4. The admin clicks on the send button.

Figure 5.1.2 Open Land Map Use Case Diagram

5.1.3 Required Functionalities and the Activity Diagram

There are some required actions and functionalities which have to be made available by our application as those were specified in the use cases (see section 5.1.2). In order to illustrate these needed functionalities which are in turn necessary to let users commence the validation activities, we can consider two scenarios: first lets us consider the case when there is an expert user who want to execute some advanced GIS functions, and then, we are going to show a citizen scientist scenario in which the user is an amateur participant contributing to the validation task, and has no knowledge of the advanced GIS processing functionalities. Note that the basic steps (Registering and login) are not illustrated in detail, in both scenarios, since they are common functionalities used in many *Desktop* and *Web* applications, and there is nothing in particular as the users provide some required information during the registration step, then use their credentials in order to access the platform and use its provided functionalities.

So first imagine that a user who is expert in the domain, in such a way able to deal with GIS services and functions, accesses the application using a valid email and a password previously obtained during a registering step to get the credentials. Let's consider that the expert user has a recent land cover map of a region or a country, the map is in vector format, and he/she wants to use it in order to classify a global land cover issued by a certain organization. After logging into the application, the user can use *Upload Maps* function to upload his/her map in to the application, after that, he/she can convert the vector map into a raster map using *Process Maps* function which exploit GRASS capabilities via web processing service (WPS) to process the map . Now the user can publish the raster map within GeoServer using GeoServer API, the published map can be then retrieved using *Manage Maps* function that permits also to manipulate the layers e.g. turn off, turn on, move to front, move back, etc. When the user has the two maps, the global one and his/her map, in raster format, there is a probability that the classes of the two maps are not comparable or better classified in heterogeneous way, there is a *Match Classes* function which allows users to classify both maps according to CORINE land cover categories (see CORINE programme section 1.6), once this step is done, the user will be able then to analyze the maps and start to evaluate the differences and create a differences map via *Process Maps* function. The user can proceed with the validation tasks, he/she creates markers on the layers pointing out the misclassified areas, and these markers will be stored in a database for accuracy indexes calculation and statistics

purposes. The user can also upload cartographic data or orthophotos via *Upload Cartographic Data* function, these data are somehow useful when the users start to validate the classified land covers. The system also provides a geospatial Metadata handler in order to manage the descriptive information of the resource data.

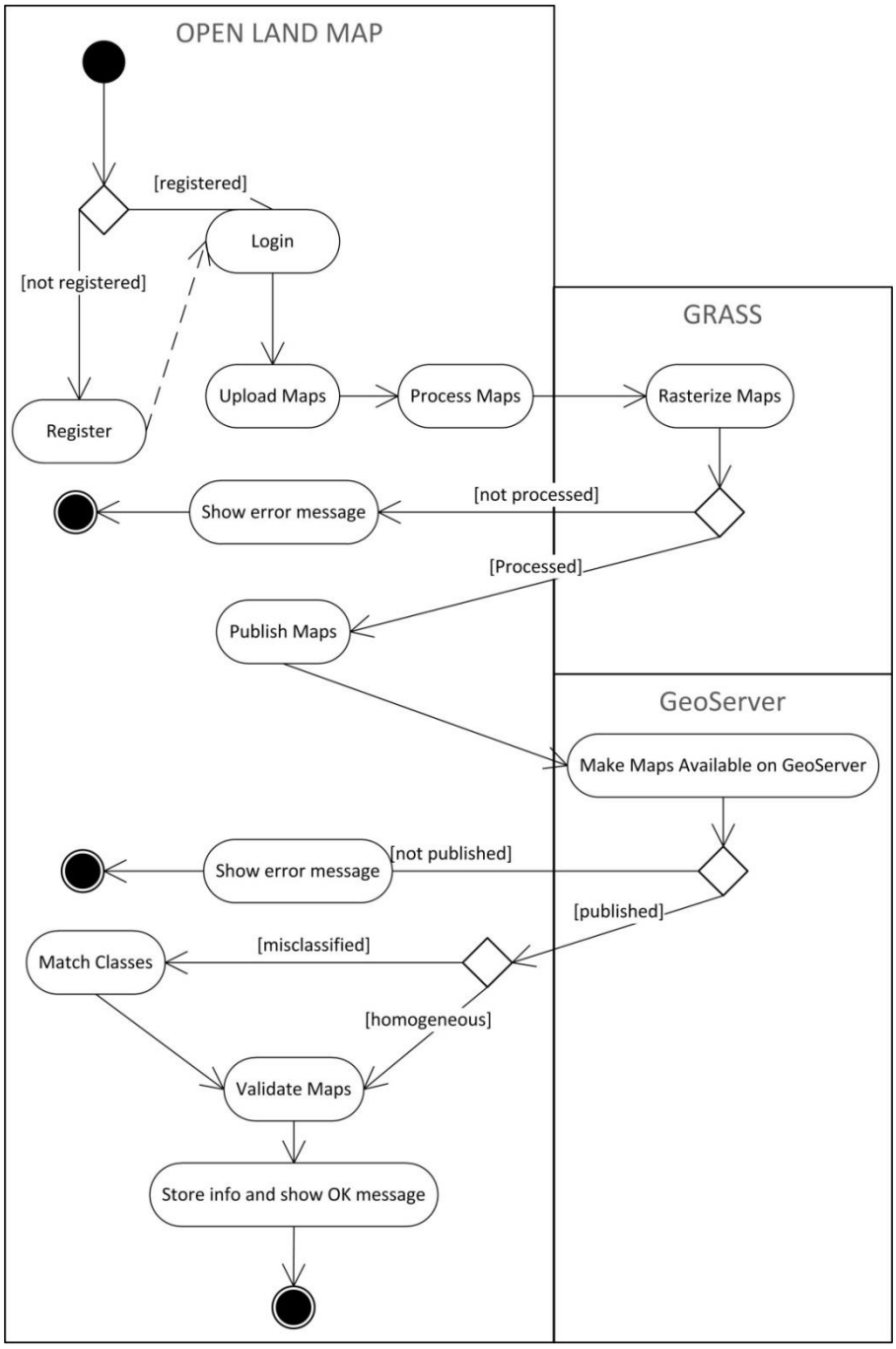


Figure 5.1.3.I Expert user most important steps activity diagram

Citizen Scientists (Volunteers), instead, contributes to the project with their intellectual effort, surrounding knowledge or with their tools and resources to improve the accuracy of land cover datasets via citizen science validation campaigns. Once a volunteer registers and logins into the applications, he/she can participate in the validation task by make use of the *Validate Pixels* function which permits to assign a particular land class to a specific pixel on the map, they can also vote some validated points by the other users using *Vote Pixels* function that enables the users to give his/her opinion about a previously validated area by selecting the right land class, if they think that the area was misclassified, and sending their votes. Moreover, they can also add multimedia files e.g. video, images, etc. of the considered area in order to consolidate both the validation and the voting tasks (see figure 5.1.3.II).

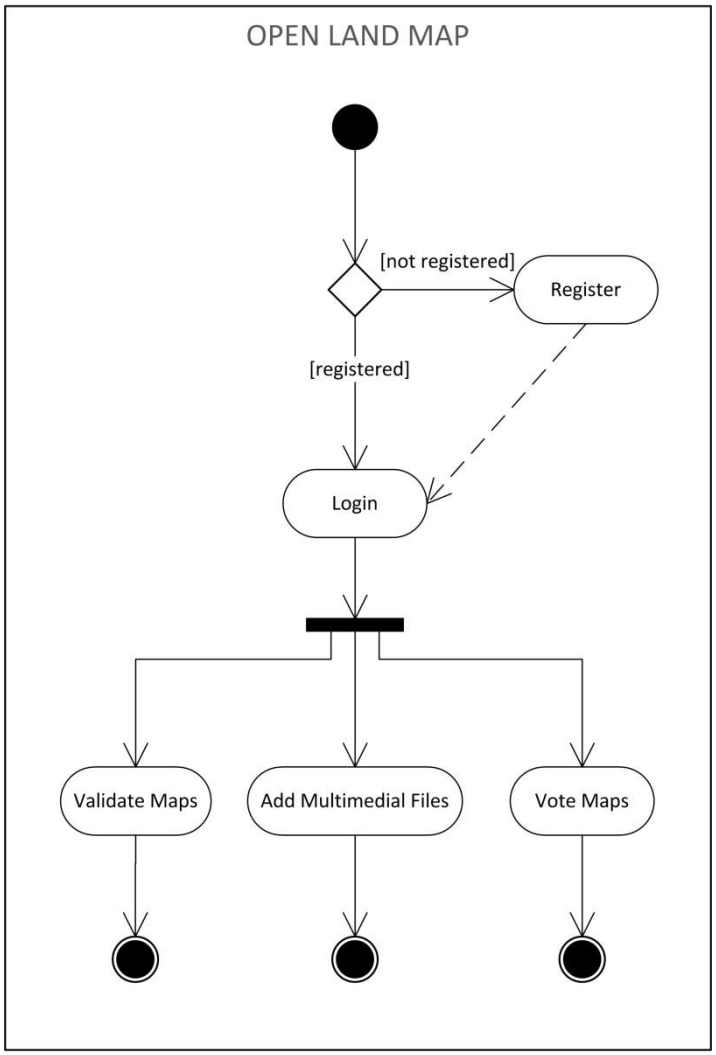


Figure 5.1.3.II Citizen Scientist activity diagram

5.1.4 Land Cover Map Statechart Diagram

Statechart diagram is used to model dynamic nature of the system. It defines different states of an object during its lifetime. And these states are changed by events. So Statechart diagram is useful to model our reactive system; where reactive systems, can be defined as systems that respond to external or internal events. Statechart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. So the most important purpose of Statechart diagram is to model life time of an object from creation to termination.

We want to illustrate (as seen in figure 5.1.4) the various lifecycle states of the object: Land Cover Map (or LCM for abbreviation). LCM can have different states, first it has to be *uploaded* into the system. It can be a vector land cover map, and Open Land Map helps to convert it to a raster image map, so it will be *rasterized*. Then it can be published on GeoServer via its related API having a *published* state. After that, if the LCM is not classified according to CORINE land cover categories, the map has to be processed and a matching step is required to assign new land cover classes (LCM *matched*). Finally, the map is subject to a validation phase, and at the end we can have a *validated* land cover map.

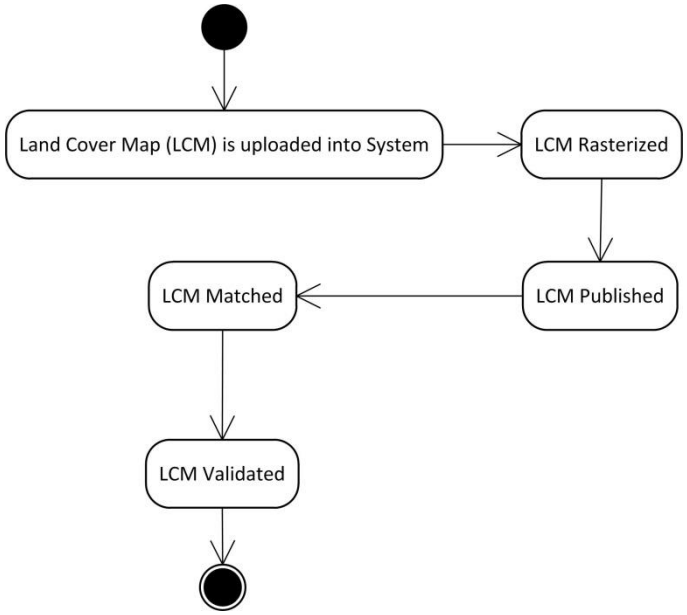


Figure 5.1.4 Land Cover Map (LCM) statechart diagram

5.1.5 Data Model

The data model describes the data or information aspects of a business domain or its process requirements, in an abstract way that lends itself to ultimately being implemented in a database such as our used relational database. The main components of the model are the entities (things) which may have various properties (attributes) that characterize them. Beside the attributes, there are also relationships that can exist among the entities.

The following will be a brief description of the entities and their relationships used in Open Land project. Each entity has a primary key, some attributes, and it can be related to some other entities. Of course the relationships between entities can be one-to-one, one-to-many, or many-to-many according to the system specific requirements.

Entity: User
Primary key: integer user identifier
Attributes: name, surname, e-mail, password, country and a registering date.
Relations:

- A User can add many video entities
- A User can add many image entities
- A User can add many comment entities
- A User can create many MarkerPixel entities
- A User can create many MarkerPolygon entities
- A User can vote many MarkerPixel entities
- A User can vote many MarkerPolygon entities
- A User can use many Land entities in the MarkerPixel voting task
- A User can use many Land entities in the MarkerPolygon voting task

Entity: MakerPixel
Primary key: integer marker pixel identifier
Attributes: latitude, longitude, mapName, and a time of creation
Relations:

- A MarkerPixel is created by one User entity
- A MarkerPixel is classified to a one Land entity
- A MarkerPixel can be voted by many User entities
- A MarkerPixel can be voted to many Land entities

Entity: MakerPolygon
Primary key: integer marker polygon identifier

Attributes: mapName, and time of creation
Relations:

- A MarkerPolygon is created by one User entity
- A MarkerPolygon is classified to a one Land entity
- A MarkerPolygon can be voted by many User entities
- A MarkerPolygon can be voted to many Land entities

Entity: Land (land cover category)
Primary key: integer land identifier
Attributes: land description
Relations:

- A Land can categorize many MarkerPixel entities
- A Land can categorize many MarkerPolygon entities
- A Land can be used to vote many MarkerPixel entities
- A Land can be used to vote many MarkerPolygon entities
- A Land can be assigned by many User entities in MarkerPixel voting task
- A Land can be assigned by many User entities in MarkerPolygon voting task

Entity: Video
Primary key: integer video identifier
Attributes: title, video, uploading time, pixel identifier, and polygon identifier
Relations:

- A Video can be uploaded by many User entities

Entity: Image
Primary key: integer image identifier
Attributes: title, image, uploading time, pixel identifier, and polygon identifier
Relations:

- An Image can be uploaded by many User entities

Entity: Comment
Primary key: integer comment identifier
Attributes: comment, writing time, pixel identifier, and polygon identifier
Relations:

- A Comment can be uploaded by many User entities

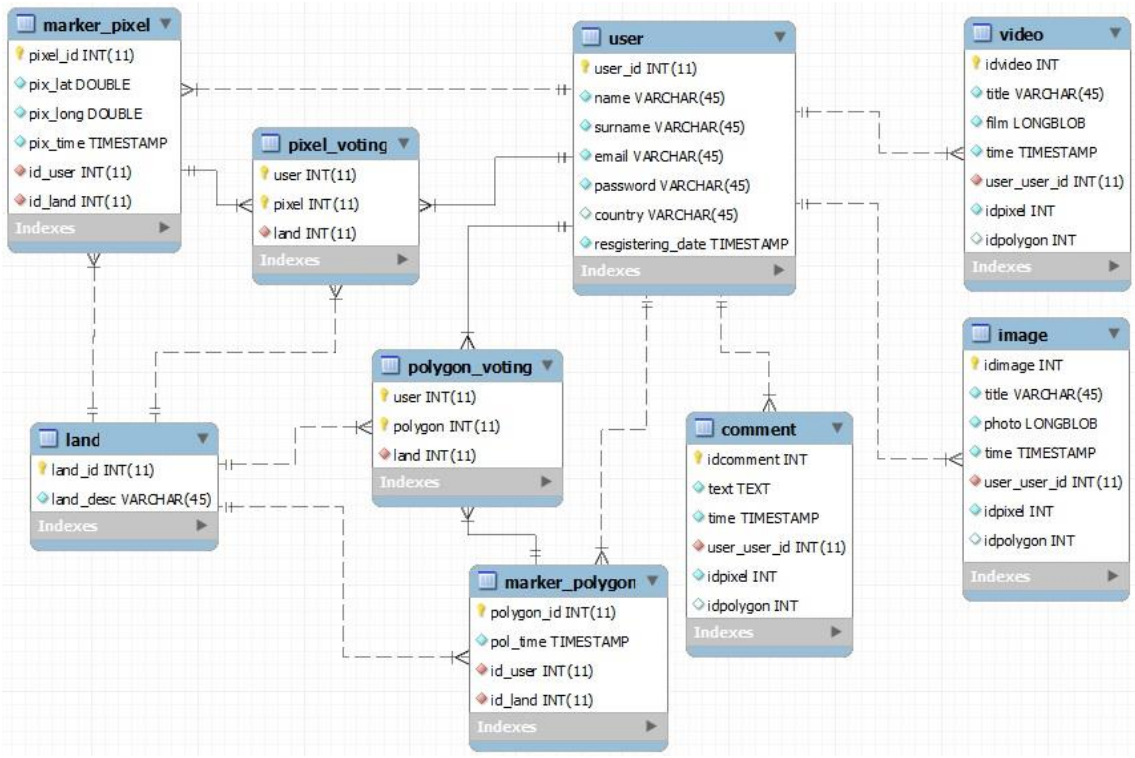


Figure 5.1.5 Data Model of Open Land Map project

5.2 Technological Choices (Server Side)

5.2.1 GeoServer

The choice of GeoServer among the existing geospatial web servers derives from a handful of considerations. First of all the author's knowledge and familiarity with the software, which is one of the most well-established and used geospatial servers worldwide, have been consolidated for a long time. GeoServer is also chosen for its general ease of use and above all for its high WMS performance. The latter is certified by the result of the last OSGeo WMS Benchmarking, performed during the 2011 FOSS4G conference in Denver and based on the comparison of the performance of multiple WMS servers. Written in Java, GeoServer consists of a standalone servlet running in servlet container applications. By default it packages Jetty as an embedded Web server, but it can be deployed on any common servlet container. Java is obviously required to be also installed on the server.

GeoServer provides an implementation of WMS, WFS and WCS specifications. It also includes the Transaction and LockFeature WFS operations to allow geospatial data editing from remote. It features an integrated AJAX viewer based on OpenLayers to usefully enable data preview. User interface, available into a number of different languages, provides easy-to-use configuration tools which free users from writing complex configuration files. GeoServer is able to read data in a variety of both raster and vector formats and has a mature support for Spatial DBMSs. It also provides full SLD support and multiple data output formats, including the KML for an easy integration with Google Maps and Google Earth.

5.2.2 GRASS GIS

Geographic Resources Analysis Support System, commonly referred to as GRASS GIS, is a Geographic Information System (GIS) used for data management, image processing, graphics production, spatial modeling, and visualization of many types of data. It is Free Software/Open Source released under GNU General Public License (GPL). GRASS GIS is an official project of the Open Source Geospatial Foundation. This software is chosen to process land cover maps before proceeding with classification and validation activities. In particular, GRASS will be used to convert vector maps into raster ones, and produce a map which highlight the differences between two input maps in order to easily discover the misclassified lands. The request of processing is achieved via Web Processing Service (see section 2.2.6 for more details about WPS).

5.1.2.1 GRASS GIS Capabilities

- *Raster analysis:* Automatic rasterline and area to vector conversion, Buffering of line structures, Cell and profile data query, Color table modifications, Conversion to vector and point data format, Correlation / covariance analysis, Expert system analysis , Map algebra (map calculator), Interpolation for missing values, Neighborhood matrix analysis, Raster overlay with or without weight, Reclassification of cell labels, Resampling (resolution), Rescaling of cell values, Statistical cell analysis, Surface generation from vector lines.
- *3D-Raster (voxel) analysis:* 3D data import and export, 3D masks, 3D map algebra, 3D interpolation (IDW, Regularized Splines with Tension), 3D

Visualization (isosurfaces), Interface to Paraview and POVray visualization tools.

- *Vector analysis:* Contour generation from raster surfaces (IDW, Splines algorithm), Conversion to raster and point data format, Digitizing (scanned raster image) with mouse, Reclassification of vector labels, Super positioning of vector layers.
- *Point data analysis:* Delaunay triangulation, Surface interpolation from spot heights, Thiessen polygons, Topographic analysis (curvature, slope, aspect), LiDAR.
- *Image processing:* Canonical component analysis (CCA), Color composite generation, Edge detection, Frequency filtering (Fourier, convolution matrices), Fourier and inverse Fourier transformation, Histogram stretching, IHS transformation to RGB, Image rectification (affine and polynomial transformations on raster and vector targets), Ortho photo rectification, Principal component analysis (PCA), Radiometric corrections (Fourier), Resampling, Resolution enhancement (with RGB/IHS), RGB to IHS transformation, Texture oriented classification (sequential maximum a posteriori classification), Shape detection, Supervised classification (training areas, maximum likelihood classification), Unsupervised classification (minimum distance clustering, maximum likelihood classification).
- *DTM-Analysis:* Contour generation, Cost / path analysis, Slope / aspect analysis, Surface generation from spot heights or contours.
- *Geocoding:* Geocoding of raster and vector maps including (LiDAR) point clouds.
- *Visualization:* 3D surfaces with 3D query (NVIZ), Color assignments, Histogram presentation, Map overlay, Point data maps, Raster maps, Vector maps, Zoom function.
- *Map creation:* Image maps, Postscript maps, HTML maps.
- *SQL-support:* Database interfaces (SQLite, PostgreSQL, mySQL, ODBC).
- *Geostatistics:* Interface to “R” (a statistical analysis environment), Matlab.

5.2.3 Glassfish

Glassfish is an open source application server project started by Sun Microsystems for the Java EE platform and now sponsored by Oracle Corporation. The supported version is called Oracle Glassfish Server. Glassfish is free software, dual licensed under two free software licenses: the common

development and distribution license (CDDL) and the CNU general public license (GPL) with the class path exception. Glassfish is the reference implementation of Java EE and as such supports Enterprise JavaBeans, JPA, Java Server Faces, JMS, RMI, Java Server pages, servlets, etc. This allows developers to create enterprise applications that are portable and scalable, and that integrate with legacy technologies. Optional components can also be installed for additional services. Built on a modular kernel powered by OSGi, Glassfish runs straight on top of the Apache Felix implementation. It also runs with Equinox OSGi or knopflerfish OSGi runtimes.

Glassfish is based on source code released by Sun and Oracle Corporation's Toplink persistence system. It uses a derivative of Apache Tomcat as the servlet container for serving Web content, with an added component called Grizzly which uses Java New I/O (NIO) for scalability and speed.

5.2.4 MySQL

The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack (and other 'AMP' stacks). LAMP is an acronym for "Linux, Apache, MySQL, and Perl/PHP/Python." Free-software-open source projects that require a full-featured database management system often use MySQL. MySQL is a relational database management system (RDBMS), and ships with no GUI tools to administer MySQL databases or manage data contained within the databases. Users may use the included command line tools, or use MySQL "front-ends", desktop software and web applications that create and manage MySQL databases, build database structures, back up data, inspect status, and work with data records. The official set of MySQL front-end tools, MySQL Workbench is actively developed by Oracle, and is freely available for use. MySQL is offered with two different editions: the open source MySQL Community Server and the commercial Enterprise Server. MySQL Enterprise Server is differentiated by a series of commercial extensions which install as server plugins, but otherwise shares the version numbering system and is built from the same code base. The developers release minor updates of the MySQL Server approximately every two months. The sources can be obtained from MySQL's website or from MySQL's Bazaar repository, both under the GPL license.

5.3 Technological Choices (Client Side)

5.3.1 NASA World Wind

Released under the NASA Open Source Agreement, World Wind is written in multi-platform Java language and thus runs on Linux, Windows and Mac OS X. First formal release was 1.2 in July 2011; World Wind is available as a highly extensible SDK which allows a full customization of the developed applications. It can be run either as a desktop Java application, or into a Web browser as a Java applet or a Java Web Start application. It integrates both Swing and Abstract Window Toolkit (AWT) Java toolkits and the Java Open Graphics Library (JOGL) for maximizing graphics capabilities. World Wind accommodates any desired data format and provides open-standard interfaces to GIS services and databases. It can be deployed as a WMS server and enables to locate on or above the globe both 2D objects (e.g. lines, polygons, markers, callouts, and multimedia viewers) and 3D objects built up from geometric primitives (e.g. parallelepipeds, spheres, and extruded polygons).

A rich collection of spatial datasets is natively provided by World Wind. This includes both satellite imagery with multiple resolutions (e.g. BlueMarble, SGS Orthophoto/Urban Area Orthophoto, and Microsoft Virtual Earth Imagery) and standard Digital Elevation Models such as SRTM, ASTER, and USGS National Elevation Dataset (NED). Both imagery and DEMs are dynamically served by NASA and USGS WMS servers. However World Wind allows users to access any other OGC-compliant WMS, serving both geo-referenced images or data to be projected on the globe, and also DEMs to be superimposed on the geoid model implemented within the platform. Full control of the terrain model strongly distinguishes World Wind from the majority of virtual globes.

5.3.2 Policrowd

PoliCrowd is a citizen science application developed by some students from Politecnico di Milano including me. PoliCrowd's purpose is to engage the general public in reporting and describing Points-Of-Interest (POIs) related to the fields of tourism, culture, sports, and transportation. It allows almost any person to synthesize his/her personal knowledge about local features and

facts, that is a highly-precious source of geographic information unavailable using global mapping techniques like satellite imagery. Everyone can thus use PoliCrowd without any specific background knowledge, provided of course the availability of a mobile device and the ability to use it to make reports.

PoliCrowd2.0 is available from http://geomobile.como.polimi.it/policrowd2.0
As mentioned above, user interaction with citizen field-collected data is provided through both traditional bi-dimensional applications and a three-dimensional platform featuring also some advanced collaboration-enabling functionalities.

One module of PoliCrowd, which is developed by me, allows users (citizen scientists) to provide comments and upload multimedia files like images, videos and audios related to a specific Point-Of-Interest (POI). This aspect is also used in Open Land Map in order to consolidate the validation and classification functions e.g. a user can classify an area to a particular CORINE land cover category, and beside the classification task, the user can also upload an image, a video, or provide a comment on what he/she observes locally about the considered land, and by doing that, users support their validation claims.

5.4 Open Land Map Architecture

The architecture of a system is a representation which is organized in a way that supports reasoning about the structures and behaviors of it. Our system architecture is divided in two modules, a server side and a client side (see figure 5.3.3). The client side, built on top of NASA World Wind core and build as a desktop application, allows publishing land cover maps via GeoServer API, and permits to retrieve the published maps as raster layers using Web Map Service (WMS). It also provides a connection to GRASS via Web Processing Service (WPS) to process the land cover maps e.g. rasterization of vector maps and producing maps which highlight the differences between two cover maps. In order to store information about users, validation data, votes and other stuff, the client side interacts with Open Land Map server side which is responsible of managing and controlling the data via RESTful CRUD (create, read, update and delete) services. Open Land Map server part is written in JAVA and runs within GlassFish application server. The data are stored in a database managed by MySQL DBMS (Data Base Management System).

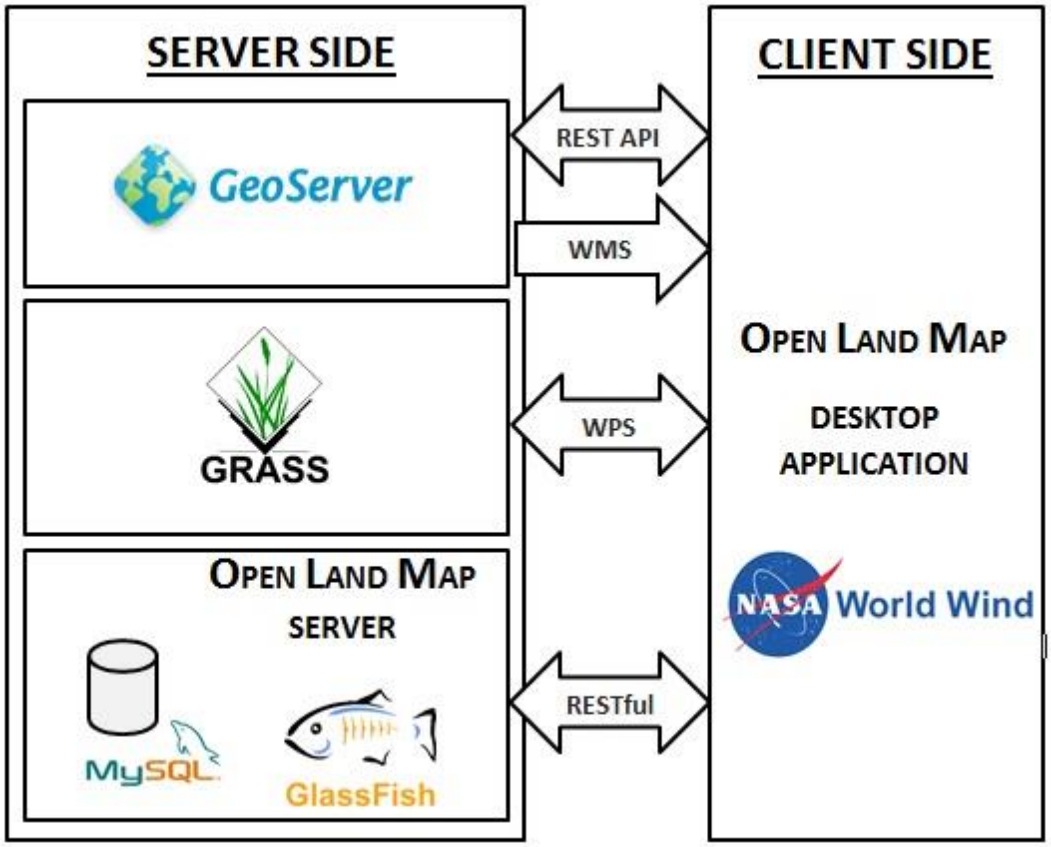


Figure 5.4 Open Land Map system architecture

5.4.1 The Server Side

The server side components have been chosen in order to achieve the needed behavior, GeoServer provides a RESTful interface through which clients can retrieve information about an instance and make configuration changes. Using the REST interface's simple HTTP calls, clients can configure GeoServer without needing to use the Web Administration Interface. REST is an acronym for "Representational State Transfer". REST adopts a fixed set of operations on named resources, where the representation of each resource is the same for retrieving and setting information. In other words, we can retrieve (read) data in an XML format and also send data back to the server in similar XML format in order to set (write) changes to the system. Operations on resources are implemented with the standard primitives of HTTP: *GET* to read; and *PUT*, *POST*, and *DELETE* to write changes. Each resource is represented as a URL.

The OGC Web Processing Service (WPS) interface standard provides rules for standardizing how inputs and outputs (requests and responses) for invoking geospatial processing services as a web service. GRASS GIS can be used as backend in WPS service frameworks to process land cover maps.

We decided to build the Open Land Map server part in somehow to be based on using RESTful web services. RESTful web services are built to work best on the Web applications. Representational State Transfer (REST) is an architectural style that specifies constraints, such as the uniform interface, that if applied to web service induce desirable properties, such as performance, scalability, and modifiability, that enable services to work best on the Web. In the REST architectural style, data (entities) and functionality are considered resources and are accessed using Uniform Resource Identifiers (URIs), typically links on the Web. The resources are acted upon by using a set of simple, well-defined operations. Open Land Map client and server sides exchange representations of resources by using HTTP standardized communication protocol. We decided to develop with GlassFish Server as it is the reference implementation of Java EE and as such supports Enterprise JavaBeans, JPA, JavaServer Faces, JMS, RMI, JavaServer Pages, servlets, RESTful web services, etc. This allows creating enterprise and web applications that are portable and scalable, and that integrate with legacy technologies. The open source MySQL server is chosen in order to manage the data related to the users, multimedia files, and validation information.

5.4.2 The Client Side

The client side is built on top of NASA World Wind core and available as a desktop application. We decided to use Java Swing API which provides a Graphical User Interface (GUI) for our java application. Swing is a platform-independent, Model-View-Controller GUI framework for Java, which follows a single-threaded programming model. Additionally, this framework provides a layer of abstraction between the code structure and graphic presentation of a Swing-based GUI.

The client side communicates with GeoServer via GeoServer REST API in order to retrieve information about some instances, make configuration changes without needing to use Web Administrator Interface, share and edit geospatial layers. The client side accesses the data from GeoServer in the KML format and renders them as rasters on NASA World Wind virtual globe using

Web Map Service. It also interacts with GRASS GIS using Web Processing Service to handle and process the land cover maps. It transmits the actions performed by the users to the server side using RESTful web services, these services in turn execute some specific control actions to change the state of the entities stored in the database. The user can upload land cover maps and cartographic data e.g. orthophotos into the application using the client side, of course the cartographic auxiliary data would support the user in the validation activities.

5.5 Application View

- The application Public view represents default access window. Includes only registration and login functionalities.
- The application Citizen Private view (only for registered citizen scientists). This part allows citizen scientists to perform some specific and easy functions.
- The application Expert Private view (only for registered expert users). This instead, allows expert users to perform all the functionalities offered by Open Land Map.
- The application Admin view for the Administrator. Provides an interface to manage registered users (e.g. delete and send notification messages to users).

I) Application view for the initial access

Name: application view public
Description: represents the default entry window of the application, where users can register to the application or login.
User Group: not registered users, registered expert users, registered citizen scientists and admin.
Use Cases: “register” and “Login”.

II) Application view for registered citizen scientists

Name: application view citizen private
Description: it allows registered citizen scientists to do some easy activities that don’t require a good knowledge in GIS domain.
User Group: registered citizen scientists.
Use Cases: “validate pixels”, “vote pixels” and “add multimedia files”.

III) Application view for registered expert users

Name: application view expert private

Description: it gives registered expert users the privilege to use all the functionalities of the application.

User Group: registered expert users.

Use Cases: “validate pixels”, “vote pixels”, “add multimedia files”, “manage layers”, “upload maps”, “publish maps via GeoServer”, “match classes”, “process maps” and “calculate accuracy indexes”.

IV) Application view for Admin

Name: application view admin

Description: a view for the administrator of the application, it allows to manage the list of registered users e.g. delete or notify a user.

User Group: admin.

Use Cases: “view users” and “manage users”.

Chapter 6

Land Cover Accuracy Assessment

The main objective of accuracy assessment is to derive a quantitative description of the accuracy of global land cover map. This is a nontrivial task, and it must be recognized that there is no one universal “best” method of accuracy assessment, but rather a suite of methods of varying value and applicability for any given map and purpose. The selection of an approach for map accuracy assessment should recognize both the limits of the data (e.g. impacts of mixed pixels) and purpose of the accuracy assessment (e.g. the different accuracy requirements of diverse user communities or the needs of map producers in evaluating mapping methods etc.). The basis of accuracy assessment is simply the comparison of the class labeling derived from an image classifier against some ground reference data set. It can, however, be a distinctly challenging analysis and one that is often undertaken poorly by the geosciences and remote sensing community.

Accuracy assessment has evolved considerably history of remote sensing. The issue is, however, complex, partly because of the great diversity of motivations and objectives in accuracy assessment as well as a set of difficulties that are widely encountered. For example, interest may focus on the accuracy of the classification as a whole or on just a sub-set of the classes mapped, and then also from the users and producers perspective depending on the importance of different types of errors. There may also be variations relating to issues such as the cost of different errors which should be integrated into the analysis. Consequently, there is no single universally accepted approach to accuracy assessment, but a variety of approaches that may be used to meet the varied objectives that are encountered in remote

sensing research. There are however, some general issues that are common to accuracy assessment. Indeed, two broad types of accuracy assessment are popular within remote sensing related research.

First, non-site specific accuracy which involves an evaluation of the similarity of the predicted and actual land cover representations in terms of the areal extent of classes in the mapped region. The focus of this type of accuracy assessment is, therefore, on the quantity or coverage of the land cover classes within the region. While this can sometimes be a useful approach to accuracy assessment it is insensitive to the geographical distribution of the classes in the region mapped. Thus, a classified image which contained the classes in correct proportions but in incorrect locations would be deemed perfect. This limitation to the non-site specific approach to accuracy assessment often renders it unsuitable for use in validation programs and so it is used relatively infrequently.

Instead, the second type of approach to accuracy assessment, based on site-specific measures, is more widely used. Site-specific accuracy assessment involves the comparison of the predicted and actual class labels for a set of specific locations within the region classified. Thus, for example, for a typical remote sensing scenario, the actual and predicted class label information for a sample of pixels drawn from the region mapped are compared. This comparison is typically based upon the cross-tabulation of the actual and predicted class labels. This latter cross-tabulation provides the error or confusion matrix which should provide a wealth of information to summarize the quality of the classification. Indeed the confusion matrix may be used to derive a suite of quantitative measures to express classification accuracy, on both an overall and per-class basis. Site specific accuracy assessment is extremely popular in remote sensing and there is a large literature that promotes it as a ‘best practice’.

6.1 Issues and Constraints of Concern

There are many issues to be considered in an accuracy assessment (e.g. Congalton and Green, 1999; Foody, 2002), but the following are of particular concern.

1. It is effectively impossible to produce a land cover map that is completely accurate and satisfies the needs of all (Brown et al., 1999). The different viewpoints and components of classification accuracy also act to ensure that

there is no single all-purpose universal measure of accuracy. The purpose of the map should, therefore, be considered in its production and assessment. In most mapping applications and map evaluations, interest is focused on overall map accuracy. It may, however, be more appropriate in some circumstances to focus on other features (Lark, 1995; Boschetti et al., 2004). This has important implications to the evaluation of map accuracy. Commonly, a relatively subjectively defined target of greater than 85 percent overall accuracy with reasonably equal accuracy across the classes is specified, but this need not be appropriate for all maps or applications.

2. To avoid bias, a sample of pixels independent of that used to train a classification should be used in the accuracy assessment (Swain, 1978; Hammond and Verbyla, 1996). The sample design used to acquire the testing set of samples used to evaluate classification accuracy is of fundamental importance and must be considered when undertaking an accuracy assessment and interpreting the accuracy metrics derived (Stehman and Czaplewski, 1998; Stehman, 1995, 1999a).

3. Since the accuracy assessment is based on a sample of cases, confidence intervals should ideally accompany the metrics of accuracy contained in an accuracy statement (Rosenfield et al**.,** 1982; Thomas and Allcock, 1984).

4. The nature of the techniques used to map land cover from the remotely sensed imagery has important implications. For example, with some classifiers it is relatively easy to derive a measure of the uncertainty of the class allocation made for each pixel (e.g. maximum likelihood classification), while with others the ability to derive an uncertainty metric is limited (e.g. parallelepiped classification).

5. The use of site-specific approaches to accuracy assessment based on the confusion matrix requires accurate registration of the map and ground data sets. Some degree of tolerance to misallocation can be integrated into accuracy assessment (Hagen, 2003); although most assessments assume implicitly that the data sets are perfectly registered. The importance of misregistration as a source of nonthematic error in the confusion matrix is most apparent in regions where the land cover mosaic is fragmented (Estes et al., 1999; Loveland et al., 1999).

6. For conventional (hard) classifications, in which each image pixel is allocated to a single class, it is assumed that the pixels are pure (i.e. each pixel

represents an area that comprises homogeneous cover of a single land cover class). Any hard class allocation made for a mixed pixel will, to some extent, be erroneous, and alternative approaches to accuracy assessment should be adopted if the proportion of mixed pixels is large. In general, the proportion of mixed pixels increases with a coarsening of the spatial resolution of the imagery.

7. Errors are commonly treated as being of equal magnitude. If some errors are more damaging than others, it may be possible to weight their effect in the assessment of classification accuracy.

8. The ground or reference data may contain error and thus misclassification does not always indicate a mistake in the classification used to derive the map. In reality, therefore, the assessment of maps commonly undertaken is one of agreement or correspondence with the ground data rather than strictly of thematic accuracy. In some instances, it may be useful to include some measure of confidence in the ground data used (Scepan, 1999; Estes et al., 1999).

9. The pixel is the basic spatial unit of the analysis. Maps could be produced using other spatial units. For example, the minimum mapping unit could be set at a size larger than the image pixel size. The use of large units may help in reducing the effect of spatial misregistration problems. With soft/fuzzy classifications and with super-resolution mapping, where the aim is to map at a scale finer than the source data, the problems of spatial misregistration in conventional approaches to accuracy assessment are likely to be large.

10. The same set of class definitions/protocols should be used in the image classification as in the ground data; that is, the class labels used in both data sets should have the same meaning. Approaches to explore and accommodate differences in the meaning of class labels may be useful if the classes have been defined differently in the data sets (Comber et al., 2004). If different classification schemes have been used, it is still possible to evaluate the level of agreement between a map and the ground data using a cross-tabulation of class labels.

11. The confusion matrix should be presented as well as the summary metrics of accuracy derived from it. To avoid problems associated with normalization (Stehman, 2004a), the raw matrix should be provided and the sample design used in its generation specified.

6.2 Basic Approach

The basis of the suggested approach to accuracy assessment is the confusion or error matrix. This matrix provides a cross tabulation of the class label predicted by the image classification analysis against that observed in the ground data for the test sites (Figure 6.2). The confusion matrix provides a great wealth of information on a classification. It may, for example, be used to provide overall and per-class summary metrics of land cover classification accuracy as well as to refine areal estimates or aspects of the classification analysis in order to meet specific user requirements. Moreover, the confusion matrix is relatively easy to interpret and is familiar to both the map user and producer communities.

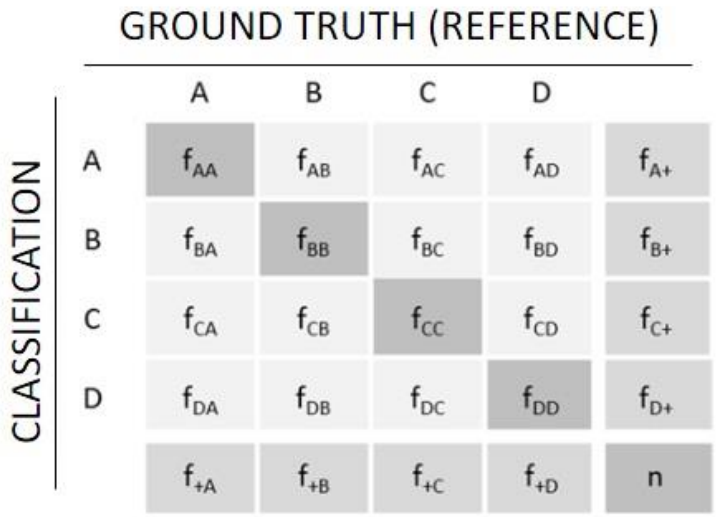


Figure 6.2 Layout of a typical confusion or error matrix.

The use of the confusion matrix in accuracy assessment applications is based on a number of important assumptions. In particular, it is assumed that each pixel can be allocated to a single class in both the ground and map datasets; and also these two data sets have the same spatial resolution and are perfectly registered. All of these assumptions are often not satisfied in remote sensing. In some instances, deviation from the assumed condition is relatively unimportant (e.g., if testing pixels are drawn from very large homogenous regions of the classes then the impact of misregistration of the data sets is unlikely to have a major impact on accuracy assessment) but in other

situations they may lead to significant error and misinterpretation (e.g., if the land cover mosaic is very fragmented and mixed pixels are common).

Interpretation of the confusion matrix also requires consideration of the sample design used to acquire the testing set. Since the testing set is a sample, its relationship to the population (the map) is important. Confusion matrices and associated metrics of accuracy derived from a land cover map using simple random or stratified random sampling may, for example, differ markedly if there are interclass differences in the accuracy of classification; ideally a probability sample design should be used.

Map accuracy may be assessed using a variety of units (e.g., pixels, blocks of pixels or polygons such as land parcels). For the purposes of our study the accuracy assessment is based on pixels. Given that the pixel is the smallest spatial unit, assessing map accuracy on a per-pixel basis is somewhat ambitious. A coarser minimum mapping unit may be more appropriate, but pixel-based assessment is common and, providing its limitations are realized, can be useful. Given that there is a trade-off between accuracy and spatial resolution, with aggregation acting to reduce misregistration errors, knowledge of the relationship between accuracy and resolution may help in the specification of an appropriate cell size for a map.

6.3 Thematic Accuracy

For global land cover maps, accuracy assessment aims to provide an index of how closely the derived class allocations depicted in the thematic land cover map represent reality. In essence, the summary metrics of accuracy provide a measure of the degree of correctness in the class allocations in the map. Attention is, therefore, focused on thematic accuracy. The confusion matrix is well suited to this task (Figure 6.2). The cases that lie on the main diagonal of the matrix represent those correctly allocated, while those in the off-diagonal elements represent errors. Two types of thematic error, omission and commission, are possible and both may be readily derived from a confusion matrix. An error of omission occurs when a case belonging to a class is not allocated to that class by the classification. Such a case has been erroneously allocated to another class, which suffers an error of commission.

A major problem in the use of the confusion matrix and associated accuracy metrics, however, is that it may contain nonthematic error. In particular, error

due to misregistration of the data sets is commonly included. It is important to be aware of this source of error, as the error due to misregistration may be larger than the thematic error actually present in the map. Sometimes it may be appropriate to spatially adjust locations of testing sites to account for known misregistration effects or to attempt to directly include some tolerance to spatial misregistration effects into the accuracy assessment.

6.3.1 Measures of Accuracy

A variety of measures of overall and per-class accuracy can be derived from the confusion matrix. Metrics of overall accuracy provide an indication of the quality of the entire land cover map. For overall accuracy, attention is focused on the main diagonal of the confusion matrix (Congalton 2009).

$$OA = \frac{\sum f_{ii}}{n} \qquad (i = A,B,C,D)$$

Equation 1: Overall Accuracy.

Sometimes interest is focused on the accuracy with which a particular land cover class is represented. Metrics to describe per-class accuracy can be readily derived from the confusion matrix. Clearly, this may be approached from two perspectives, depending on whether the data in the confusion matrix are read vertically or horizontally (Story and Congalton, 1986). If attention is focused on the accuracy of the map as a predictive device, concern is with errors of commission. In this situation what is generally termed user's accuracy, UA, may be derived, which is based on the ratio of correctly allocated cases of a class relative to the total number of testing cases allocated to that class(Congalton 2009).

$$UA_i = \frac{f_{ii}}{f_{i+}} \qquad (i=A,B,C,D)$$

Equation 2: User Accuracy.

$$CE_i = 1 - UA_i \ (i=A,B,C,D)$$

Equation 3: Commission Error.

The resulting metric provides an indication of the probability that a pixel allocated to a particular land cover class actually represents that class on the ground. Reading the matrix in the alternative way, from the map producer's perspective, the focus is on errors of omission. What is generally termed producer's accuracy, PA, may be derived from the ratio of cases correctly allocated to a class to the total number of cases of that class in the testing set (Congalton 2009).

$$PA_i = \frac{f_{ii}}{f_{+i}} \qquad (i=A,B,C,D)$$

Equation 4: Producer Accuracy.

$$OE_i = 1 - PA_i \ (i=A,B,C,D)$$

Equation 5: Omission Error.

Many summary metrics may be derived from a confusion matrix to express accuracy. The two most widely used measures of land cover map accuracy are the percentage of correctly allocated cases and the kappa coefficient of agreement. These give a guide to the overall quality of the map. Kappa, introduced to remote sensing in the early 1980s, in particular, Congalton and Green (2009, p. 105) state that 'Kappa analysis has become a standard component of most every accuracy assessment and is considered a required component of most image analysis software packages that include accuracy assessment procedures'. Indeed, Kappa is published frequently and has been incorporated into many software packages.

Kappa is the proportion of agreement after chance agreement is removed. From the error matrix, Kappa calculated as following:

$$k = \frac{\frac{1}{n}\sum f_{ii} - \frac{1}{n^2}\sum f_{i+}f_{+i}}{1 - \frac{1}{n^2}\sum f_{i+}f_{+i}}$$

$$k_i = \frac{\frac{f_{ii}}{n} - \left(\frac{f_{i+}}{n} * \frac{f_{+i}}{n}\right)}{\frac{1}{2}\left(\frac{f_{i+}}{n} + \frac{f_{+i}}{n}\right) - \left(\frac{f_{i+}}{n} * \frac{f_{+i}}{n}\right)} \quad (i = A, B, C, D)$$

Equation 6: Kappa Standard.

The value of Kappa is between 1 and –1, the higher the value, the stronger the agreement. Although the kappa coefficient has been widely promoted for accuracy assessment, there are sufficient concerns with its use that it cannot be recommended as general measure of map accuracy. Foody (2008) exposed some of the conceptual problems with the standard Kappa, the arguments used to promote the use of the kappa coefficient are fundamentally flawed First, chance agreement is of no particular concern to classification accuracy assessment; it does not matter if a pixel is allocated correctly by chance or design. Thus, chance correction is unnecessary. Even if chance correction was desired the standard method to calculate the agreement due to chance is inappropriate for the typical remote sensing scenario and alternatives that are not dependent on the confusion matrix's row marginal may be used.

Critically, however, chance correction is unnecessary and the derived coefficient just a downward scaled version of overall accuracy. Second, although only a minor and possibly pedantic issue, the kappa coefficient does not actually use all of the matrix's elements directly but rather only its marginal values. Third, the existence of popular scales for the evaluation of kappa may be useful but these scales are necessarily arbitrary and not of universal applicability. Fourth and finally, the kappa coefficient is not unique in relation to comparisons. In order to rigorously compare two accuracy values all that is normally required are appropriate estimates of the accuracy and the variance of the accuracy for each classification. It is also recommended to replace these indices with a more useful and simpler approach that focuses on two components of disagreement between maps in terms of the quantity and spatial allocation of the categories.

Allocation disagreement (AD) can be considered as difference between classified data and reference data due to incorrect spatial location of pixels on the classification. Allocation disagreement is always an even number of pixels because allocation disagreement always occurs in pairs of misallocated pixels.

$$AD_i = 2 * min\left(\frac{f_{+i}}{n} - \frac{f_{ii}}{n},\ \frac{f_{i+}}{n} - \frac{f_{ii}}{n}\right) \qquad i = (A, B, C, D)$$

$$AD = \frac{\sum AD_i}{2} \qquad i = (A, B, C, D)$$

Equation 7: Allocation Disagreement.

Quantity disagreement (QD) is defined as the difference between the reference data and classified data based upon mismatch of class proportion.

$$QD_i = \left|\frac{f_{+i}}{n} - \frac{f_{i+}}{n}\right| \qquad i = (A, B, C, D)$$

$$QD = \frac{\sum QD_i}{2} \qquad i = (A, B, C, D)$$

Equation 8: Quantity Disagreement

The total disagreement is the sum of the Quantity disagreement and Allocation disagreement (Pontius and Millones 2011).

Other metrics of overall and per-class accuracy can be derived from a confusion matrix. Each metric focuses on different aspects of accuracy and may vary in utility between map users. Since it is impossible to anticipate the needs of all users, the confusion matrix itself should be provided so that the user may derive a specific measure of interest. To maintain flexibility, the raw and not a normalized matrix should be provided (Stehman, 2004a). Values of

acceptable classification accuracy presented in literature differ considerably. The accuracy over 70 % is considered somehow as adequate, whereas Foody (2002) recommends values over 85 %. Landis and Koch (1977) proposed categories for assessment of the classification performance measured by Kappa value as poor (<0.41), moderate (0.41-0.61), good (0.61-0.81), and excellent (>0.81).

Conclusion

Open Land Map is an online open geo-application for visualization, validation and crowd-sourcing of land cover datasets using NASA World Wind platform. It enables easier, more efficient data sharing and information service helping organizations, communities and individuals to visualize, process, compare, validate, evaluate and assess their land coverage maps.

Open Land Map adopts crowdsourcing in order to defy the large spatial discrepancies between land cover products by taking advantage of the observations and evaluations of the users. The application offers functionalities both to citizen scientists and expert users; the citizen scientists are employed in the land cover validation tasks, letting them provide in-situ observations and participate actively to assess validated data in order to produce more accurate and consistent maps. Expert users instead can use various tools for processing their data; this to make their heterogeneous land cover data comparable and standardized in somehow to be able to proceed with the validation mission because it is not possible to compare diverse resolution land cover maps or maps which have different number or type of classes without doing some initial processing and matching steps; Open Land Map actually offers all the necessary tools to overcome such problems of incoherency.

After land cover validation phase, the application permits to make a sort of comparison between the original maps and the classified ones regarding the accuracy indexes. Of course one goal of this application is to produce more accurate land cover maps in order to let the organizations have a better view of the land cover which is important in the analysis of environmental processes and problems that have to be comprehended if living standards are to be improved or maintained at some required levels.

In future, to reach out to a wider audience and build a sustainable community around, Open Land Map will require a step change in the application, i.e. addition of social networking tools and feedback mechanisms that motivate individuals to participate. The gaming aspect (gamification) will also be interesting to monitor as it has the potential to massively increase the amount of data collected in a very different way to community building that is built around a common goal of improving land cover datasets.

Acknowledgments

Special Thanks To …

Dr. BROVELLI Maria
Dr. HOGAN Patrick
Dr. UMUHOZA Eric
Dr. ZAMBONI Giorgio

… For Their Support and Collaboration

References

1. Running, S. W. 2008. “Ecosystem Disturbance, Carbon, and Climate.” Science 321: 652–3. Scepan, J. 1999. “Thematic Validation of High-Resolution Global Land-Cover Data Sets.” Photogrammetric Engineering and Remote Sensing 65: 1051–60.

2. Bounoua, L., R. Defries, G. J. Collatz, P. Sellers, and H. Khan. 2002. “Effects of Land Cover Conversion on Surface Climate.” Climatic Change 52: 29–64.

3. Ge, J., J. Qi, B. M. Lofgren, N. Moore, N. Torbick, and J.M. Olson. 2007. “Impact of Land Use/Cover Classification Accuracy on Regional Climate Simulations.” Journal of Geophysical Research 112: D05107.

4. Hibbard, K., A. Janetos, D. P. Van Vuuren, J. Pongratz, S. K. Rose, R. Betts, M. Herold, and J. J. Feddema. 2010. “Research Priorities in Land Use and Land-Cover Change for the Earth System and Integrated Assessment Modeling.” International Journal of Climatology 30: 2118–28.

5. Imaoka, K., H. Fujii, H. Murakami, M. Hori, A. Ono, T. Igarashi, K. Nakagawa, T. Oki, Y. Honda, and H. Shimoda. 2010. “Global Change Observation Mission (GCOM) for Monitoring Carbon, Water Cycles, and Climate Change.” Proceedings of the IEEE 98: 717–34.

6. Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, J. H. M. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder. 2005. “Global Consequences of Land Use.” Science 309: 570–4.

7. Jung, M., K. Henkel, M. Herold, and G. Churkina. 2006. “Exploiting Synergies of Global Land Cover Products for Carbon Cycle Modeling.” Remote Sensing of Environment 101: 534–53.

8. Jun Chen. 14 January 2014, Genève. GlobalLand30: 30m Global Land Cover Data Sets, National Geomatics Center, NASG CHINA, ISPRS GEO SB 02-C1.

9. Foody, L. See, S. Fritz, M. Van der Velde, C. Perger, C. Schill, D. S. Boyd. 2013. Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project, Transactions in GIS, 17(6): 847–860.

10. European Commission, issue 9. December 2013. Science for Environmental Policy, In-depth report: Environmental Citizen Science.

11. Linda See, Alexis Comber, Carl Salk, Steffen Fritz, Marijn van der Velde, Christoph Perger, Christian Schill, Ian McCallum, Florian Kraxner, Michael Obersteiner. 31 July 2013. Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts, Tobias Preis, University of Warwick United Kingdom.

12. Jun Chen. April-May 2014. "ISPRS supports Future Earth initiative with Global Land Cover Information", GEO Informatics magazine for Suveying, mapping and GIS Professionals.

13. Hao Wu. May 16th 2014. Work Plan on GLC Information Service System, National Geomatics Center of China, Suzhou China.

14. Songnian Li, Jun Chen. 15 January 2014 Genève. GLC information portal: Concepts and International collaboration. National Geomatics Center, NASG CHINA, GEO SB 02- C1 ISPRS.

15. Nebert, D., ed. (2004). Developing Spatial Data Infrastructures: The SDI Cookbook. Version 2.0. Global Spatial Data Infrastructure Association. [online]. http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf (2014-08-25).

16. Masser, I. (1999). All shapes and sizes: the first generation of national spatial data infrastructures. International Journal of Geographical Information Science 13(1), 67–84.

17. Craglia, M., Goodchild, M. F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S. and Parsons, E. (2008). Next-Generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. International Journal of Spatial Data Infrastructures Research 3, 146–167.

18. European Parliament and Council (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union L108, 50, 1–14.

19. Steiniger, S. and Hunter, A. J. S. (2012). Free and Open Source GIS Software for building a Spatial Data Infrastructure. in E. Bocher and M. Neteler, eds, Geospatial Free and Open Source Software in the 21st Century. Berlin Heidelberg: Springer-Verlag. 247–261.

20. International Organization for Standardization (2003). ISO 19115:2003 Geographic Information – Metadata. Geneva: ISO.

21. International Organization for Standardization (2005). ISO 19119:2005 Geographic Information – Services. Geneva: ISO.

22. Scharl, A. and Tochtermann, K., eds (2007). The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 Are Shaping the Network Society. London: Springer-Verlag.

23. Open Geospatial Consortium (2007a). Styled Layer Descriptor profile of the Web Map Service Implementation Specification. [online]. http://portal.opengeospatial.org/files/?artifact_id=22364 (accessed: 2014-08-25).

24. Open Geospatial Consortium (2007b). OpenGIS® Catalogue Services Specification. [online]. http://portal.opengeospatial.org/files/?artifact_id=20555 (2014-08-25).

25. Hazzard, E. (2011). OpenLayers 2.10: Beginner's Guide. Birmingham, UK: Packt Publishing.

26. Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial Data Infrastructure in the world of Web 2.0. International Journal of Spatial Data Infrastructures Research 2, 24–32.

27. Onsrud, H. J. (2007). Research and theory in advancing spatial data infastructure concepts. Redlands, CA: ESRI Press.

28. Estes, J. E. and Mooneyhan, D. W. (1994). Of maps and myths. Photogrammetric engineering and remote sensing 60(5), 517–524.

29. Howe, J. (2006a). The rise of crowdsourcing. Wired magazine 14(6), 1–4.

30. Howe, J. (2006b). Crowdsourcing: A Definition. [online]. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html (accessed: 2014-08-25).

31. Harvey, F. (2013). To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information. in D. Sui, S. Elwood and M. Goodchild, eds, Crowdsourcing Geographic Knowledge. Dordrecht: Springer. 31–42.

32. Acohido, B. (2011). Privacy implications of ubiquitous digital sensors. USA Today 26/01/2011, P1B.

33. Haklay, M. (2013). Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. in D. Sui, S. Elwood and M. Goodchild, eds, Crowdsourcing Geographic Knowledge. Dordrecht: Springer. 105–122.

34. Grey, F. (2009). Viewpoint: The age of citizen cyberscience. CERN Courier . April 29, 2009. [online]. http://cerncourier.com/cws/article/cern/38718 (accessed: 2014-08-25).

35. Comber, A. Fisher, P. and Wadsworth, R., 2003 Actor Network Theory: a suitable framework to understand how land cover mapping projects develop? Land Use Policy, 20 299–309.

36. Harvey, F. and Chrisman, N. 1998. Boundary objects and the social construction of GIS technology. Environment and Planning A, 30 1683-1694.

37. Hoeschele, W., 2000. Geographic Information Engineering and Social Ground Truth in Attappadi, Kerala State, India. Annals of the Association of American Geographers 90 293-321.

38. Robbins, P. 2001. Fixed Categories in a Portable Landscape: The Causes and Consequences of Land Cover Categorization. Environment and Planning A. 33 161-179.

39. Bowker G.C., 2000. Mapping biodiversity. International Journal of Geographical Information Science, 14 739-754.

40. Bowker G.C. and Star, S. L., 1996. How things (actor-net)work: Classification, magic and the ubiquity of standards.

41. Rosch, E.H., 1978. Principles of categorization. in Cognition and Categorization, editors E. Rosch and B. Lloyd, (Lawrence Erlbaum Associates, New Jersey) pp27-48.

42. Varzi, A.C., 2001. Introduction. Topoi 20 119-130.

43. Smith, B. 1995. On Drawing Lines on a map. Spatial Information theory: Lecture Notes in Computer Science, 988 475-484.

44. Smith, B., 2001. Fiat Objects, Topoi, 20 131-148.

45. Smith, B. and Mark, D., 2001. Geographical categories: an ontological investigation. International Journal of Geographical Information Science, 15 591- 612.

46. Fisher, P., Wood, J. and Cheng, T. 2004. Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. Transactions of the Institute of British Geographers, 29 106-128.

47. Smith, B. and Mark, D.M., 1998. Ontology and Geographic Kinds. In Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98), editors T. K. Poiker and N. Chrisman, (International Geographical Union, Vancouver) pp308-320.

48. Woodcock, C.E., and A.H. Strahler. 1987. "The factor of scale in remote sensing", Remote Sensing of Environment, 21 311-332.

49. Chavez, P.S., 1992. Comparison of spatial variability in visible and near-infrared spectral images. Photogrammetric Engineering and Remote Sensing, 58 957-964.

50. Fisher, P., 1997. The pixel: A snare and a delusion. International Journal of Remote Sensing, 18 679-685.

51. Forshaw, M.R.B., A. Haskell, P.F. Miller, D.J. Stanley, and J.R.G. Townshend. 1983. Spatial resolution of remotely sensed imagery. A review paper, International Journal of Remote Sensing, 4 497-520.

52. Lillesand, T.M. and Keifer, R.W., 1987. Remote sensing and image interpretation, (John Wiley & Sons, New York).

53. Richards, J.A. and Jia, X., 1993. Remote sensing digital image analysis. (Springer, Berlin).

54. Verstraete, M.M., Pinty, B. and Myneni, R.B., 1996. Potential and limitations of information extraction on the terrestrial biosphere from satellite remote sensing. Remote Sensing of Environment, 58 201-214.

55. Cherrill A, McClean, C., 1995. An investigation of uncertainty in-field habitat mapping and the implications for detecting land-cover change. Landscape Ecology, 10 5-21.

56. Fisher, P.F., 1998. Is GIS hidebound by the legacy of cartography? The Cartographic Journal, 35 5-9.

57. Saura, S., 2002. Effects of minimum mapping unit on land cover data spatial configuration and composition. International Journal of Remote Sensing, 23 4853-4880.

58. Comber, A. Fisher, P. and Wadsworth, R., 2003 Actor Network Theory: a suitable framework to understand how land cover mapping projects develop? Land Use Policy, 20 299–309.

59. Hunter, G.J., 2002. Understanding Semantics and ontologies: they're quite simple really – if you know what I mean. Transactions in GIS, 6 83-87.

60. Smith, J.H., Stehman, S.V., Wickham, J.D. and Yang, L., 2003. Effects of landscape characteristics on land-cover class accuracy, Remote Sensing of Environment, 84 342-349.

61. Griffiths, G.H., Lee, J. and Eversham, B.C., 2000. Landscape pattern and species richness; regional scale analysis from remote sensing, International Journal of Remote Sensing, 21 2685–2704.

62. Forman, R.T.T., 1995. Some general principles of landscape and regional ecology. Landscape Ecology, 10 133-142.

63. Jones, S., 2002. Social constructionism and the environment: through the quagmire. Global Environmental Change, 12 247–251.

64. Carrillo, G. (2012). Web mapping client comparison v.6. [Online]. http://geotux.tuxfamily.org/index.php/en/component/k2/item/291-comparacion-clientes-web-v6 (accessed: 2014-09-02).

65. Congalton, R. G. and K. Green (1999). Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Boca Raton, Lewis Publishers 160 pp.

66. Foody, G. M. (2002). "Status of land cover classification accuracy assessment." Remote Sensing of Environment 80: 185-201.

67. Brown, J. F., T. R. Loveland, D. O. Ohlen and Z. Zhu (1999). "The global land-cover characteristics database: The user's perspective." Photogrammetric Engineering and Remote Sensing 65: 1069-1074.

68. Lark, R. M. (1995). "Components of accuracy of maps with special reference to discriminant analysis on remote sensor data." International Journal of Remote Sensing 16: 1461-1480.

69. Boschetti, L., S. P. Flasse and P. A. Brivio (2004). "Analysis of the conflict between omission and commission in low spatial resolution dichotomic thematic products: The Pareto Boundary." Remote Sensing of Environment 91: 280-292.

70. Swain, P. H. (1978). "Fundamentals of pattern recognition in remote sensing." In: Remote Sensing: The Quantitative Approach. P. H. Swain and S. M. Davis, Eds. New York, McGraw Hill, pp. 136-187.

71. Hammond, T. O. and D. L. Verbyla (1996). "Optimistic bias in classification accuracy assessment." International Journal of Remote Sensing 17: 1261-1266.

72. Stehman, S. V. and R. L. Czaplewski (1998). "Design and analysis for thematic map accuracy assessment: Fundamental principles." Remote Sensing of Environment 64: 331-344.

73. Stehman, S. V. (1995). "Thematic map accuracy assessment from the perspective of finite population sampling." International Journal of Remote Sensing 16: 589-593.

74. Stehman, S. V. (1999a). "Basic probability sampling designs for thematic map accuracy assessment." International Journal of Remote Sensing 20: 2423-2441.

75. Rosenfield, G. H., K. Fitzpatrick-Lins and H. S. Ling (1982). "Sampling for thematic map accuracy testing." Photogrammetric Engineering and Remote Sensing 48: 131-137.

76. Thomas, I. L. and G. M. Allcock (1984). "Determining the confidence level for a classification." Photogrammetric Engineering and Remote Sensing 50: 1491-1496.

77. Hagen, A. (2003). "Fuzzy set approach to assessing similarity of categorical maps." International Journal of Geographical Information Science 17: 235-249.

78. Estes, J., A. Belward, T. Loveland, J. Scepan, A. Strahler, J. Townshend and C. Justice (1999). "The way forward." Photogrammetric Engineering and Remote Sensing 65: 1089-1093.

79. Loveland, T. R., Z. Zhu, D. O. Ohlen, J. F. Brown, B. C. Reed and L. Yang (1999). "An analysis of the IGBP global land-cover characterisation process." Photogrammetric Engineering and Remote Sensing 65: 1021-1032.

80. Comber, A. J., P. F. Fisher and R. A. Wadsworth (2004). "Identifying land cover change using a semantic statistical approach." In: GeoDynamics:

Modelling Spatial Change and Process. P. M. Atkinson, G. M. Foody, S. E. Darby and W. F., Eds. Boca Raton, FL, CRC Press, 440 pp.

81. Stehman, S. V. (2004a). "A critical evaluation of the normalized error matrix in map accuracy assessment." Photogrammetric Engineering and Remote Sensing 70: 743-751.

82. Congalton, R G and Green, K., 2009 Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, 2nd edition, Boca Raton, Lewis Publishers.

83. Foody, G. M., 2008, Harshness in image classification accuracy assessment, International Journal of Remote Sensing, 29, 3137– 3158.

84. Robert Gilmore Pontius Jr & Marco Millones (2011): Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment, International Journal of Remote Sensing, 32:15, 4407-4483.

85. Lndis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics. 33, 159-174.

Appendix A

Application Screenshots

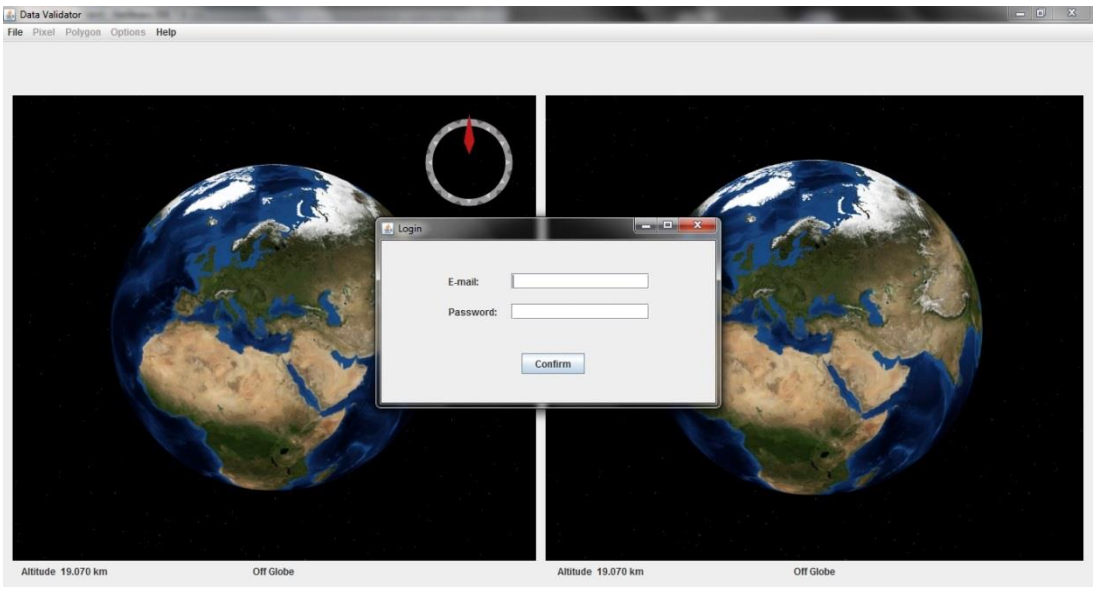


Figure (A) Main Application Window (note the main menu items are inactive until the user logs into the application)

Figure (B) Login frame which asks for the email and a password

Figure (C) Registering frame, it asks some information about the user

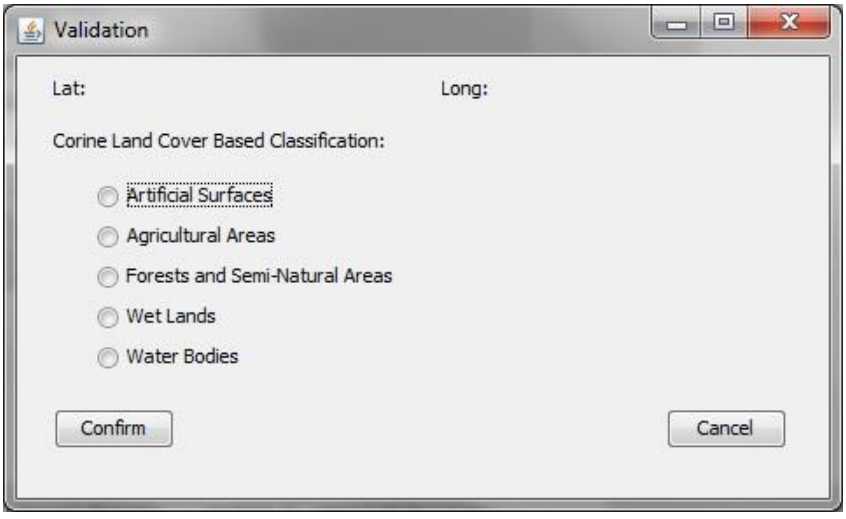


Figure (D) Validation frame, it assigns a land cover class to a certain marker

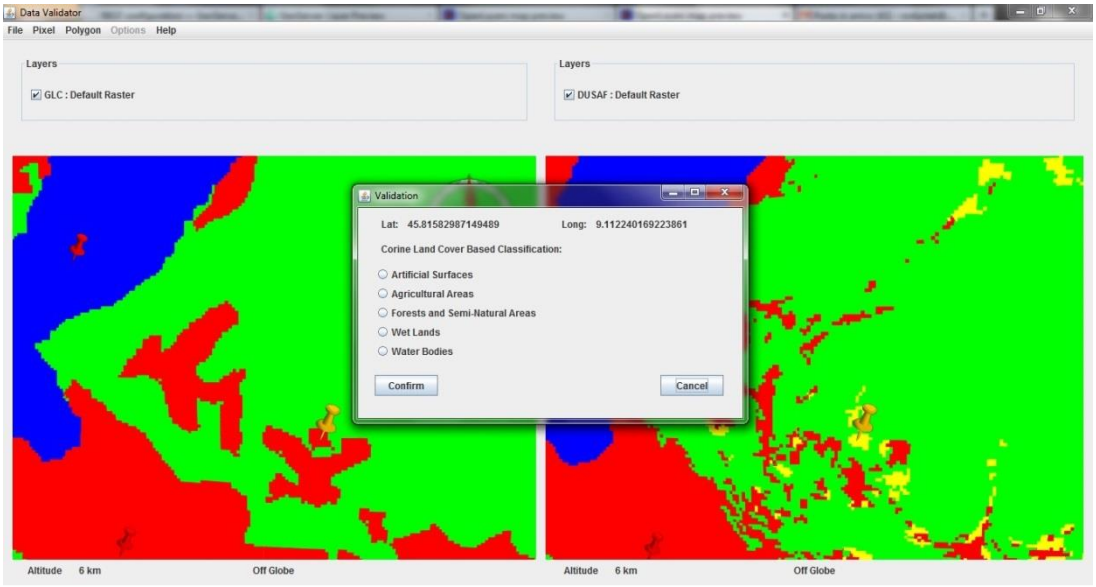


Figure (D) Validation Window (note the difference in land classifications between the two land cover maps right and left)

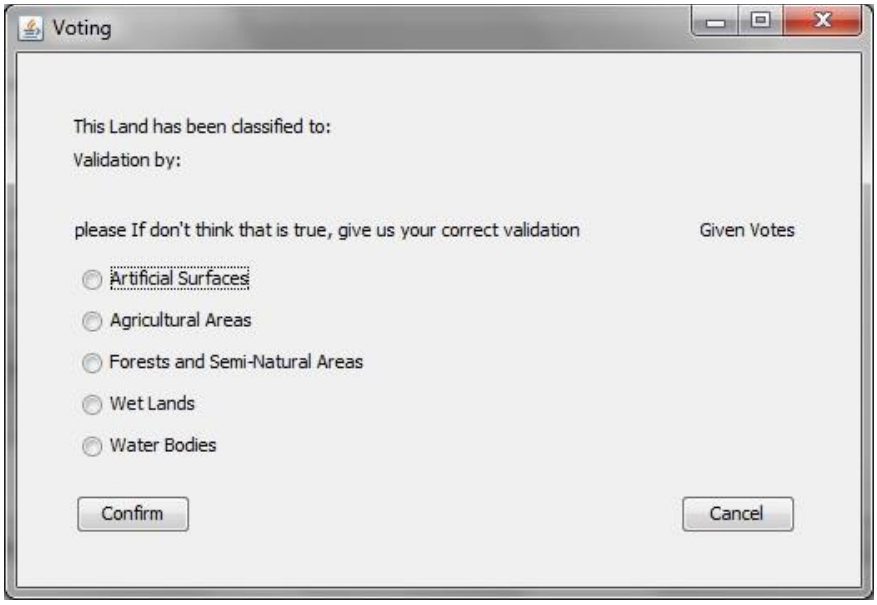


Figure (E) Voting frame, it asks to vote a previously validated markers

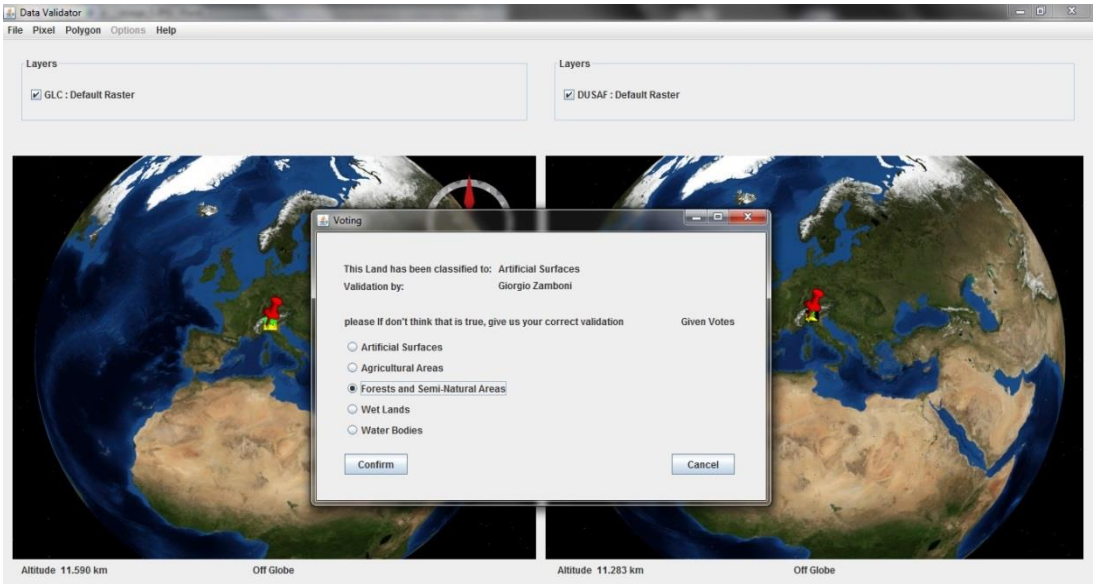


Figure (F) Voting Window

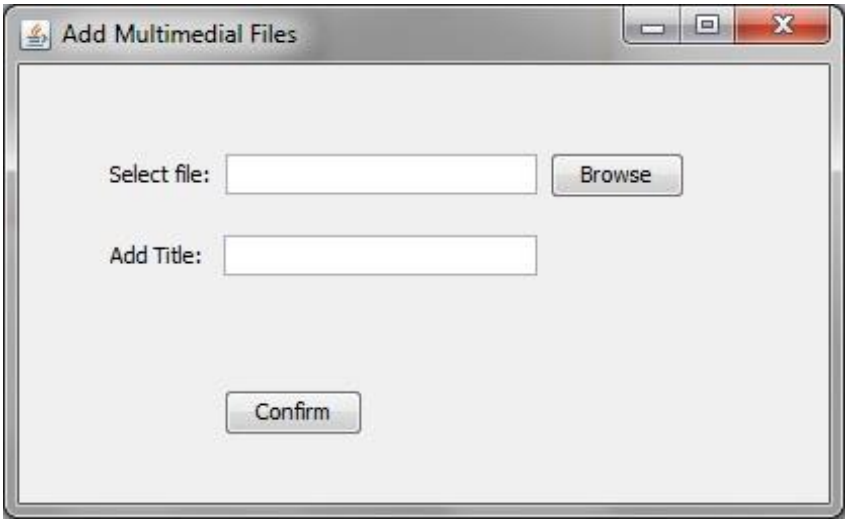


Figure (G) Multimedia files uploading frame

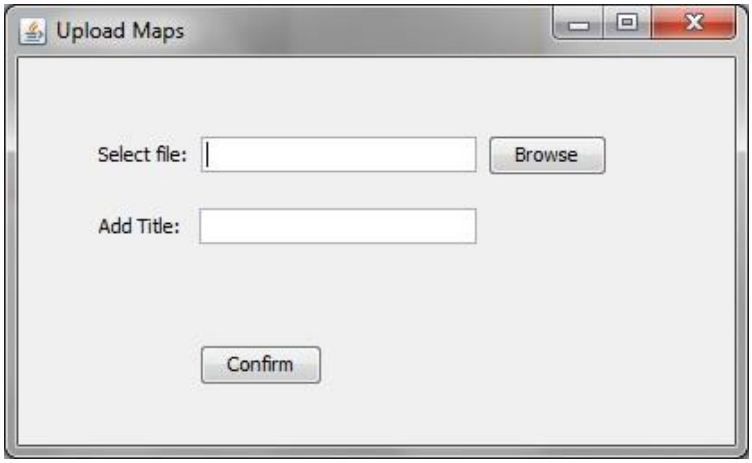


Figure (H) Uploading Maps frame

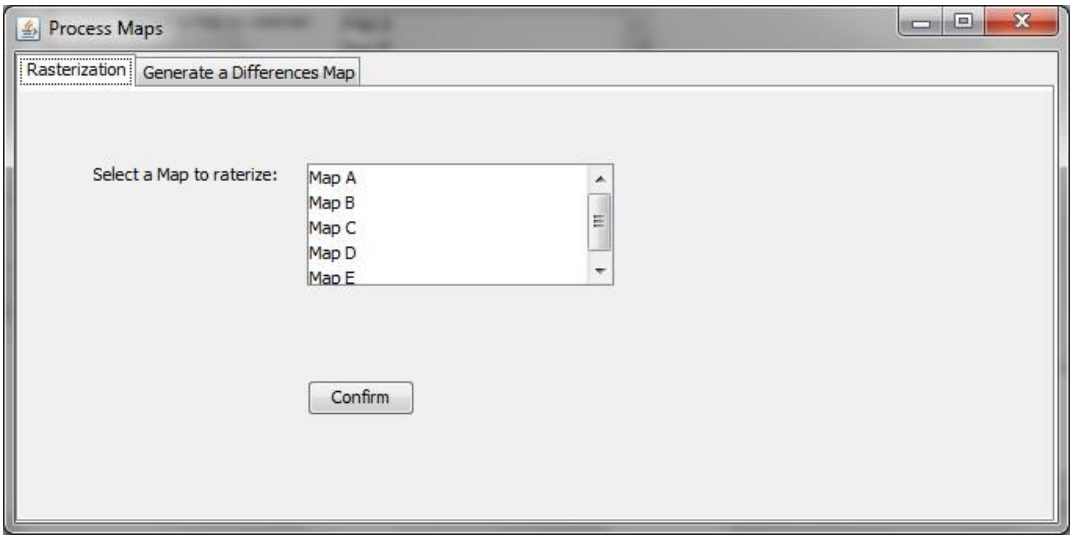


Figure (I) Processing maps frame, this pane allows to do rasterization

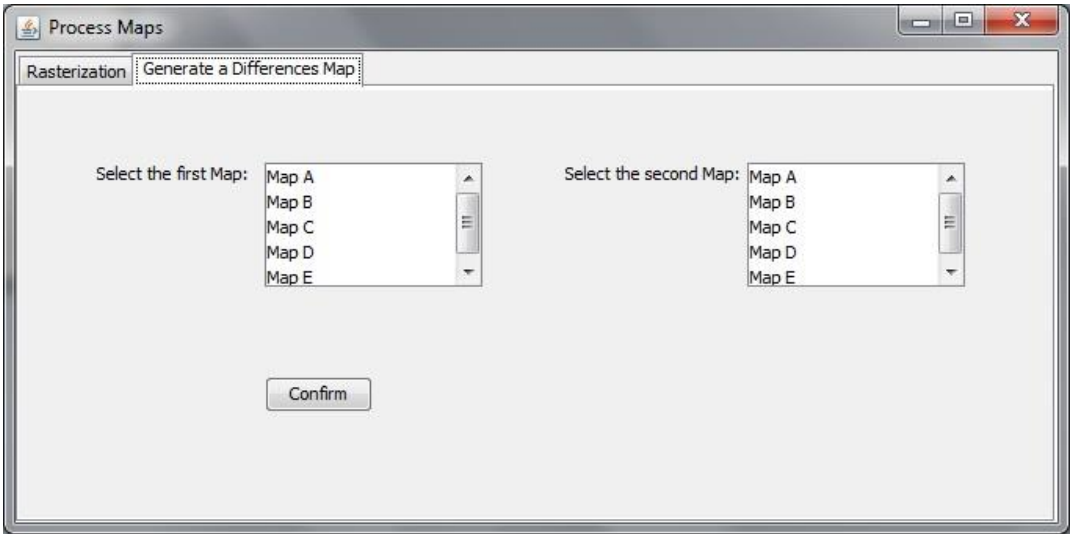


Figure (J) Processing maps frame, this pane allows to create a differences map

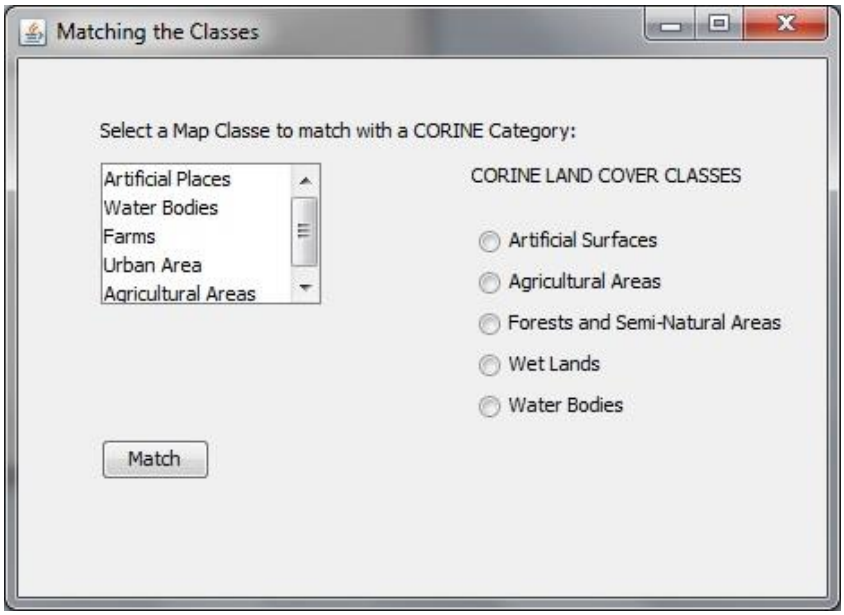


Figure (K) Matching classes frame, it allows to standardize land cover classes matching them to CORINE land cover categoriess.

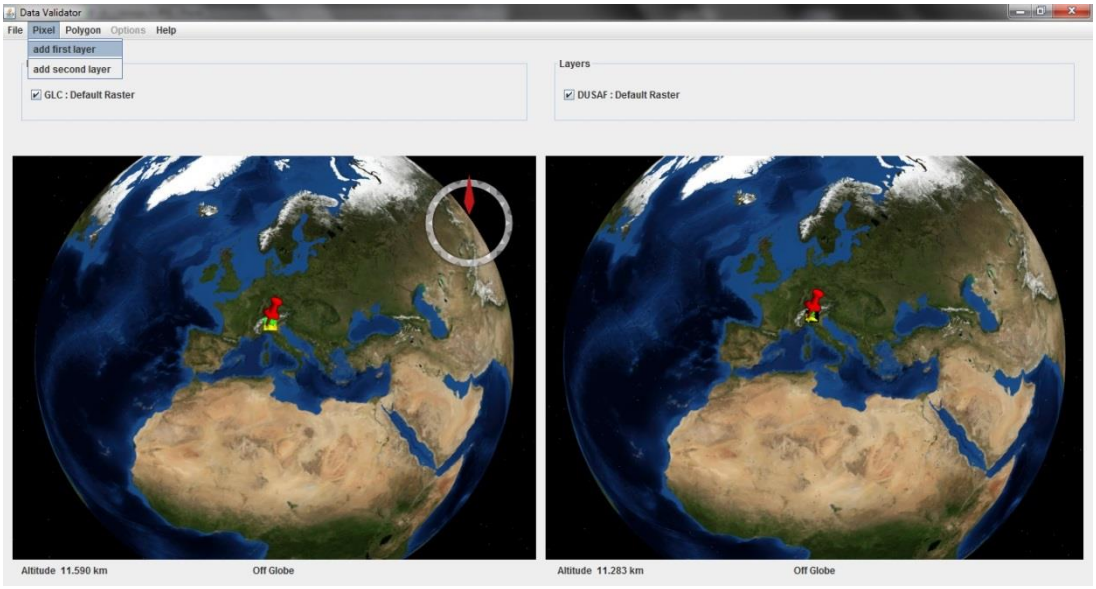


Figure (L) add land cover layers on the globes)

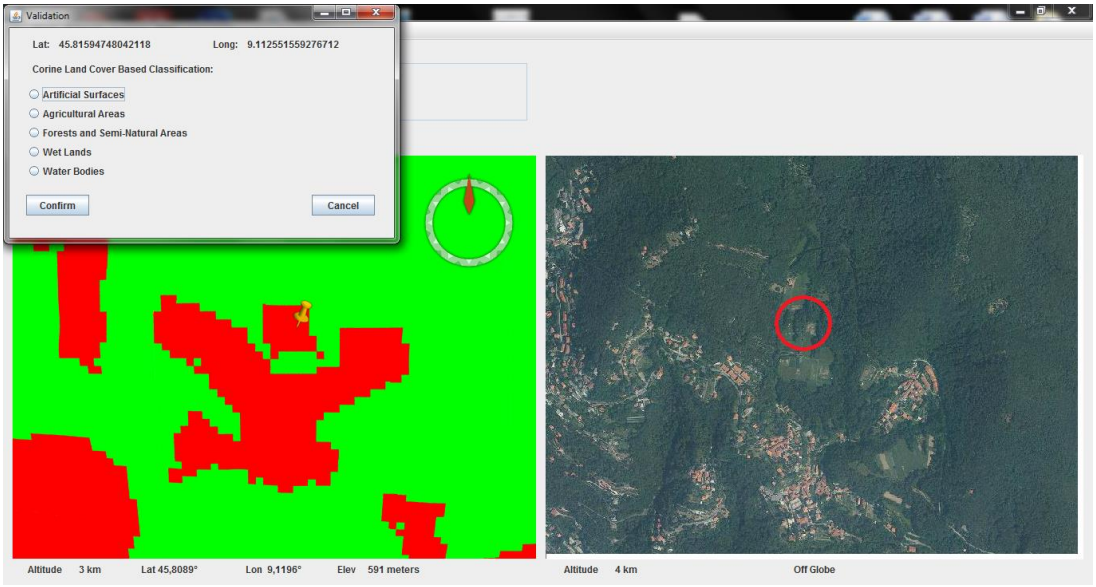


Figure (M) Validation Maps Window (left, the land is classified as Artificial Surface "red", on the right, the red circle shows the same area as seen from satellites)