Regionales Innovationszentrum für nachhaltiges Wirtschaften und Umwelt-/Geoinformation

Band 2

Hardy Pundt, Lars Bernard (eds.)

Proceedings
1st AGILE PhD School

March 13-14, 2012
Wernigerode, Germany

Shaker Verlag
Aachen 2012
Preface

The Association of Geographic Information Laboratories in Europe (AGILE, http://www.agile-online.org/) seeks to ensure that the views of the geographic information teaching and research community are fully represented in the discussions that take place on future European research agendas. AGILE also provides a permanent scientific forum where geographic information researchers can meet and exchange ideas and experiences at the European level.

The activities of AGILE are managed by an eight person council elected by the member organizations. Its main tasks are to develop an organizational structure to release the goals of AGILE, to further develop with the help of the members a European research agenda, to instigate and stimulate AGILE initiatives, to network with other organizations working in the field of Geoinformation and concerning technologies and to organize the yearly AGILE conference on Geographic Information Science.

During one of the last meetings, the Council discussed to strengthen the focus on young researchers who are doing challenging scientific work within the more than 80 member laboratories. Therefore, the Council decided to implement a new initiative that is focused specifically on PhD candidates. This decision was the advent of the AGILE PhD School. The first of its kind has now been carried out at the University of Applied Sciences in Wernigerode, Germany. PhD students and senior scientists met in Wernigerode to present their subjects, exchange ideas and to discuss critically experiences, results and challenges.

Within this booklet, the selected papers of the PHD students are included, all of them presenting work in progress in different stages of PhD projects covering various fields of Geoinformation Science.

The AGILE Council would like to thank the participants who issued in a first step position papers, and afterwards long papers, and took part in the PhD school to give valuable input. The PhD school is financed by AGILE; furthermore the participants get a grant from AGILE which should help to alleviate their costs.

Special thanks go to Martin Raubal (University of Zurich, Switzerland) who supported the selection of papers, and Bénédicte Boucher (Institut Géographique National, France) who confronted the PhD students with the theoretical and practical requirements of a National Mapping Agency.

Have new insights and make new experiences while reading the challenging papers included in these proceedings.

Lars Bernard, Hardy Pundt
# Table of Contents

New hypotheses on serial offender’s spatial behaviour *Marie Trotta* .................................................. 1

Integration of temporal and semantic components into the Geographic Information *Willington Siabato, Bruno Martins* .................................................................................................................. 10

Information Overload Problem in Decision Making *Ricardo Filipe Silva* ........................................... 29

Semantic description of generalisation web services for on-demand mapping *Nicholas Gould* ........... 38

Geospatial Ontology Harmonization *Heshan Du* .................................................................................. 48

Development of a Framework for Uncertainty-Aware Land Cover Change Analyses with Visual Analytics Methods *Christoph Kinkeldey* .................................................................................................. 56

Modeling and Validation of Spatial Uncertainty as a Context to Provide Ambient Services *Anahid Bassiri* ........................................................................................................................................... 65

Towards Understanding The Factors of Integrating Spatial Data Infrastructures and Crowdsourced Geographic Data *Mark Iliffe* ................................................................................................................. 71

Spatial Scale in Digital Soil Mapping *Stefano Cavazzi* .............................................................. 77

Ad-hoc-Geoprocessing in Spatial Data Infrastructures – Formalizations for Geooperators *Johannes Brauner* .................................................................................................................................................. 89

Towards a CityGML based data model for the GRB, the 2D-large scale topographic inventory of Flanders *Ine de Cubber, Jos van Orshoven* ................................................................................................................. 99

Modeling and analysis of migration movement of pre-colonial cultures in the Amazon basin *Carolin von Groote-Bidlingmaier, Sabine Timpf* ........................................................................................................ 110

Crowdsourcing Building Interior Models *Julian Rosser* .................................................................... 118

Taking VGI Seriously: The Portuguese Administrative Modernization Agency (AMA) Case Study *Antonio José Silva, Jorge Gustavo Rocha* ............................................................................................................. 126

Context-based geoinformation technologies in e-Tourism *Thomas Spangenberg* ......................... 134
New hypotheses on serial offender’s spatial behaviour

Marie Trotta
University of Liège, Belgium
E-Mail: marie.trotta@ulg.ac.be

Abstract

Geographic profiling is a methodology used to delineate a prior search area for a serial offender. Two new hypotheses integrating recent researches on offender’s behaviour are proposed to broaden the applicability of the techniques of geographic profiling to commuters and irregular patterns of crimes. The journeys-to-crime variance and place attractiveness will be successively integrated in GP methods. In order to evaluate the variance of the journeys-to-crime on real road network, a method based on least square adjustment in raster mode is developed. A real case of serial rapes illustrates the effectiveness of those assumptions for a non-uniform pattern of crimes.

Keywords: Geographic Profiling, Crime attractors and generators, distance decay heuristic, Raster analysis

1 Geographic profiling: its origin, methods and limits

In the 1980’s, environmental criminologists stressed the importance of the spatial dimension in crime analysis, and encouraged spatial analysis to become an investigative tool to narrow the offender search area (Brantingham and Brantingham, 1981, LeBeau, 1987). Theories such as routine activity (Cohen and Felson, 1979) and rational choice (Cornish and Clarke, 1986) were used to describe the criminal’s spatial behaviour, which is supposed to be directly influenced by the configuration of its immediate environment. According to the theory of routine activities, an offender is taking advantage of a criminal opportunity during his usual activities such as going to work, visiting relatives, etc. The rational choice postulates that an offender will try to minimise the effort and the risks to commit an offence while maximising the profit.

Subsequently, Rossmo transposed Brantingham’s theories into practice with the creation of the geographic profiling (GP). GP was firstly defined as a methodology of investigation that uses the locations of a series of connected crimes to determine the criminal’s most probable area of residence (Rossmo, 2000). It is based on the heuristic that offenders do not travel far to commit their crimes so that the likelihood to commit an offence deceases with the distance to his residence; this is called the “distance decay effect”. This heuristic is supported by both the rational choice and routine activity theories. It requires several crime locations to work effectively, and, therefore, it focuses primarily on crime series. This method is particularly effective at narrowing large pool of suspects (van Koppen et al., 2011) or at constraining an area for DNA investigation. Therefore, GP helps reducing the resources and time spent in investigations.
However, Rossmo’s definition of GP is too restrictive with respect to the evolution of the discipline. Indeed, the methodology can highlight any kind of anchor point of the serial offender such as work place, a relative’s residence, etc. Secondly, several locations can be connected to a single offence: the initial contact scene, the offence and disposal/release sites, etc., rendering GP applicable to that kind of single cases (Knabe-Nicol and Alison, 2011). Within this new canvas, GP can be defined as “a specialized subset of behavioural analysis directed at examining the geographical and temporal decision-making of an offender, in order to provide an investigative advice” (Knabe-Nicol and Alison, 2011 p127). This definition emphasizes the relationship between the time and space aspects in any decision process. As the investigator tries to understand this process, GP requires making hypotheses on the offender’s behaviour. This hypothesis is often the constraint of proximity between the offence locations and the offender’s anchor point. Besides, this definition broadens the scope of GP to any useful information that could help the investigators.

In practice, the ‘spatial distribution’ and ‘spatial interaction’ models are the most commonly used to determine a prior offender’s search area. While the former model focuses on the geographic arrangement of crimes (central tendency and dispersion), the later analyses the offender's geographic behaviour through its mobility characteristics (Brantingham and Brantingham, 1981, Canter and Larkin, 1993, Rossmo, 2000, Kent and Leitner, 2009). GP-dedicated softwares provide basic spatial statistics such as the centroid, the centre of minimum distances, and the harmonic mean or the deviational ellipse to study the spatial distribution. Amongst spatial interaction models, the journey-to-crime (JTC) estimation techniques are mainly used to estimate the offender’s anchor point. The form of the distance decay function does not seem to affect GP effectiveness. However, the function needs to be calibrated with solved cases on the study area.

GP techniques require specific conditions: the offence locations should be uniformly distributed around the anchor point and the offender should not travel too far to commit his crimes. This uniformity hypothesis is usually met in modern cities and grid-like road networks. This partially explains why GP has been successfully applied in Canada and USA but is still conspicuously lacking in European countries, excepting the UK, which are characterized by a small and highly populated territory with a complex road network. Indeed, researches in the Netherlands (van Koppen et al., 2011) and in Belgium. (Trotta, 2011) have shown that the uniform hypothesis is often not met. More concerns about the influence of the geographical features on the journeys-to crime are therefore needed to implement GP techniques in European cities.

New Bayesian approaches (Levine and Lee, 2009, Mohler and Martin, 2011) allow to take into account some of those geographical influences. Unfortunately, these methods mainly rely on Euclidean distances; therefore, they do not consider that the offender’s journey can be constrained by some geographic barriers such as the road network and landuse.

Moreover, current GP methodologies are based on the heuristic of the distance decay function. As mentioned above, it requires an offender acting from his anchor point and committing his crimes uniformly around it. This behaviour is what defines a ‘marauder’ (Canter and Larkin, 1993). On the other side, less commonly, the behaviour of the ‘commuter’ is characterized by the non-respect of the uniformity and proximity conditions. There are, however, other heuristics that can be tested in order to explain those behaviours (Bennell et al., 2009).
Herewith, I propose two new hypotheses integrating recent researches on offender’s behaviour to broaden the applicability of GP techniques to commuters and irregular patterns of crimes. JTC variance and place attractiveness will be successively integrated in GP methods.

2 New hypotheses for serial offender’s spatial behaviour

The first new heuristic is that a serial offender, by his repetitive behaviour, tends to minimize the variance in the distances travelled from his anchor point to the crime sites. This heuristic can be seen as a limit case of the distance decay effect where the buffer area is quite large in comparison with the distance an offender is willing to travel. As a result, the offender always travels just behind the ‘too-risky area’ and minimizes the variance of his journeys-to-crime. This heuristic presents several advantages. First, it is only based on the offender’s own spatial behaviour. There is no distance decay that needs to be calibrated with already solved cases on the same study area. This is in line with recent observations that variance is much higher for inter-offenders JTC than for intra-offender ones (Lundrigan et al., 2010, Townsley and Sidebottom, 2010). Besides, this hypothesis could be specifically applied to commuter’s behaviour where the variance between the journeys is small in comparison to the travelled distances (Mohler G and Martin, 2011). This heuristic should not be confused with the centre of minimum distance, which focuses on the place where the mean of the JTCs is the smallest.

Another concept needs to be integrated in GP methodologies: place attractiveness. (Brantingham and Brantingham, 2008) distinguished three kinds of places: crime attractor places, crime generator places and neutral places. Both the crime attractors (places known for their numerous criminal opportunities) and generators (locations with high concentration of both potential offenders and targets) are related with a higher attractiveness. This means that offenders travel longer distances to reach those locations than they do for neutral places. However, the measure of the absolute place attractiveness can be very difficult to estimate. At a small scale, urban hierarchy techniques can help to determine the places concentrating the highest levels of services that potentially attract people from ‘remote’ places. But at a larger scale, the influencing factors are numerous. The absolute attractiveness can also be approached by the Bayesian techniques with the origin-destination matrix. However, it is easier to estimate the relative attractiveness of places. This measure really makes sense in the context of a crime series as each offence location can be compared to the others.

According to Brantingham (2008), a distance decay effect can only be observed for neutral place. However, this affirmation needs to be nuanced: the slope of the distance decay actually depends on the place attractiveness; the more attractive is a place, the less steep is the slope (Trotta, in prep).

It is then possible to combine the assumptions concerning the distance variance and the place attractiveness. The places presenting the same profile in terms of attractiveness should reflect the same offender’s decision process. The offender’s anchor point should then be located at the place minimising the variance of offence locations with the same attractiveness profile.

The following sections of this article will successively present how to test the variance heuristic on a real environment when road network is used to estimate the travel distances.
Then, a real serial rape series illustrates how place attractiveness can be combined with this new hypothesis.

3 A least square adjustment as a method

While integrating place attractiveness only requires a variation of the slope for the distance decay function, the first heuristic, minimizing the variance of the JTCs needs a new methodology. The problem can be mathematically defined as follow:

The observations correspond to the position of the n crime locations \( x_i \) presenting a similar attractiveness. The objective is to compute the value of the 2 unknowns: the constant travel distance on road network from the anchor point to each crime location \( d \) and the origin of all these journeys \( x_o \).

The location of the anchor point should then ideally solve the following equation system: for \( i=1 \) to \( n \),

\[
x_i = x_o + d
\]

However, if an offender is presenting a constant behaviour, small variations in human behaviours always exist. A \( \nu_i \) term is then introduced in the above equation: for \( i=1 \) to \( n \),

\[
x_i + \nu_i = x_o + d
\]

The local solution in pixel \( j \) is then given by the mean of the \( n \) \( d_{ij} \). The global solution is finally given, following the least square adjustment, by the local solution \( j \) that minimises the following equation.

\[
\sum_{i=1}^{n} (\bar{d} - d_{ij})^2
\]

The raster process has the advantage of enabling the calculation on any point of the territory (more exactly on any cell). By contrast, the vector approach is constrained by the network nodes. The raster process can be described as follow. Firstly, a cost surface is created with the mask of the road network. A crossing cost of 1 is assigned to every pixel of the network while an infinite cost is given to pixels outside it. Any GIS software working with raster process (ex: GRASS, ARCGIS) is able to generate the cost distance from one specific location (in this case, a crime location) to any cell of the cost surface (the road network) (ex: the function “CostDistance” in Arcgis). See (Trotta et al., 2011) for the description of the algorithm. The raster calculator computes then for each pixel the mean of the \( d_{ij} \) and the local values of equation 3 corresponding to the sum of the squares of the residuals (SSR).
A major question of this methodology is to provide an upper bound above which the hypothesis of constant travel distance cannot be accepted. The variance on the travelled distances \( S^2 \) is given by the ratio \( \text{SSR}/n \). As the residuals of the least square adjustment are assumed to follow a normal distribution, the a-priori variance corresponding to the tolerated variation existing in all human behaviours restricted the value for \( S^2 \) (Dixon and Frank, 1983).

\[
\frac{S^2}{\chi^2_{1-0.5\alpha}/df} < \sigma^2
\]  

(4)

where \( \sigma^2 \) is the a-priori variance, \( S^2 \) is the variance computed with the ratio \( \text{SSR}/n \) and the degree of freedom for the \( \chi^2 \) Test is given by \( n-1 \). It is then possible to build the following test:

\[
S^2 < (\sigma^2 \times \chi^2_{1-0.5\alpha}/df)
\]

(5)

which means that, for a chosen variance \( (\sigma^2) \), the hypothesis of constant travelled distance is validated for all the pixels with a value inferior to \( n^*\sigma^2 \times \chi^2_{1-1/2\alpha}/df \).

4 Implementation on a real crime series

This section illustrates how this new method can be implemented on a non-uniform pattern of unsolved crimes attributed to the same unknown offender.

I use data concerning 18 rapes committed by a single offender on a short period of time (between April 2004 and May 2008) obtained via the Federal Police of Belgium. The objective was to delineate a priority area around crime locations for DNA testing as the offender was assumed to be local. The series presents a pattern focusing on two different city centres and the uniform distribution principle required to apply previous methodologies is not respected. Besides, as Belgian police does not have a long tradition in crime mapping, there is no information concerning already solved cases in the area available. The figure 1 presents the distribution of the crime sites with each number corresponding to its position in the chronology of events. All the crimes were located near a major road or near one of the two cities centres, not far from pubs or nightclubs. The first event (in chronological order) presents the particularity to be located in a small village where no attractive location could be identified.

Based on this data, two different profiles of place attractiveness are identified: the attractive city centres and the isolated rural area. The hypothesis of constant travelled time is then tested for all the crimes except the first one. One difficulty is to estimate the a-priori variance chosen to determine the upper-bound under which the hypothesis could be accepted. As the distance between the two sub-patterns is about 10 km, we consider that 10% of this distance can approximated the offender’s variability, corresponding to a variance of \( (1000)^2 \), with a resolution in metres. The figure 2 illustrates the area delineated by the computation of \( \text{SSR}/n < \sigma^2 \times \chi^2_{1-1/2\alpha}/df \) for with \( \alpha = 0.05 \).
As the first event is isolated and neutral in term of attractiveness, a linear distance decay function is then applied from this unique crime location. In order to combine both results to create a prioritized search area, each pixel under the upper-bound of the first treatment is multiplied by its distance to the first crime location. (Figure 3)

The actual residence of the offender was finally located very close from the first crime location and inside the area delineated by the minimization of variance. Those results favour the existence of another spatial hypothesis underlying the decision process of serial’s offenders (minimisation of the JTCs variance) and the necessity to nuance the classical assumption of distance decay with regard to place attractiveness.

Figure 1 The pattern of crime locations is far to be uniformly distributed
Figure 2 The $\chi^2$ test, considering an a-priori variance of $(1000)^2$ delineates a search area between the two sub patterns of offence locations.

5 Conclusion and perspectives

This paper discusses the opportunity to develop new hypotheses to describe the serial offender’s journeys-to-crime when the crime pattern is not uniform. Two hypotheses: constant travel distance and a distance-decay effect varying with place attractiveness are proposed to describe the decision process underlying such pattern. An original approach in raster mode, based on least square adjustment, is then proposed to test the hypothesis of constant travelled distance. An unsolved case of serial rapist provided by the Belgian police illustrates how the combination of both assumptions results in a prioritized search area for DNA testing.
Figure 3 The distance decay function from the first event prioritizes the results provided by the hypothesis of constant travelled time, the solution of the LSA

Several improvements could be made on this methodology. Firstly, the relative attractiveness of each crime location could be estimated with a matrix of origin-destinations of solved cases. Secondly, if a lot of researches have focused on the average distance travelled in journey-to-crime, literature concerning the variance is very limited. A better understanding of the variability of human’s choices under the same decision process would help to choose the best a priori variance. Finally, human decision processes are also influenced by their perception of space. Time distances could then better describe the offender’s journey-to-crimes.

References


TOWNSLEY, M. & SIDEBOTTOM, A. 2010. All Offenders are equal, but some are more equal than others: variation in journeys to crime between offenders* _Criminology_, 48, 897-917.


Integration of temporal and semantic components into the Geographic Information: Methodology

Willington Siabato
Technical University of Madrid, Spain
E-Mail: w.siabato@upm.es

Bruno Martins
Instituto Superior Técnico, Porto Salvo, Portugal
E-Mail: b.martins@ist.utl.pt

Abstract
The overall objective of this research project is to enrich geographic data with temporal and semantic components in order to significantly improve spatio-temporal analysis of geographic phenomena. To achieve this goal, we intend to establish and incorporate three new layers (structures) into the core of the Geographic Information by using mark-up languages as well as defining a set of methods and tools for enriching the system to make it able to retrieve and exploit such layers (semantic-temporal, geosemantic, and incremental spatio-temporal). Besides these layers, we also propose a set of models (temporal and spatial) and two semantic engines that make the most of the enriched geographic data. The roots of the project and its definition have been previously presented in Siabato & Manso-Callejo 2011. In this new position paper, we extend such work by delineating clearly the methodology and the foundations on which we will base to define the main components of this research: the spatial model, the temporal model, the semantic layers, and the semantic engines. By putting together the former paper and this new work we try to present a comprehensive description of the whole process, from pinpointing the basic problem to describing and assessing the solution. In this new article we just mention the methods and the background to describe how we intend to define the components and integrate them into the GI.

Keywords: T-GIS; Temporal GIS; spatio-temporal analysis; Information Retrieval; GIR; TIR; geosemantics.

1 Introduction
In Siabato & Manso-Callejo 2011, we presented the formal definition of the research project entitled Integration of temporal and semantic components into the Geographic Information through Mark-up Languages. Outlining this proposal, the project aims to define a set of methods, rules, and restrictions for the adequate integration of semantic, temporal, and spatiotemporal components into the primary elements of the Geographic Information (theme, location, and time) to improve spatio-temporal analysis of geographic phenomena. Although these primary elements were formally defined more than three decades ago (Sinton 1978), they are still used in different systems as they were originally defined, namely single vector formats, database storage methods, spatial database engines, among others.
Although these components certainly work quite well in different scenarios and solve an important amount of spatial problems, they alone seem to be insufficient for achieving tough spatio-temporal analyses. We consider that the lack of the semantic and temporal elements in the current storage structures of Geographic Information (GI) is the main reason for the spatio-temporal analyses to be deficient. Due to this, we state that the proposal of a new storage model for incorporating an independent temporal structure and a set of semantic components (temporal and spatial) would optimise such storage and therefore would improve the retrieving, processing, and analysis capabilities of GI into spatio-temporal scenarios.

Into this context, our proposal defines work oriented to the modelling, storage, and retrieval of dynamic GI taking into account the three above mentioned components. In order to integrate such components into the GI, we intend to establish and incorporate three new layers (structures) into the core of data storage process by using mark-up languages as well as defining a set of methods and tools to exploit the new layers. The ultimate objective is the modelling, querying, and retrieval of dynamic geographic features, establishing the necessary mechanisms to store incremental geometries enriched with a temporal structure and a set of semantic descriptors detailing (i) the nature of the represented phenomena, (ii) their temporality, and (iii) their temporal and meaning (semantic) relations. Figure 1 and Figure 2 show and overview of these concepts.

This research project is primarily based on concepts and studies related to space and time; Temporal GIS (T-GIS); semantic, geosemantics and semantic interoperability; annotation of temporal expressions; Geographic Information Retrieval (GIR) and Temporal Information Retrieval (TIR); Spatio-temporal Databases; and other cross-related topics. We do not include in this paper the state of the art since we have already prepared a completed survey of literature about space and time in GIS (Siabato & Manso-Callejo 2012b) and a preliminary study of the T-GIS foundations is also available in (Siabato & Manso-Callejo 2011). In addition to this, we have prepared and made available online an up-to-date dynamic bibliography about the above mentioned topics and others such as Moving Objects, Information Retrieval, Spatial Databases; even some standards quite related to this research have been listed. Moreover, we are also preparing an infographic study of this bibliography. All these elements and updated information about the evolution of this research project are available in (Siabato & Manso-Callejo 2012a).

---

1 The research plan includes the publication of six incremental papers, each one of which will describe a specific part of the project. This is the second of the series. Part III, IV, V, and VI will describe the models, the semantic layers, the semantic engines, and the integration of the proposed model.
Figure 1 Basic components of the GI and new proposed layers. The three layers relates the three primary components, the point of convergence is the time. A specific geosemantic layer describes nature of features.

Figure 2 Schematic overview of the proposal as a whole system. In this scheme, we represent the main parts of the proposal (models, layers, and engines) and the flow of the process.

The basic geographic data will be enriched through the three layers, which are in turn based on the temporal and spatial models. Once the enriched geographic data is ready, it is necessary to deploy a set of methods for exploiting them, part of these methods are the semantic engines. These engines can interact directly with both the data and the user to solve and identify temporal and geographic expressions. In this sense, user can also provide
any kind of textual entrance (e.g. documents, queries, tables) for processing it; this characteristic allows the handling of textual documents through GIS. By putting this together, user can interact with new data structures and make the most of the enriched data through new methods of analysis that will be based mainly on temporal interval algebra.

Thus, in this new article we present part of the foundations and the methodology that we will follow to define and to implement the components and methods that are expected to enrich and to improve geographic information. The following sections describe in detail the methodology for the design and implementation of three components. As we just describe the methodology, we do not go into detail about the design and characteristics of each component. In section 2, we present the Semantic Engine of Temporal Expressions (SEoTE) and the Spatio-Temporal semantic engine (STing). Then, we briefly introduce the temporal model on which this research will be based, in this section we introduce how the temporal behaviour of geographic features can be modelled through The Cube of Time and the Timing Points. We mention the spatial incremental storage model. Finally, in section 3, we present some preliminary conclusions and the next steps of this research.

2 Methodology and background

In this section, we present the foundations of the methods that are used for the definition of the semantic engines as well as the core concepts for their integration into our proposal. Although the formal definition of each component is out of the scope of this paper, we intend to show clearly the background and its justification.

Figure 3 shows how, since a theoretical viewpoint, different features of different nature will be able to interact directly with others and establish relations by themselves. These relations will enable data to identify features of same type and nature (circles) in addition to follow links or relations (intersections) that can be established e.g. by some heuristics or simple rules. This could make the system more flexible since it is not dependent on relational rigid models. Given this, the main component are the semantic taggers, through of which data will know who they are, what they are, and their temporal references, in this sense, task of semantic annotation plays an important role in this research.
Figure 3. Schematic representation of the integration of data due to the semantic taggers.

With the semantic layers, data will be aware of their nature and will be able to create relations by themselves. So, our vision in a very utopian world is being able to see how data can interact in a system just by incorporating them into the system. We want to provide data with the capability of interact without having to be immersed in a static model.

2.1 Semantic components

Semantic is a very general term used in different fields and which can refer to different significances. It is also often too vague. In this project, we refer to the term semantics in its purest definition, i.e. the meaning of the things. In this sense, the semantic components look for defining what the geographic elements and the temporal elements stand for, since a natural language viewpoint. In fact, this approach is quite related to the linguistic semantics, which refers to the study of meaning used to understand human expression through language. The idea of the semantic components and the layers of the system (see Figure 2) is to translate that meaning sense from language to data. Due to this meaning, data could be able to interact as shown in Figure 3.

In this research project, the definition of the semantic components comprises in essence two single tasks: tagging and processing. The tagging task includes the annotation of the data with specific tags, either temporal or spatio-semantic, as well as grant data the possibility of identifying and recognizing semantic expressions onto a text e.g. through a corpora (Corporation 2004, Ferro et al. 2010, Group 2008, Group 2010, Pustejovsky et al. 2003) or by using semantic dictionaries or ontologies (Lutz & Klien 2006). The processing task includes the identification and disambiguation of the expressions. In the strict sense, first task is completed when data are created or modified and the second one when data are queried or analysed. In conclusion, we must firstly define the semantic layers to be able to use the semantic engines for searching and analyzing data collections or user inputs (see Figure 2).

For the tagging task we will use a set of mark-up languages (e.g. TIMEX2, TimeML, and SpatialML). We will also define some general heuristics for the resolution of specific cases;
these heuristics will also improve the accuracy in the identification and disambiguation of the expressions (F-measure).

For the definition of the processing task we will use the Natural Language Processing techniques, specifically the Named Entity Recognition (NER) (Nadeau & Sekine 2007). The methods on which we will base come from the Information Retrieval area; we include some specific techniques applied into the Geographic and Temporal Information Retrieval fields. Over the last few years the temporal reference resolution problem has been addressed by many different researchers (Ahn et al. 2007, Mani & Wilson 2000, Miller 2003) as well as the geographic references (Goldberg et al. 2009, Kornai & Sundheim 2003, Markowetz et al. 2005, Martins et al. 2010). The processing task will be conditioned by the semantic engines and their training subtasks which depend on annotated corpora, gazetteers, rules, and ontologies. In this project, we will define and create a specific ontology for temporal expressions. For the spatial entity recognition, there exist huge and very reliable geographic collections such as the GeoNames\(^2\) gazetteer among others, this research project will be supported on this comprehensive projects and collections. This task is initially just related to identification of geographic entities in the natural language queries defined by users, although it can be also extended to general use purposes. In fact, these engines will support the tagging of geographic data as well as permit the handling of textual documents through GIS. Keeping all these basic concepts in mind, in the next sections we describe the semantic engines.

2.1.1 Semantic Engine of Temporal Expressions. SEoTE

Temporal expressions by definition have multiple granularities (finer and coarser), models, representations and other different possibilities depending on context and source such as durations and reference points. These expressions are also pervasive since every document or conversation has a high probability of contain references to particular calendar dates, clock times or duration periods (Loureiro et al. 2011). This means that temporal reference identification presents different non-trivial problems due to the inherent ambiguity and contextual assumptions of the natural language discourse, e.g. the expression “at the beginning of the Cainozoic era” could means a thousand of years, while the expressions “at the beginning of the renaissance” could means 10 years. The definition of these intervals implies to know the period of the events and what the expression at the beginning of means.

For the resolution of temporal references, we will follow the method proposed by Loureiro et al. (2011). In their work, authors proposed a supervised machine learning approach. The method is an instance of the well-known stacked learning paradigm proposed by Wolpert (1992). Stacked generalization is a way of combining multiple models, i.e. introduces the concept of a meta-learner or second learner. The procedure comprises four general steps: (i) split the training set into two disjoint sets; (ii) train a significant amount of base learners on the first part; (iii) test the base learners on the second part; and (iv) using these predictions

\(^2\) [http://www.geonames.org/](http://www.geonames.org/)
as the inputs and the correct responses as the outputs, train again a higher level learner. In our case, the first learner, based on Conditional Random Fields (CRF), is used to recognize and classify temporal references, and the second learner, based on Support Vector Machine (SVM) regression, is used to rank the possible candidate disambiguation for the temporal references that were initially tagged. In addition to this, we introduce a novel characteristic in this model by using the Information Retrieval techniques based on ontologies. Although the accuracy of Loureiro’s method ($F_1 = 0.61$)\footnote{The $F_1$ rate is equal to the harmonic mean between precision and recall. Precision is the percentage of correct references identified/disambiguated by the system and recall is the percentage of references present in the test collection that are identified/disambiguated by the system.} can be considered good enough to be applied into our project as is, we expect to improve this metric and increase the identification of the temporal expressions. We will measure such expected improvement by using the same gold-standard collections (Corporation 2004, Ferro et al. 2010, Group 2008, Group 2010, Pustejovsky et al. 2003) and textual documents to establish accurately how this modification affects the cited method. This implies that we will need to train some datasets apart from create the temporal ontology.

2.1.1.1 SEoTE proposal

In order to define what a temporal expression is and how one should be disambiguated, we follow the TIMEX2 standard. In a previous work (Nieto et al. 2010), authors made an initial attempt for the integration of TimeML into GIS. The main idea was, as in our case, to enrich GIS with temporal capabilities. However, one of the conclusions in the ongoing work was that TimeML is a powerful Mark-up language with a large semantic that goes far beyond the identification of temporal expressions, with a finer granularity, and due to this it is too much just for the identification of the expressions. For this reason and based on this conclusion, we have decided to use TIMEX2 instead of the TimeML standard (-ISO- 2007). In fact, one of the foundations of TimeML is TIMEX, although this standard implements TIMEX3.

Here we provide a very brief description of the TIMEX standard. There are six attributes defined for TIMEX2, the values of these attributes express the semantics of a temporal expression. (i) VAL, that contains a normalized value for the date; (ii) MOD, that captures temporal modifiers using values close to Allen’s algebra expressions such as before, after, less than, more than, equal or less, start, mid, end or approx; (iii) ANCHOR VAL, that contains a normalized value for an anchoring date or time; (iv) ANCHOR DIR, that captures the relative direction or orientation between the VAL and ANCHOR VAL attributes, as in within, starting, ending, as of, before or after (It is used to express information about when a duration is placed); (v) SET, that identifies expressions denoting sets of times; and (vi) COMMENT, that contains any comment made by the annotator but it is ignored in the processing. The complete description of the TIMEX2 is available in (Ferro et al. 2005). As this document indicates, although one often refers to this standard as “the TIMEX2 standard”, the original name stands for the TIDES program in which was originally developed. It was first documented in Ferro et al. (2000, 2001). Although Loureiro et al. only take into account the VAL and SET attributes, ignoring the remaining TIMEX2 information, for the proposed
semantic engines we will consider the full set of annotations supported in the TIMEX standard since the identification of the relations would be particularly important.

In the methodology, the first step is the identification of the temporal expressions. In order to do this it is necessary to implement a supervised method. Nadeau and Sekine (2007) states that the currently dominant technique in NER is the supervised learning, mainly with models such as the Hidden Markov Model (HMM) or Conditional Random Fields (CRF). Supervised learning requires a pre-annotated set of documents for training a classifier capable of identifying entities in text, for doing this training data must be tagged manually. For the SEoTE and the STing engines, the first learner will be based on the CRF method.

![Figure 4](image-url) Schematic representation of the semantic temporal engine. The raw data could be any kind of text, e.g. input user text, or attributes of the database. The annotated could be stored in the semantic layers or be used for querying the data. The evaluation method 1 generates the initial temporal references, while method 2 is in charge of identify problems and solve it (disambiguation process).

Lafferty et al. (2001) define CRF as a framework for the construction of probabilistic models to segment and label data given in sequence. Sutton & McCallum (2012) asserts that CRF are essentially a way of combining the advantages of discriminative classification (meaning that it is based on a model of the conditional distribution p(y|o)) and graphical modeling, combining the ability to compactly model multivariate sequential data with the ability to leverage on a large number of input features for prediction. In a sequence tagging problem, such as NER, CRF carries out logistic regression over the possible sequences of tags. The conditional distribution P(\lambda|\theta) is modelled as shown in Equation 1. In other words, CRF is basically a conditional probability associated with an undirected graphical model.

\[
P(\lambda|\theta) = \frac{1}{Z(\theta)} \exp \left( \sum_{i=1}^{n} \sum_{t=1}^{m} \lambda_i f_i(y_i, 1, y_0, \theta_i) \right)
\] (1)

---

4 There are also semi-supervised approaches, where only a small degree of supervision is needed to start the training process, as well as unsupervised learning approaches, which use techniques that rely on lexical resources, patterns and statistics to recognize the entities. (Nadeau & Sekine 2007)
The parameter $Z\lambda$ is a constant for normalization, and it is defined as shown in equation 2.

$$Z\lambda(\omega) = \sum_{\gamma \in \Gamma} \exp \left( \sum_{n=1}^{N} \sum_{i=1}^{p^n} \lambda_{i} f_i(y_i - 1, y_i, \omega_{i}) \right)$$

(2)

With this model, the identification of temporal references starts with the procedure indicated in Figure 4. By using the BIO (Begin, Inside, and Other) encoding for tagging, we will identify seven types of references (classes) as in (Loureiro et al. 2011): recurrences, durations, points, ambiguous past references, ambiguous present references, ambiguous future references, miscellaneous. This implies to separate the BIO tags for each of the seven references. The classification model tags words given in a sequence according to the above tags, from which the system then generates the final results for the temporal reference recognition step.

The recognition process starts with a tokenization of the sequence to be analyzed. We use a scheme that deterministically breaks an input text into a sequence of independent tokens. This collection of tokens corresponds to the (o) observation sequence. What we are looking for, it is probability of classifying each symbol (o) (tokens) with a class $y$ among the set of classes we have defined (Y). Tagging these tokens is a non-trivial classification problem, since if there are 7 different tags, then a sequence of length $N$ has up to $7^N$ possible sequences. To get through this problem, we assume the CRF probabilistic approach.

Once identified the temporal expression, next step is disambiguation. For doing this, the methodology includes three steps: (i) the proposal of a set of disambiguation candidates by using a generative approach based on rules and the ontology of temporal terms, (ii) scoring possible candidates using a Scalable Vector Machine (SVM)$^5$ regression model, and (iii) selecting the best candidate. The SVM regression is based on the overlap between the true temporal period associated with the expression and the temporal periods of the candidates. This way, the selected candidate will be closest to the real period. For the definition of the candidates, the identification is based on (i) pattern comparison and (ii) ontology comparison. Once the candidates have been identified the highest probability will be function of the both comparisons. Patterns define rules that may refer to a small lexicon of names, units, and numeric words, these patterns could be use both in the text and the ontology. It is necessary also to define a tagger element for the VAL and SET TIMEX2 elements. In general terms, the process involves evaluating the rule’s pattern against the text of the temporal expression and, in case of a successful match, executing the tagger element to identify the candidate already disambiguated. As usual in TIR, deictic and anaphoric terms imply a higher level of complexity and additional treatments must be included as identified in (Nieto et al. 2010). In order to compare expressions properly it is necessary a temporal reference system, we will use the Gregorian calendar and ISO 8626 standard (this will be presented in Temporal component, see Section 2.2 below).

---

$^5$ Support Vector Machine is a set of machine learning approaches used for classification and regression. This method was developed in the mid 90’s by Vapnik (2000) at the AT&T Bell Labs.
A temporal reference will be correctly identified only if it exist an exact match with the corresponding temporal reference given (o) in the test collections (corpus and ontologies), and correctly disambiguated only if the assigned disambiguation is an exact match with the one that is given in the TIMEX2 annotation (i.e., the VAL and SET attributes have the same values). Thus, correct disambiguation can only occur when a correct recognition has first taken place. The final step of the method involves selecting the candidate with the largest estimated overlap as the result for the disambiguation. Loureiro et al. (2011) asserted that the achieved results of the stacked proposed method are of an acceptable quality to be used in the subsequent processing stages of many different types of Temporal Information Retrieval applications, in our case, the construction of SEoTE and its integration into a GIS. We expect to improve the accuracy in our semantic engine by adding the ontology-based recognition layer and enriching the rules, pattern and heuristics. Even, it could be valid to modify the standard annotated corpus to increase the recognition of valid expressions that have been wrong annotated in the gold-standard collections.

2.1.2 Spatio-temporal semantic engine. STing

By using the same techniques is possible to identify geographic references, and places with a temporal scope. As in the temporal expression recognition process, to identify geographical references it is not a trivial task and it could be even considered a harder challenge, e.g. some lat-long coordinates refer to a single location but it can be recognized as different geographic references across the centuries: Byzantium (667 BC), Constantinople (330 AD), Istanbul (1930). One city or another will be right depending on the temporal context of the data. As this, there are multiple scenarios that require strong support and comprehensive knowledge-repositories such as gazetteers. The ambiguity of geographic references is high, e.g. place names often have other non geographic meanings, different places can be referred to by the same name, and the same places are often referred to by different names (Martins et al. 2010). In general, the concepts we will use for the geographic entity recognition come from the Geographic Information Retrieval field. The approach followed for the definition of the spatio-temporal semantic engine is also identification/disambiguation. The main aim of the STing engine is to identify geographic references included in user query expressions. Nonetheless, the engine could be classified as a general purpose semantic application.

System accuracy for the identification of geographic references is higher than in the TIR systems. F1 scores around the 90% were already reported by Tjong-Kim-Sang and Meulder (2003). In the last years techniques have evolved in process speed although accuracy has not improved too much. Although Initial approaches for the identification of entities (NER) were based on manually constructed patterns and/or dictionary lists of entity names, the current trend in NER is to use machine-learning approaches, relying mainly on features extracted from training data. Machine learning approaches are more attractive in that they are trainable and adaptable (Martins et al. 2010). Different techniques such as Support Vector Machines, Maximum Entropy, Hidden Markov Models (HMM), and the already mentioned
CRF, mainly support GIR. The expected geographic semantic engine will be based on a stacked method with two learners CRF and SVM, just as in the temporal engine (see Figure 4).

While the identification of the entity can be made to rely entirely on internal features of the documents, place reference disambiguation requires always external knowledge in the form of a dictionary (e.g. a gazetteer) for translating place names into geospatial footprints (Siabato et al. 2008). In our case, we will use the GeoNames gazetteer service (Wick 2011), which includes over 8 million geographic references. An important characteristic of this service is that entities evolve through time, as we exemplified above. In an early stage of this research, we were thinking to use a personal gazetteer in which we included GeoNames and other gazetteers services (Manguinhas et al. 2009). Nonetheless, this proposal is out-of-date and GeoNames service has grown and evolved significantly making it by far a better option for the disambiguation of geographic references. In his doctoral dissertation, Leidner (2007) surveyed different approaches for identifying place references on textual documents. He concluded that most methods usually rely on gazetteer matching for performing the identification, together with a set of heuristics.

Some remarkable examples of GIR systems are the MetaCarta RSS GeoTagger⁶ and the Yahoo! Placemaker⁷. Martins et al. (2010) presented a comparison of these commercial services against a proposed machine learning approach for the recognition and disambiguation of place references based on HMM. Experiments with labelled datasets in three different languages showed that the machine learned method outperforms the two commercial state-of-the-art systems.

### 2.1.2.1 STing Proposal

The approach that we will use in this engine is the same that we presented for the temporal one: stacked learning based on CRF and SVM. In this case, the (SVM) regression to rank a set of possible disambiguation candidates relies on the gazetteer service.

For recognition task, once again the first step is the tokenization of the text. The candidates (recognized tokens) must be tagged following general structures such as the BIO encoding or BMEWO+ encoding. For the geographic entities we will check which encoding system is more appropriate. The recognition problem must include a method for solving the non-geo/geo and geo/geo problems.

The disambiguation process involves (i) the identification of candidates by querying the gazetteer, (ii) the classification of possible candidates using the SVM regression’s scores, and (iii) the identification of the highest scoring candidate. Following Martins et al. (2010) methodology, the scoring criteria used in the second step are based on estimating the

---

⁸ LingPipe’s BMEWO+ encoding distinguishes the following set of tokens: begin-of-entity (B); mid-entity (M); end-of-entity (E); non-entity (O) and single-token entities (W). There are also sub-classifications and constraint rules.
distance between the true geospatial footprints and the geospatial footprints of the candidates by using algorithms such as the convex hull and the Levenshtein distance. The selected candidate will be the top scoring, which corresponds to the one with the least estimated distance in terms of geographic reference approximation. Due to try to identify geographic references by exact location is not possible at all, mainly for the precision of the different datasets, the disambiguation process will take into account a radio of tolerance, the value of this radio must be determined by comparing precision of the involved datasets.

The datasets for training data will be the ones defined in the Conference on Computational Natural Language Learning -CoNLL- (Tjong-Kim-Sang & Meulder 2003). Both editions of the CoNLL NER evaluation, the organizers provided a set of data files for each language (Spanish in the 2002 edition and English in 2003).

2.1.3 Implementation

These engines are expected to be implemented through the CRF and SVM algorithms defined in LingPipe\(^9\) and the Weka Machine Learning toolkit\(^10\). These libraries are robust and portable enough to implement the engines as though software pieces to be included into the GIS software.

2.2 Temporal component

The temporal component is the most important part of this research due to the whole developed concepts fall on it. In this section we introduced very briefly the temporal model that we intend to define and to establish.

2.2.1 Temporal model. The cube of Time.

Here we present some basic and initial elements about what we have called The Cube of Time. This concept differs from the space-time prism defined by Hägerstraand (1970) and Lenntorp (1976) and which is commonly used for spatial analysis (e.g.Miller 1991). Unlike the space-time prism which is composed by two spatial axes and one temporal (Z), the cube of time (C\(_t\)) comprises three temporal axes: database recording time, object changing time, and object creation time (see Figure 5). The point defined by the three temporal coordinates is called the Timing Point. A sequence of timing points describes the evolution of the object through time. This sequence, which can be described mathematically, shows the evolution of an object through time by following a linear function (we will define if the linear function is good enough or it will require a different type of function). A set of functions permit to establish relations between objects and to identify for instance how a single change in a specific object can affect another or others. The aim of this cube, and therefore of the model, is to provide mathematical definitions that describe the evolution of single objects through time and permit to analyze and identify how they are correlated.

\(^9\)http://alias-i.com/lingpipe/
\(^10\)http://www.cs.waikato.ac.nz/ml/weka/
Even though Miller (2005) has established some mathematical definitions for the space-time path and prism, space-time lifelines, bundles, intersections, and other components of the space-time prism framework, and the definitions we will propose for our model are quite linked to Miller’s proposal since both of them relate spatial objects in the background; the prism and the cube are quite different in concept mainly because our functions express purely temporal relations and not space-time relations.

In the late nineties and based on the concept of bitemporal element (BTE) (Snodgrass 1992), Worboys (1998) defined the spatio-bitemporal model for geographic information. In this model he considered event-time and database-time along two orthogonal axes. He defined the ST-simplex primitives (0-simplex, 1-simplex, 2-simplex) from which ST-complexes and ST-objects were derived. Every simplex can be combined with a bi-temporal element to form an ordered pair space-time, this pair defines geographic objects with both spatial and bi-temporal extents.

![Three temporal Axes of the Time Cube](image)

**Figure 5** Three temporal Axes of the Time Cube. The cube is described in a $\mathbb{R}^3$ Euclidean space ($E^3$). Each coordinate of the *Timing Point* is related to a single object’s state. The origin $\theta t$ can be positioned in any range of $R$. Although the functions that represent the evolution of the objects can be circumscribed in any octant, due to any database recording time would never be less than the first recorded value $\theta t(z)$, the functions will be always over the plane $Zt=0$, and therefore, into the four upper octants.

This is a significant difference between our proposal and Worboy's model due to $Ct$ is an abstract composite for establishing temporal relations and do not include neither spatial information nor elements. In fact, it is not possible to state comparisons between these approaches. Nonetheless, it is necessary to bear in mind the bi-temporal model because it contains two of the three axes we have defined in our cube. Although it seems quite logic to use only two axes and to omit the object creation time axis mainly because this coordinate could be considered as the first event of change, hence not to think the temporal composite as a cube but as two orthogonal axes just as Worboys (and Snodgrass) defined, we will show how by defining the three times (axes) independently and by treating the *Timing Points* in different and parallel planes will be possible to model temporal relations in a different way.
and to make temporal analysis easier, richer and more flexible. The main objective is to enrich temporal descriptions and temporal relations of the objects (features). On the other hand, this requires more complex definitions due to vectors that describe the behaviour of the object are defined in $\mathbb{R}^3$, but this added complexity level is just for the model definition and do not affect neither implementation nor performance.

As we stated in the definition of this research project (Siabato & Manso-Callejo 2011), how to perceive and model time is a mainstay. In this study, the mechanical and mathematical models will be considered due to through them all geographic features perceived by our senses (rivers, ways) or abstractly modelled (airways) may be ordered and described from a temporal viewpoint. Moreover, it is possible to define elements such as measure, interval, dimension, modification, instant, and position among others. Particularly the applied perspective will be the Newtonian, where time is envisioned as an independent dimension though widely related and similar to the spatial dimension. This approach is used in view of the flexibility it provides for independent handling of the temporal component allowing that space and time can be measured independently (relativist perspective (Ott & Swiaczny 2001)).

In general, (i) the three axes that define the cube are mechanical-temporal axes; (ii) the three planes defined by the axes are bi-temporal planes; (iii) the origin of the cube ($\mathbf{0}_t$) is determined by the first record in the system (e.g. Spatial database), what implies that the cube in the very first record will be a square (XZ bi-temporal plane). Finally, the time registered in the axes follows the Gregorian calendar for the identification of calendar days and the ISO representation of dates and times (-ISO- 2004). We will evaluate to adapt Coordinated Universal Time for the Timing Points values, this will permit evaluate and operate them easier.

2.3 Spatial component

An initial description of the model for the spatial component is available in (Siabato & Manso-Callejo 2011). The proposed incremental spatial storage model is also based on mathematical foundations that determine born tasks, evolutional tasks and dying tasks. These three set of tasks are also defined in the temporal model. In this sense, temporal evolution and spatial evolution will be modelled independently but they are quite related as in any spatio-temporal model. So, we will integrate the whole concepts in a metamodel, that’s the main reason we assumed the mechanical model described above. As presented in (Siabato & Manso-Callejo 2011, p.404), the spatial model is based on the +/-$\delta t$ operator. We do not extend this concept due to the fact that it has been described quite enough in such work.

---

11 A complete discussion of types of time in GIS was presented by Frank 1998.
12 The formal definition of the temporal model will be submitted for peer-reviewed to the ICCSA 2012 Conference. This model corresponds to third paper of this series.
2.3.1 Spatio-temporal analysis

Finally, to provide the system with analysis capabilities we rely on Allen’s interval algebra (Allen 1981, Allen 1983). For doing this, we will develop an xQuery library to exploit the model, the layers and the proposed storage data model. This library could be considering as third component over and above the semantic engines.

3 Preliminary conclusion

Although this paper just presents the formal methodology that we will follow up into our research, it is possible to state some conclusions. Bearing in mind that the overall goal of this research is to significantly enhance geographic data (i.e. making it more intelligent and flexible) in order to improve current spatio-temporal analysis methods, and to provide GIS with semantic tools that allow system interaction following the flow User $\rightarrow$ System $\leftarrow$ Data; in this paper we have briefly described the methodology and foundations on which we will relay to achieve such objective.

After analyze different proposals and literature reviews, it is clear that for the semantic engines (SEoTE and STing) the Machine Learning approach turns in the best option for the identification and disambiguation of temporal expressions and geographic references. We will follow the stacked method by taking into account two learners based on the Conditional Random Fields (CRF) and Support Vector Machine (SVM) regression. We decided to use these learners due to in the reported works this approach has achieved the highest quality score $F_1$. Furthermore, these techniques are well-known and there exist a set of algorithms and robust software libraries on which we will stand to create the engines. These libraries guaranty a tough system as well as its portability and scalability. Although the aim of our research is not to create new methods for identifying or disambiguating the expressions, we expect to improve the $F_1$ score by introducing a new ontology-based layer. About this, it is not possible to conclude something until implement the prototype and measure the resolution of references (temporal and geographic) using the same gold-standard document collections and conditions under were tested the proposals on view. We intend to create general purpose semantic engines, in this sense, their integration into the GIS will be just a single use case through which such engines will be validated in a geographic context and system.

Regarding this matter in the case of the STing engine, and due to once included in the GIS software the geographic context will be already defined in somehow, we will have to deal with this specific characteristic and, even more importantly, to use it for the disambiguation process. In fact, this should be considered as the main source for this process since the system supposes to have in most cases trust data, and these data will define free of error the geographic scope of the system in that scenario.

Although Louveriro et al. concluded that for addressing more advanced temporal processing a next step in their proposal could be the use of a more powerful mark-up language as
TimeML; based on our previous results we strongly believe that simple annotation is more practical and useful for the integration of GIR and TIR methods in GIS and other systems. Due to this, we will use simple TIMEX2 annotations. Probably, TIMEX3 would be considered if the spatial and temporal integration does not report the expected results.

The described temporal model plays a fundamental role in the research. We think that representing the dynamic behaviour of geographic data through parallel planes, on which each feature has “drawn” its evolution, will permit us to define a comprehensive set of operations to analyze temporal data in a completely novel way. Moreover, once defined the foundations of The Cube of Time, it will possible to create scenarios for n-dimensional analysis by creating new mathematical definitions for a Universal cube (U-Ct) in which the basic cubes are contained into a Universal set. This will permit to analyse how changes in one system impacts or changes the behaviour of another. Nonetheless, much more have to be done about this proposal to be considered as a general model, once again, just until implementation tasks have been completed the real impact and reach will be analyzed.

By putting together the former position paper and this new work, we have tried to describe the whole process, from pinpointing the basic problem to describing and assessing the solution. In the Timebiography web site (Siabato & Manso-Callejo 2012a), there are available over 150 references in the sections Semantic, IR and GIR and Annotating Time related to the development of the semantic engines. There are 50+ references related to Allen’s interval algebra and time intervals in the section AI and Logic. Ongoing work and next steps couldn’t be others than implementation of the engines and definition of the mathematical foundations of the models.

Acknowledgments

This work was partially supported by the Doctoral Program of the Technical University of Madrid -UPM- (Grant ref.CH/056/2008) and by the UPM Training and Mobility of Researchers Programme (Resolution 23/02/2011). The work is also supported through SInteliGIS, a research project at INESC-ID financed by the Fundação para a Ciência e Técnologia (grant PTDC/EIA-EIA/109840/2009). Willington Siabato is on leave from UPM and he is currently a visiting researcher at INESC-ID, a research lab associated to the Technical University of Lisbon.

References


Wick, M. (2011/09/12), GeoNames, [Online], Available from: <http://www.geonames.org/> [05/02/2012].


ref_end
Information Overload Problem in Decision Making:
The Important Relies on Changes

Ricardo Filipe Silva
Universidade Nova de Lisboa, CENTRA, Portugal
E-Mail: ricardofcsasilva@gmail.com

Abstract
Nowadays databases are storing huge volumes of data which are referenced in space and time, representing spatio-temporal dynamism with multivariate connections, containing implicit structures, relationships and interactions. When we are dealing with complex data in Business Intelligent (BI) systems, it is likely that the decision maker does not know what useful information (interesting patterns, relationships and correlations) can or should be obtained from the dataset available to him. This problem is mentioned in Geovisual Analytics (GA) area as the information overload problem (IOP). To address it, a conceptual analysis mode emerge “Analyze first, show the important, zoom, filter analyze further, details on demand”, which has underlying a model that combines the strengths of automatic data analysis and interactive visualization. This paper proposes an approach to address the IOP by providing the user with syntheses of changes that could be significant as a way to show the important. Finally, we will sketch a research path discussing how this proposal should be evaluated, its applicability as well its potential for further research.

Keywords: geovisual analytics, spatio-temporal, information overload problem

1 Introduction
Nowadays, Decision Support Systems (DSSs) in organizations are mostly Data-driven DSSs (D-DSSs) also known as Business Intelligent (BI) systems, which emphasizes the access and manipulation of the time-series data of internal organization, and sometimes external data, helping the decision maker to gain insights from historical data (Power 2008).

With the technological advances, organizations are producing and storing huge volumes of data (D. A. Keim, Mansmann, & Thomas, 2010) (Yang et al., 2011) which in most cases are referenced in space and time, representing spatio-temporal dynamism with multivariate connections, containing implicit structures, relationships and interactions. Such environment has a strong impact on the D-DSSs usability. Originally, those systems were designed to be used on discovery mode, i.e. “Give me what I say I want, then I can tell you what I really want.” (Inmon, 2005). However, our capability to process, analyze and exploit it was outpaced (Compieta, Martino, Bertolotto, Ferrucci, & Kechadi, 2007).

When we use the same analysis mode for complex data, it is very likely that the user get lost in data. The user does not know what useful information can or should be obtained from the dataset available to him, which in turn let him lost in the instruments, methods and
representations available. Furthermore, in general, the user has a trial-and-error behavior until he gets interesting information from his viewpoint. Plus, such interaction can be tedious for him (Giacometti, Marcel, Negre, & Soulet, 2011) and, after the analysis process, it is likely that useful, hidden information and implicit structures of data will remain to be discovered (Ham, 2010). Thus, the discovery mode is not the most appropriate for complex data.

This issue is mentioned in Visual Analytics/Geovisual Analytics (VA/GA) area (G. L. Andrienko et al., 2007) (G. L. Andrienko, Andrienko, Keim, MacEachren, & Wrobel, 2011) as the information overload problem (IOP): (i) the data is processed in an inappropriate way; (ii) the data is presented in an inappropriate way; (iii) even the data is processed and presented in proper manner could be irrelevant for the analyst task.

To solve the IOP, VA proposes a model (D. A. Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008) (D. Keim, Andrienko, et al., 2008) based on the cooperation between the human and automated data analysis. On one hand, it takes advantage from algorithms and vast computational power of modern computers in order to provide useful data syntheses. On the other hand, it integrates human background knowledge and intuition to overcome the automatic analysis limitations (e.g., the parameters of data mining algorithms and the corresponding complex process to tuning those parameters). Furthermore, it incorporates appropriate visualizations and interaction techniques, bringing the user to the analysis process where data can be interactively manipulated, instead of him being just left with the results.

The proposed model has an implicit analysis approach: “Analyze first, show the important, zoom, filter analyze further, details on demand” called the VA Mantra (D. Keim, Andrienko, et al., 2008). The purpose of the analyze first and show the important steps is to overcome the IOP by providing the user with a synthesis that gives him an overview and a concise description of a phenomenon. Such task relies on the ability of evaluate and detect what is important from the available spatio-temporal data to a given analyst and at given context so that a user could conduct his analysis based on relevant data. However, evaluate and detect what is important given a spatio-temporal dataset is a truly challenging task.

In this paper, we propose to address the information overload problem by providing the user with syntheses of changes that might be significant as a way to show the important.

This paper is organized as follows. Section 2 presents the current issues associated to the information overload problem. Section 3 details our proposal. Section 4 concludes with a discussion about the new questions that our proposal triggers as well how this proposal should be evaluated and its applicability.
2 Related Work

In order to solve the IOP, VA proposes a model that combines the strengths of automatic data analysis and interactive visualization. Underlying to this model, there is a diversity of challenges. Once we are dealing with spatio-temporal datasets, particular issues arise related to the IOP as temporal and spatial data have its own properties and structures. In order to handle with those issues (discussed below), a new recently research sub-area emerged, called Geovisual Analytics (GA) (G. L. Andrienko et al., 2007) (G. Andrienko, Andrienko, Keim, MacEachren, & Wrobel, 2011), as a sub-area of VA.

Spatial and temporal data are characterized by its dependency and heterogeneity (Sahli & Jabeur, 2010). Regarding spatial data, dependency can be explained through the Tobler's first law: “everything is related to everything else but nearby things are more related than distant things” (Tobler, 1970). In other words, the characteristics observed at close locations tend to be more correlated and similar. Likewise, the dependency property is also verified by temporal data. By contrast, the heterogeneity refers to high difference between the observations. For instance, the behavior of people who live near each other could differ a lot; or radical changes may be caused by an event.

Although spatial and temporal data have in common those properties, they have different semantics/structure (Galton, 2009) (G. Andrienko et al., 2010) (Sahli & Jabeur, 2010). Time has unique scaling and granularity properties while space verifies topological relations between spatial objects which could have different structures/representations. Thus, the combination of spatiality and temporality results in complex spatio-temporal phenomena (Sahli & Jabeur, 2010).

In almost areas of human activity, spatio-temporal data are produced and stored like in telecommunications, environmental, criminology, travel and tourism, among others. More specifically, an analyst may want understand, for instance, the animal behavior (e.g., movements, migration), the evolution of the telecommunications carries' clients, the evolution of risky areas for a given crime, or the changes verified in the country's population (using census data, for example). Thus, spatio-temporal data represent spatial entities evolving over time.

A spatio-temporal object may change in three different ways over time (Natalia Andrienko, Andrienko, & Gatalsky, 2003): (i) existential changes, i.e. appearance and disappearance (ii) changes of spatial properties, i.e. location, shape or/and size, orientation; (iii) changes of thematic properties, i.e. change on descriptive attributes.

Changes in spatial objects trigger modifications of their relationships over time with other objects. Spatio-temporal predicates are important to understand those interactions and they express developments occurred over time (Martin Erwig & Schneider, 2002). Considering two objects represented by regions, an example of a development predicate can be Leave defined as: Cross: disjoint $\rightarrow$ meet $\rightarrow$ overlap $\rightarrow$ inside $\rightarrow$ overlap $\rightarrow$ meet $\rightarrow$ disjoint. For
instance, we could use it to answer to the follow query (Vaisman & Zimányi, 2009): *for each district and polluting cloud, give the duration of time when the cloud passed over the district.*

Hereupon, spatio-temporal predicates are essential to capture the changes, evolution or even spatio-temporal patterns present in spatio-temporal data (M Erwig & Schneider, 1999). However, in general, spatio-temporal phenomena are in constantly changing which causes a unexpected number of possible developments and interactions (Martin Erwig & Schneider, 2002) (Galton, 2009). Consequently, how to retrieve syntheses of meaningful changes where everything is changing is non-trivial. Besides that, much more issues arise when we are interested in extract syntheses from large spatio-temporal datasets.

Spatial, temporal or spatio-temporal data mining algorithms should be used, once classical data mining algorithms do not attend to the properties that make spatio-temporal data a special data type (Miller & Han, 2009) (Leung, 2010). For instance, global parameters could not reflect well the phenomenon occurred at a particular location as we need to consider the heterogeneity property. Additionally, the dependency property makes the standard data mining techniques insufficient as it are assumed independence among the observations (Sahli & Jabeur, 2010).

Another concern is the scale at which the analysis is performed. The term scale refers to the level of detail that the data is observed. Different scales could lead to different spatio-temporal syntheses. The patterns and relationships at one scale may not exist or be detected while in other scale such pattern or relationship emerges. In short, the choice of temporal and spatial scale influence the data achieved which in turn affects our understanding about the phenomenon under study (G. L. Andrienko et al., 2007) (Leung, 2010). Presently, how to discover knowledge in data with multiple spatial and temporal scales is a challenge in the development of data mining techniques (Leung, 2010).

In order to become conceivable anticipate meaningful syntheses from spatio-temporal datasets, there is a need to develop methods capable: *(i)* to detect at which analysis' scale important information given a spatio-temporal phenomenon is visible (G. Andrienko et al., 2010); *(ii)* able to provide effective data syntheses in an efficient way from spatio-temporal datasets of large size, high dimensionality and complex (Mennis & Guo, 2009). However, overcome the issues mentioned so far solve one part of the IOP: *how to process spatio-temporal data in an appropriate way.*

A purpose of the visualizations methods is to provide a simply way to display data/syntheses, allowing the user understand them effortless. There are several visualization methods developed to represent spatio-temporal phenomena (M. Kraak, 2003) (D. A. Keim, Schneidewind, & Sips, 2004) (N Andrienko & Andrienko, 2006) (Ham, 2010). The number of visualization methods is quite big. Different visualization methods have different purposes. Independently from the visualization methods available, when dealing with spatial data the use of cartographic displays is indispensable for enabling analysts to deal with spatial information as it allows better and faster global perception of the information, including the
possibility to discover correlations between phenomena, understand spatial relationships, among others (Bédard, Rivest, & Proulx, 2006). Yet, the power of maps is limited (Silva, Moura-Pires, & Yasmina Santos, 2011), especially representing multi-attribute and dynamic data (Gennady Andrienko, Andrienko, et al., 2011). To overcome those limitations, usually it is combined maps with other visualizations methods. However, it is not enough to solve another part of the IOP: how to visualize spatio-temporal data in an appropriate way. The role of the visualization is still an open issue when we want to deal with spatio-temporal phenomena described through huge amounts of raw data characterized by its high dimensionality (Gennady Andrienko et al., 2010). Easily the visualization methods get cluttered and difficult to analyze.

All the issues discussed so far mainly concerns about how to process and present spatio-temporal data. In order to solve the IOP, one more issue needs to be addressed: how we should anticipate relevant syntheses given a spatio-temporal dataset to the decision maker. Therefore, let us discuss how this can be addressed.

3 The Important Relies on Changes

In general, natural and man-made phenomena change over time and space (Leung, 2010). For instance, in country’s population changes occur in birth/mortality/ unemployment rate; in health, the number of persons affected by a determined disease; in business, the evolution of the company customers (e.g., which customers they are losing or acquiring) is important to define new strategies; in pharmacovigilance science, changes/detection of adverse drug reactions are essential to determine whenever a drug should stays on market or not, and so on.

Our world is dynamic and a lot of changes are constantly occurring. However, not all of the changes are equally important for the analyst. We may have unexpected changes; expected changes but with a larger intensity than homologous time periods; long-term tendencies which accentuate over time; trend reversal; and so on. So, the issue is to detect relevant changes for an analyst at a given context.

Even so, by doing this restriction, the identification of what is important remains non-trivial. First, we envisage a framework domain independent which brings a greater complexity as different domains have different ways to change/evolve. Second, underlying spatio-temporal dataset everything is changing. Often the possible changes of a given phenomenon are unpredictable and only the most relevant should be considered. A spatio-temporal object may or may not change over time but if any change occurs, it can change its attributes, spatial properties or in fact disappear/appear as mentioned previously. So, there is an unpredictable possibilities of spatio-temporal objects evolve and interact with each other. Furthermore, a change can occur at different rates, i.e. it could be a long-term (e.g., change climate) or short-term (e.g., Portugal became a member of the European Union). This brings us to the follow issue: a change can occur at different scales either in spatial or temporal
viewpoint. As a result, what at a one scale may be relevant, at another may not have an impact/importance. This factor can be useful when we are dealing with huge volumes of data as we can adapt the relevant changes to the analysis’ scale level pretended by the decision maker.

Moreover, the relevant changes should be context dependent. By context we means external factors like (i) user or users involved (user profile based on historical interactions); (ii) temporal context i.e. the moment in which the interaction is taking place (e.g., Christmas season); (iii) an eventual intention expressed by users. Also, the significant changes’ identification should take into account all useful metadata like multidimensional model information, for instance. Furthermore, some domain specific knowledge should be considered if it is available.

Another concern emerges when several significant changes are returned. In such scenarios, approaches are needed to select one result from a set of possible ones. A possible approach is to rank the results based on a relevance function. Whatever be the relevance function, it should model the decision maker notion of relevance. Another possible approach is to agglomerate similar results, allowing the user to have access to a diversity of results. Actually, through the diversity of results may be a way for the user to find unexpected relevant information.

![Figure 1 The approach proposed](image)

Thus, based on the previous discussion, one approach is proposed and it is synthesized in Figure . Our proposal focus on analyze first and show the important steps specified in the VA Mantra. In order to overcome the information overload problem, analyze first step evaluates and detects possible relevant changes to the user. Its detection depends on the context, domain specific knowledge, user participation and metadata. Those four inputs were defined as a way to meet the relevance notion of a user in order to display the most important relevant change.

A long path is envisaged in order to turn this conceptual proposal into a well-specified framework. Hence, in the following section, we will sketch a research path discussing how this proposal should be evaluated, its applicability as well its potential for further research.
4 Discussion

Our proposal triggers new research questions. So far, we have used the term relevant change as referring the important change for an analyst at a given context. The first step towards to the development of a framework capable to anticipate relevant changes is a relevant change modeled concept. Such concept should model a significant changes from phenomena with different inherent characteristics (G. L. Andrienko et al., 2007) like: spatial extent (local, regional, global); temporal extent (time-critical, long-term, one-time decision, long-term sequence of linked decisions); domain (telecommunications, government, health, transportation, environmental, among others).

Furthermore, in order to anticipate information based on the relevant change concept given a dataset of complex data, it is important design how to compute and visualize such information. With this in mind, we should take the VA model as basis. A framework composed by an analytical component and visual component is envisaged. The former is meant to detect and compute relevant changes from available spatio-temporal datasets while the latter assigns them a visual representation.

Regarding the analytical component, it is envisioned two ways to detect and compute relevant changes: (i) use or develop spatio-temporal data mining techniques with the relevance change concept embedded; (ii) or, a multi-resolution based approach with the relevance change concept embedded. However, research is needed in order to evaluate these two alternatives, discussing the benefits and limitations of each one.

Concerning the visual component, appropriate visualization(s) method(s) should be used in order to allow him easily understand the significant changes. According to the relevant change detected/computed, appropriate visualization methods should be used. There are several visualizations methods, and the user should not have to think about how he should visualize data. Thus, in the visualization component should be a module that suggests which visualization method or what combination of visualizations methods is suitable to represent a determined relevant change so that the user does not get lost in visualization methods.

Moreover, it is through the visualization that the user participation will take place. It is envisaged an important role of the user participation. Hence, an interaction model should be defined in order to take advantage from the user strengths in those different tasks, i.e. (i) How the decision maker will intuitively participate in the anticipation process without specifying any attributes or metrics? (ii) How the decision maker will participate in the computation process? (iii) How the user will interact with the visualization suggestion model?

Once the proposal is defined a proper evaluation of the framework proposed is needed. We consider adequate to have two of levels of evaluation. First, we should perform some individual evaluations like to evaluate the integration of the relevance change concept in the existent spatio-temporal data mining algorithms; to evaluate the visualization suggestion model. Then, the framework should be evaluated as whole in order to evaluate the level of users' satisfaction.
We believe through our proposal the decision maker will not get lost on data available to him, since he has important data as a starting point of his analysis. Furthermore, we aim a framework applicable to different applications domains as we believe such framework can be useful in different domains.

References


Abstract
This paper discusses the need for the semantic description of web services that implement map generalisation algorithms to allow the services to be selected automatically. Particular focus is placed on need to define a common set of parameters that can be used by any number of algorithms.

Keywords: automatic generalisation; semantic web services; on-demand mapping

1 Introduction
This project focuses on automated mapping techniques. The development of Google Maps has led to a vast number of “mashups” where users can overlay their own data on a Google Maps background and make the result available to others. The problem with this approach is that the creator is limited to using the background map as supplied and there is no opportunity to vary the content depending upon the context. The lack of integration of user-supplied data leads to cartographic problems such as road names being obscured by overlaid cycle routes, for example (Figure ).

What is required is a system to allow data from a variety of sources to be mapped at a variety of scales. Since the possible combination of datasets and scales is too numerous to be pre-defined, on-demand generalisation (deriving smaller scale maps from larger scale maps) integrating user-supplied data is needed. There have been attempts to provide online on-demand maps (Kopf, et al., 2010) but such systems have been developed by applying a fixed sequence of generalisation operations to a known set of data.
The aim of the project is to design a system for the automatic generation of workflows for on-demand mapping. This will involve the automatic selection, sequencing and execution of generalisation web services based on the OGC’s Web Processing Service (WPS) (Open Geospatial Consortium, 2011).

The steps involved in creating and executing the workflow are abstract composition, concrete composition, and execution.

The abstract composition phase will involve taking the user’s requirements and the descriptions of the data sources and generating an abstract workflow that consists of generalisation operators (simplification, smoothing, amalgamation etc.) The concrete composition phase will involve finding the appropriate web services for those operators and the execution phase will involve calling and managing those services. This paper focusses on the concrete composition phase and in particular two issues; 1) how to annotate the web services in such a way that they can be automatically selected, and 2) once a service has been selected, to supply values to its parameters automatically.

1.1 Definitions
A single WPS service can perform a number of processes. Each process will implement one or more generalisation algorithms (atomic or composite processes). An algorithm may implement one generalisation operator (e.g., building amalgamation) or a combination of operators (e.g., road simplification and smoothing).

1.2 The need for the semantic description of services
The selection and sequencing of the generalisation operators, to build the abstract workflow, will be derived from the system’s cartographic knowledge, for example, generalisation of a road network might require selection, simplification and smoothing in that order. The issue is how do we select the appropriate web services to perform these generalisation operators (concrete workflow)? We need semantic descriptions of the web services to allow for the automatic selection and parameterisation of algorithms. However, it is a well-documented problem with the OGC protocols that they provide for syntactic interoperability but not semantic interoperability (Janowicz et al., 2010; Lemmens et al., 2007). A free-text description field is not sufficient to allow for the automatic selection of a process.

The second issue is, once selected, how do we execute the services i.e. supply the required parameter values? WPS services are syntactically interoperable in that for any parameter a data type is defined (double, integer etc.). This is not sufficient; we need to know the meaning of the parameters. For example, the point clustering ISODATA algorithm (Li, 2007) has two parameters both of which represent a distance – how will the system be able to distinguish between the two?
Another problem is that some algorithms will have geographically meaningless parameters. The ‘simulated annealing’ approach to automatic feature displacement includes among its parameters an ‘initial temperature’ and a ‘temperature gradient’ (Ware et al, 2003). How can values for these parameters be derived from user requirements expressed in geographic terms such as target scale? Other algorithms such as the Douglas-Peucker line simplification have a single parameter, ‘tolerance’, which has a geographic meaning i.e., the maximum distance from the original line and the simplified line (Douglas and Peucker, 1973). However, the Visvalingam-Whyatt algorithm, which also performs line simplification, has a different concept of tolerance, a minimum area (Visvalingam and Whyatt, 1993).

Some algorithms define as many as six parameters (Revell, 2004; Ware et al 2003) others have none (Yan and Weibel, 2008). Every time a new algorithm was exposed by a web service the system’s knowledge would have to be updated. Rather than the system having knowledge of all of the parameters for every algorithm it may use, it is more sensible for the system to define a common set of parameters that all algorithms must use. It would be the responsibility of the service to translate from the common parameters to the specific parameters of the algorithm.

In summary, processes need to be described in such a way that the system selects the most suitable WPS process for each task and then provides appropriate values for their parameter. But what information is required to describe a service sufficiently? Section 3 describes a classification of generalisation algorithms and their parameters. Section 4 includes a survey of generalisation algorithms based on this classification. Section 5 discusses the conclusions and future work.

2 Related work

How are the generalisation web services to be semantically enhanced? One approach is semantic annotation (Maue et al., 2009; STI Innsbruck, 2011) where OGC standards are enhanced with semantic descriptions. A comprehensive approach is the semantic enablement layer (Janowicz et al, 2010) that includes a web ontology service that manages a processing ontology and a features ontology, and a web-reasoning service that encapsulates a semantic reasoner that aims to combine processing and features to meet the user’s needs.

Touya et al (2010) describe translator functions that take the set of constraints and set of rules held in the system knowledge and the user requirements and translates them into algorithm parameters. Each algorithm will have its own translator function. This approach embeds the translator component in the system rather than ask each generalisation service to manage its own translation.

When considering line simplification Foerster et al. (2007) defined a simplification ratio that could be used by both the Douglas-Peucker and the Visvalingam-Whyatt algorithms. The value for the ratio is dependent on the user’s choice of target scale and is based on the ratio of the number of vertices before and after simplification. The ratio is a variation on the
Radical law of selection (Töpfer and Pilliwizer, 1966) which is used to determine the number of objects to be retained on a map at reduced scale based on the number of source objects and the ratio of the source and target scales.

The Radical law is also used in a point generalisation algorithm (Yan and Weibel, 2008) as a limit on the number of iterations the algorithm performs. The same algorithm considers the importance value of the source points when deciding to retain a point. By employing the user requirements (target scale) and the source data (number of points and importance values) this algorithm requires no parameters. Could these concepts be used to provide parameter values for those generalisation algorithms that do have parameters?

### 3 Classification of Algorithms

Following discussions with the project’s partner, the Ordnance Survey of Great Britain, a set of attributes that were needed to describe generalisation algorithms, and hence any web service that implemented them, was defined and extended (Table 1).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Free text description of what the algorithm does (this will not be part of the algorithm’s semantic description but will aid its construction).</td>
</tr>
<tr>
<td>Operators</td>
<td>The generalisation operator(s) that the algorithm implements. This needs to be classified.</td>
</tr>
<tr>
<td>Feature types</td>
<td>The feature type that the process applies to; e.g. buildings. Some algorithms are either specific to, or work best on, particular feature types. We will need a classification of feature types.</td>
</tr>
<tr>
<td>Geometry of the input data</td>
<td>We need to know what type of data the algorithm uses.</td>
</tr>
<tr>
<td>Scale related information</td>
<td>Maxima and minima for target scale and scale ratio.</td>
</tr>
<tr>
<td>Nature of source data</td>
<td>Some algorithms work better on rural buildings than they do on urban buildings, for example. This will have to be classified.</td>
</tr>
<tr>
<td>Data quality</td>
<td>In particular data input quality – e.g. for algorithms operating on networks issues such as a badly formed network or duplicate vertices can cause problems.</td>
</tr>
</tbody>
</table>
Performance | Some algorithms may be restricted in the size of input dataset they can operate on; e.g., a maximum number of features.
---|---
Time factor | Some algorithms may produce high ‘quality’ results but take a relatively long time and vice versa (the selection of such services will be dependent on user requirements). This needs to be quantified.

**Table 1** Algorithm descriptions

<table>
<thead>
<tr>
<th>Model generalisation</th>
<th>Cartographic generalisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class selection</td>
<td>Enhancement</td>
</tr>
<tr>
<td>Reclassification</td>
<td>Displacement</td>
</tr>
<tr>
<td>Collapse</td>
<td>Elimination</td>
</tr>
<tr>
<td>Combine</td>
<td>Typification</td>
</tr>
<tr>
<td>Simplification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amalgamation</td>
</tr>
</tbody>
</table>

**Table 2** Operator classification (Foerster et al, 2007)

Since the abstract workflow will be based on generalisation operators we will need to classify algorithms by the operator(s) they implement and for that we will need a consistent way of classifying operators; do *filtering* and *elimination* mean the same thing, for example? The most recent attempt to classify operators was by Foerster et al (2007) (Table 2) which makes a distinction between model and cartographic generalisation. However, this classification is at too high a level of abstraction; for example, *enhancement* includes *exaggeration*, *smoothing* and *enlargement* so it is not sufficient to state that a road network requires *enhancement*, we need to be more specific.

A complement to the generalisation operator method of classifying algorithms would be to adopt the descriptions used by Li (2007) in his survey, where the nature of the source data is incorporated in the description, hence we have categories such as ‘smoothing individual line features’ and ‘transformation of individual area features’ (e.g. building simplification). Li introduces sub-categories that refer to the processing method so ‘transformation of individual area features’ can have the sub-categories ‘Boundary-based shape simplification of an area feature’ and ‘Region-based shape simplification of an area feature’. Also included are subcategories that relate to the generalisation operator they implement such as ‘Collapse of area features’. Although the categorisation used by Li has not been formalised it
offers the advantage of describing what action is being taken on what type of data and is thus relatively expressive.

3.1 Description of parameters
A method of describing parameters is required such that the workflow system can automatically provide values for the parameters based on the user requirements and the nature of the data that is to be generalised. A number of attributes for describing parameters was defined (Table 3) based on the requirement to describe a parameter sufficiently that values can be supplied automatically. The next step was to apply the classifications to a set of generalisation algorithms.

4 Survey of generalisation algorithms
The algorithms and their parameters were described using the categories listed above (Table 1, Table 3). It was decided to focus on point clustering algorithms, line simplification and smoothing, and building simplification and amalgamation. This includes point, line and polygon data. Table 4 provides an example description.

To give an indication as to the variety of parameters, the name of each parameter was extracted from the survey data and represented as a word cloud (Figure 2).

*Figure 2 Word cloud of parameter names*
The survey has exposed the large variety of parameters in just a subset of the generalisation operators. The word cloud serves to highlight certain parameters such as *minimum distance* that appear frequently. However, the meaning of what minimum distance represents varies from algorithm to algorithm.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name of the parameter e.g. <em>maximum distance</em></td>
</tr>
<tr>
<td>Description</td>
<td>Free text description of the parameter (this will not be part of the algorithm’s semantic description but will aid it).</td>
</tr>
<tr>
<td>Class</td>
<td>What the parameter represents. E.g. a <em>tolerance</em></td>
</tr>
<tr>
<td>Effect on output</td>
<td>Describes the particular impact of the parameter.</td>
</tr>
<tr>
<td>Weight</td>
<td>How important is this parameter in relation to other parameters of the algorithm? This is a quantification of the previous parameter</td>
</tr>
<tr>
<td>Necessity</td>
<td>Can the parameter be omitted or can a default be provided? This needs to be codified as a Boolean attribute.</td>
</tr>
<tr>
<td>Data type</td>
<td>Double, integer etc.</td>
</tr>
<tr>
<td>Units</td>
<td>Linear units, areal units etc.</td>
</tr>
<tr>
<td>Range</td>
<td>Range of possible values for the parameter. This can be used as a check against non-sensible values being supplied.</td>
</tr>
</tbody>
</table>

*Table 3* Description of parameters
Algorithm | ISODATA  
---|---
References | Li, 2007, p78  
Li Category | Aggregating point features (clustering)  
Operator(s) | Combine  
Description | Repeatedly split a cluster until the maximum standard deviation of distance from centre in each direction \((x,y)\) is reached.  
Application | Punctual events  
Source data | Set of points  

## Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Class</th>
<th>Effect on output</th>
<th>Necessity</th>
<th>Weight</th>
<th>Data type</th>
<th>Units</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma_{x,\text{max}})</td>
<td>Max. standard deviation from centre of any cluster in (x) direction</td>
<td>Maximum Distance</td>
<td>Fixes the maximum width of any cluster</td>
<td>Required</td>
<td>0.5</td>
<td>Decimal</td>
<td>Linear Unit</td>
<td></td>
</tr>
<tr>
<td>(\sigma_{y,\text{max}})</td>
<td>Max. standard deviation from centre of any cluster in (y) direction</td>
<td>Maximum Distance</td>
<td>Fixes the maximum height of any cluster</td>
<td>Required</td>
<td>0.5</td>
<td>Decimal</td>
<td>Linear Unit</td>
<td></td>
</tr>
<tr>
<td>(K)</td>
<td>Number of clusters</td>
<td>Goal</td>
<td>Determines number of clusters</td>
<td>Optional</td>
<td>0</td>
<td>Integer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>Number of iterations</td>
<td>Speed, Time</td>
<td>Optional</td>
<td>0</td>
<td>Integer</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4** Example description of an algorithm (ISODATA)

The challenge is to define a common set of parameters that is neither too restrictive nor too broad. Their values should be derived using the user requirements and the source data. The
value of any parameter defined by an algorithm would be a function of one or more of the common parameters. The relationship need not be a continuous function, it could be, for example, that for a range of values for a common parameter the algorithm’s parameter could take a single value.

Since the use case focuses on the clustering of road accidents, the initial investigations will be based on clustering algorithms. There are a number of clustering algorithms but two, ISODATA and K-means, both include all source points in their generated clusters. K-means has a single parameter K which represents the number of clusters to be generated. As can be seen from Table 4 this parameter is shared with ISODATA. However, for ISODATA, K is only an optional parameter and is used with n, the number of iterations (optional), to put a stop on the iterative process. Its key parameters, \(\sigma_{x,\text{max}}\) and \(\sigma_{y,\text{max}}\) define the maximum size of the resulting clusters.

Further work will be carried out to examine how the output of each algorithm varies with their respective parameters in an attempt to define one or more common parameters that is defined by both the source data (for example, the initial point density) and the required output (target scale).

5 Conclusions and further work

The next stage is to implement and formalise the algorithm classification described above by using it to semantically describe the generalisation services. This will be aided by a formalisation of the classification used by Li (2007).

The concept of a semantic enablement layer (as described by Janowicz et al., 2010) to facilitate service selection will be investigated further with a view to describing the algorithms/services using an ontology such as OWL-S (Martin et al., 2007) and a reasoning engine such as Racerpro to perform ontological matching (Lemmens et al, 2007; Fitzner et al., 2011; Janovicz et al, 2010).

Regarding the parameterisation of the algorithms, the survey has shown the diversity in their parameters emphasising the need to define a common set and work will continue to that aim.

References


Geospatial Ontology Harmonization

Heshan Du
University of Nottingham, United Kingdom
E-Mail: psxhd1@nottingham.ac.uk

Abstract
Owing to the distributed nature of ontology development, semantic heterogeneity is one of the main obstacles of information sharing and exchange in several fields. In geospatial science, there is an increasing need of using disparate and complementary geospatial information sources synergistically. Ontology alignment is a promising approach to tackling the problem of semantic heterogeneity in order to obtain harmonized and maximized information content from disparate geospatial ontologies. This project employs a logic-based approach for geospatial ontology alignment, using Distributed Description Logics as the theoretical framework to represent and reason with semantic relations. Our main ideas for establishing semantic relations from underspecified ontologies are discussed, compared with CtxMatch.

Keywords: geospatial ontology, ontology alignment, semantic mapping, distributed description logics

1 Introduction
Nowadays, a large amount of disparate information has been or is being made available by people and organizations around the world. This provides rich resources for information exchange and sharing. However, owing to the distributed and decentralized nature of information development, it is not easy to fully capture the information content in different sources. One of the main obstacles is semantic heterogeneity, which often occurs when terms are used differently. For example, people may use a same expression to mean different things, or use different expressions for the same meaning. Such issues are quite common when disparate but also related information sources exchange their data.

Ontology plays an important role in information sharing. Ontology, originated in the work of Aristotle, is a branch of philosophy which studies the existence of entities (Welty, 2003). In computer science, ontology is a formal, explicit specification of a shared conceptualization (Gruber, 1993). Compared to its origin, computer science ontology is not only about existence, but also about meaning, and making meanings as clear as possible (Welty, 2003). More specifically, ontology refers to a formal description of a domain of discourse, expressed in a language that can be used for reasoning, and intended for sharing among different applications (Noy, 2004). In artificial intelligence, ontologies are often used as important means for establishing shared explicit formal vocabularies. However, since it is not realistic to agree on a common ontology among different communities with their own preferences, ontology alignment is often needed.

The rapid developments in geospatial science emphasize the importance of geospatial information sharing and exchange. Spatial Data Infrastructures (SDI) refers to an institutional and technical foundation of policies, standards and procedures that enable organizations at
multiple levels and scales to discover, evaluate, use and maintain geospatial data (Nebert, 2004). In the last few years, the current top-down approach to SDI is being challenged by the rapid pace of technological development (Jackson, 2011). There is a need to address the separation of national and international SDI from crowd-sourced geospatial databases (Anand et al, 2010). Relying on volunteers for data collection, crowd-sourced data is less expensive than authenticated data. In addition, although typically not as complete in its coverage or as consistent in its geometric or metadata quality as authenticated data, crowd-sourced data may provide a rich source of complementary information with the benefit of often more recent and frequent update than that of authenticated data (Jackson et al, 2010). It is desirable if crowd-sourced geospatial data can be used synergistically with authenticated geospatial data.

The rest of the paper is organized as follows. Section 2 specifies the research question. Section 3 discusses existing approaches for ontology matching, particularly a logic-based method called CtxMatch. Distributed Description Logics, as the theoretical framework, will be explained in Section 4, and our main ideas will be discussed in Section 5. Finally, it provides conclusions in Section 6.

2 Research Question

The research question is how to deal with ambiguity or inconsistency caused by semantic heterogeneity in disparate geospatial ontologies, and benefit from the information difference, to obtain harmonized and maximized information content.

Semantic heterogeneity often occurs when disparate information sources use terminologies differently. This project focuses on three main kinds of semantic heterogeneity. Firstly, different expressions may refer to a same meaning. For example, a road may be renamed for historical reasons, or different parts of it may have different names, or it may have localized names. Secondly, a same expression may refer to different meanings. For example, in the US, “pavement” refers to the surface of a road, whilst it usually means a walkway beside a road in the UK. Thirdly, different ways of classification can be employed in disparate information sources. For example, one community classifies streets based on traffic and width, while the other only based on width.

The end goal of this research is to harmonize information from disparate sources, such that the information content is maximized. To achieve it, firstly, at the terminology level, it is important to establish the semantic relations between concepts (and roles) from disparate ontologies. In addition, at the instance level, it is necessary to find the correspondences among instances or data elements, and allow instances to migrate from one ontology to another.

This project falls into the field of semantic alignment, which involves several dimensions, such as mapping discovery, declarative formal representation of mappings, and reasoning with mappings (Noy, 2004). The core of the semantic alignment is to establish appropriate semantic relations between entities from different ontologies. The semantic relations include disjointness, equivalence, subsumption, overlap, etc. The technical sub-problems of this project include generating, representing, verifying the semantic relations between entities (concepts, roles, and individuals) from disparate ontologies, and using the semantic relations to translate (geospatial) information from one context to another, maximizing the
information content during this process. With respect to the central research question, the main idea is that, disparate ontologies will be harmonized by establishing semantic relations between their entities, and information content will be maximized by extracting information as much as possible.

3 Existing Approaches

Automated and semi-automated ontology matching is an active research area. A lot of tools have been developed, based on shared upper ontologies, if available, or using other kinds of information, such as lexical and structural information, user input, external resources and prior matches (Noy & Stuckenschemidt, 2005). According to Bouquet (2007), the approaches for ontology matching can be classified into three broad categories, syntactic, pragmatic and conceptual approaches.

- Syntactic approaches rely on a syntactic analysis of the linguistic expressions (the properties of labelling strings) to generate mappings. While these methods are highly effective, semantic relations between entities are not captured.

- Pragmatic approaches infer the semantic relationships between concepts from their associated instance data. Though they work well in some cases when the instance data is representative and overlapping, this kind of methods use a very strong form of induction from particular to the universal, which leads to wrong associations.

- Conceptual approaches compare the lexical representation of concepts to compute mappings. Socially negotiated meanings (e.g. in dictionaries) need to be used when generating relations between concepts, making the problem very complicated.

Many methods are hybrid, combining different approaches, and making use of structural, lexical, domain or instance-based knowledge. Most of them apply heuristics-based or machine learning techniques to various characteristics of ontology. To look at the state of the art in ontology mapping research before 2004, see the comprehensive survey (Kalfoglou & Schorlemmer, 2003). The CtxMatch will be explained below, since it is most relevant to our method.

While most methods compute linguistic or structural similarities, the CtxMatch is a logic-based algorithm, which shifts the semantic alignment problem to deducing relations between logical formulae, which is a significant improvement compared to the previous methods (Bouquet et al, 2003; Bouquet et al, 2004). It was initially designed for semantically coordinating hierarchical classifications (HC), which are structures with the explicit purpose of organizing or classifying some data (Bouquet et al, 2003; Bouquet et al, 2004), later it is extended to match hierarchical classifications with attributes (Serafini et al, 06).

The CtxMatch algorithm represents relevant lexical, world and structural knowledge as logical formulae, and uses logical reasoning to infer semantic relations. It includes two main steps: semantic elicitation and semantic comparison (Serafini et al, 06). Firstly, the semantic elicitation approximates and builds a formal representation of the (global) meaning of each node. The global meaning is generated by combining internal or structural semantics and external semantics, which may come from lexical knowledge (concepts denoted by words) and world knowledge (relations between concepts) (Serafini et al, 06). WordNet (Miller, 1995) is employed to provide both lexical and world knowledge, but it is replaceable.
Secondly, the semantic comparison process computes the semantic relations between two nodes from different HCs, by comparing their meanings via logical reasoning. Though the logic-based method ensures the consistency of generated mappings, it is often criticized for being too restrictive and missing matches which are correct but not derivable (Meilicke, 2008).

4 Theoretical Framework

Distributed Description Logics (DDL) are employed as the theoretical underpinning for semantic mapping representation and reasoning, with the benefits of treating mappings as first class entities and accommodating reasoning in a distributed manner (Meilicke, 2008). In this section, we will recall some main definitions of DDL related to this project.

Description Logics are a family of formalisms for representing the knowledge of an application domain (Baader et al, 2007). They firstly define the relevant concepts (terminology), and then use these concepts to specify properties of objects in the domain (Baader et al, 2007). Compared to traditional first-order logic, Description Logics provide tools for the manipulations of compound and complex concepts (Brachman & Levesque, 2004). Description Logics support inference patterns, including classification of concepts and individuals, which are often used by humans to structure and understand the world (Baader et al, 2007).

Distributed Description Logics (DDL), an extension of Description Logics, can be seen as a framework for formalizing the representation of a set of ontologies pairwise interconnected by semantic mappings (Borgida & Serafini, 2003).

Recalling the definition of DDL given in (Borgida & Serafini, 2003), given a set $I$ of indexes, a DDL is a collection of ontologies. The T-box (terminologies) and A-box (assertions) of ontology $DL_i$ are denoted as $T_i$ and $A_i$ respectively.

Semantic relations between entities from disparate ontologies can be encoded as bridge rules. See the definition of bridge rules below (Ghidini et al, 07).

**Definition 1.** A (homogeneous) bridge rule from $i$ to $j$ is an expression of one of the two forms:

$$i : X \sqsubseteq Y \quad (1)$$

$$i : X \sqsupseteq Y \quad (2)$$

where $X$ and $Y$ both are either concepts or roles, and $i$ and $j$ are ontologies.

Bridge rules from $i$ to $j$ express semantic relations from the *subjective* point of view of the ontology $j$, rather than some external objective point of view. The into bridge rule (1) states that, from the point of view of the ontology $j$, $X$ is more specific than $Y$. Similarly, the onto bridge rule (2) states that, from the point of view of the ontology $j$, $X$ is more general than $Y$. 
Bridge rules from $i$ to $j$ translate the concepts or roles of the foreign ontology $i$ into the local ontology $j$ under some approximation introduced by an implicit semantic domain relation.

Individual correspondences, as the individual level equivalent bridge rules, are introduced to express the correspondences between individuals from different ontologies (Serafini & Tamilin, 2005; Serafini & Tamilin, 2007). The individual correspondence $i: x \rightarrow j: y$ expresses that, from the point of view of the ontology $j$, the individual $y$ is one possible translation of the foreign individual $x$ in the local domain $j$.


5 Main Ideas

To undertake our research, disparate but related geospatial ontologies are required. The Ordnance Survey Great Britain ontology and the DBpedia ontology are selected. Ordnance Survey Great Britain has built ontologies for Hydrology and for Buildings and Places (Hart et al, 2008). DBpedia has a crowd-sourced ontology (wiki.dbpedia.org/Ontology, 2011), based on the most commonly used infoboxes within Wikipedia. These two ontologies are overlapping, but define concepts and their hierarchies differently. For example, in the Ordnance Survey Buildings and Places ontology, Airport is a subclass of Place, and has some Air Terminal Building, while in the DBpedia ontology, Airport is a subclass of Building, which is a subclass of Place. Both ontologies describe the hierarchy of concepts (or roles), rather than defining their exact meanings. In other words, both are underspecified, where equivalence axioms and disjointness axioms are usually not available. This makes it difficult to generate and verify the semantic relations between entities from different ontologies. For example, in the Ordnance Survey (OS) ontology,

$$OS : \text{School} \subseteq \text{Place}$$

$$OS : \text{School} \subseteq \exists\text{hasPurpose.Education}$$

$$OS : \text{School} \subseteq \exists\text{hasPart.(Building} \sqcap \exists\text{hasPurpose.Education})$$

while in the DBpedia (DB) ontology:

$$DB : \text{School} \subseteq \text{EducationalInstitution}$$

The symbol $\subseteq$ means “is included in”, $\exists R.C$ means “has an R-successor which belongs to C”, and $\sqcap$ means “and”.

The exact meanings of School are unknown in both ontologies. Their super concepts only indicate part but not full meanings of School. In other words, only part of the meaning of a concept can be inferred from its super concepts. Knowing the semantic relations between super classes of two concepts, except when the semantic relation is disjointness, it is not sufficient to infer the semantic relations between these two concepts. In addition, since the concept OS: School has more than one super concept, the existing technique CtxMatch,
which is designed for tree-like structures, cannot make full use of information available to
tell the semantic relations between \( OS: \text{School} \) and \( DB: \text{School} \). Based on (3) and (6), the
result generated by CtxMatch can be expressed as the bridge rule
\[
DB: \text{School} \rightarrow OS: \text{School}
\] (7). Without some transformation, CtxMatch cannot use (4)
and (5).

When dealing with underspecified ontologies where exact definitions are unavailable, using
different explicit information or using information differently can lead to different semantic
relations being generated. Realizing the uncertainty and subjectivity in the semantic relation
generation, our method will keep a record of what and how information is used for
generated bridge rules, thus know when and why each bridge rule is valid. In addition,
compared to CtxMatch, it will use the available information more fully and directly by
allowing more expressive power. While CtxMatch is generally criticized for missing matches
which are correct but not derivable (Meilicke, 2008), our method will relax the restrictions
by allowing assumptions to be added explicitly to fill in information or reasoning gaps.
Differing from facts, assumptions may be incorrect, and be suspected later with the
discovery of new knowledge. Both uncertain “facts” and unsound inference rules can be
stated as assumptions. Firstly, assuming some unknowns as the known allows more
inference and deduction to be made. In addition, uncertainty can be modelled by adding
different possible assumptions explicitly, allowing more matches to be generated under
certain assumptions.

This is a reasonable and general approach, which not only has the potential to resolve the
problems of underspecified ontologies in mapping generation and verification effectively,
but also has the capability of incorporating and formalizing some existing ways dealing with
underspecified ontologies. For example, to deal with Example 1, one may assume that, given
that both concepts are denoted by the same English noun phrase “School”, with the current
knowledge, their semantic differences can only be inferred from the semantic differences
between their super classes. This idea can be formalized as the following inference rule, and
employed as an assumption when generating bridge rules.
\[
i: A \subseteq i: C, j: A \subseteq j: D, i: C \not\rightarrow j: D \implies i: A \not\rightarrow j: A
\] (8)
where \( i: A \) and \( i: C \) are concepts in ontology \( i \), and \( j: A \) and \( j: D \) are concepts in ontology \( j \), and
\( \not\rightarrow \) means “implies”. Another example is to deal with unavailability of disjointness axioms,
which are quite useful in the semantic mapping verification. The common idea of
disjointness of sibling concepts can be formalized as the following inference rule, which can
be used as an assumption when verifying bridge rules.
\[
A \subseteq C, B \subseteq C \implies A \subseteq \neg B
\] (9)
where \( A, B \) and \( C \) are concepts.

The generated and verified bridge rules can be used for distributed query answering.
Suppose that Ordnance Survey wants to import data about \( \text{School} \) from DBpedia. From the
perspective of Ordnance Survey, in the case where the bridge rule (7) holds, not all
individuals of \( DB: \text{School} \) can migrate to \( OS: \text{School} \), since \( DB: \text{School} \) may not be a \( \text{Place} \), nor
have a \( \text{Building} \). For example, if an individual returned from DBpedia is an online school,
Ordnance Survey may not wish to treat it as a new individual of \( OS: \text{School} \). More
information (e.g. checking whether the individual in DBpedia has a location) is required to determine which individuals of DB: School can be seen as an OS: School.

6 Conclusion

To obtain harmonized and maximized information content from disparate geospatial information sources, it is necessary to discover, represent and reason with the semantic relations between entities from different ontologies. Distributed Description Logic is employed as the theoretical framework for knowledge representation and reasoning. Compared with CtxMatch, we discussed our main ideas for establishing semantic relations from underspecified ontologies. Initial plan of this project is to design a logic-based algorithm for semantic mapping generation and verification. The established semantic mapping includes semantic relations not only at the terminology level, but also at the instance level. Any suggestions or feedback are welcome.

References


Development of a Framework for Uncertainty-Aware Land Cover Change Analyses with Visual Analytics Methods

Christoph Kinkeldey
HafenCity University Hamburg, Laboratory for Geoinformatics and Geovisualization, Germany
E-Mail: christoph.kinkeldey@hcu-hamburg.de

Abstract

The detection of changes in remotely sensed images is a common way to produce change maps needed for decision making in various fields. For this, a number of methods are included in standard remote sensing software. But the production of reliable change maps remains a challenge because of the relatively low quality of the change detection results. The purpose of this paper is to present a concept to overcome these drawbacks using a Visual Analytics approach to integrate the detection as well as the analysis of changes into iterative workflows. A change analysis framework that is currently under development is presented at an early stage.

Keywords: remote sensing, change detection, change analysis, uncertainty, framework

1 Motivation

Land cover change maps serve as a basis for decision-making processes in a wide range of fields including urban planning, climate research and environmental monitoring. These maps are often produced on the basis of remote sensing data. Especially multispectral satellite data has become popular as it is provides coverage over large areas at relatively low cost. In the last decades, the spatial resolution of the imagery has improved very much and can be less than 10 meters. Besides, the temporal resolution, i.e. the revisiting frequency of modern satellites has also improved.

Methods for producing change maps are subsumed under the term change detection which can be defined as the process of identifying differences in the state of an object or phenomenon by observing it at different times (Singh 1989, p. 989). A number of change detection methods exist ranging from simple image differencing to more complex methods such as Change Vector Analysis (Coppin et al. 2004). Most of the change detection methods can be found in common remote sensing software and are used frequently in practice. However, despite much research in the past they are not yet applicable effectively and efficiently. The high amount of incorrectly detected changes remains a drawback. For instance, Pontius et al. showed in sensitivity analyses that even with highly accurate data the amount of errors in the change detection results can be severe: With user’s accuracies of 91
percent, half of the detected changes were revealed as errors (Pontius Jr and Lippitt 2008). Especially when dealing with multiple images at once (multi-temporal analysis) these effects sum up and have an even greater negative impact on the quality of the change detection results.

One reason for this situation is the high complexity and dimensionality of remote sensing data compared to other kinds of imagery. First of all, this is a matter of the content itself - the description of the Earth’s surface is highly complex (e.g., natural vs. artificial structures), dynamic (e.g., long-term processes vs. rapid developments) and variable (e.g., different appearance of vegetation). Secondly, the lack of transferability of parameters from one dataset to the other is another reason for unsatisfactory results with common methods. Furthermore, each image is just a temporal snapshot of the area from which the changes in between must be reconstructed. But most methods are limited to the comparison of two scenes only - multi-temporal change detection methods (taking more than two scenes into account) are still the exception. Another factor for the drawbacks is that uncertainty information is widely neglected though this information is usually available, e.g. in terms of class membership values or class probabilities that come with most classification algorithms. Besides, the separation between ‘change’ and ‘no change’ is usually done by simple filtering which turns out to be not sufficient for this task.

Our conclusion from this is that a new perspective on the change analysis process is needed. We see the integration of change detection and the analysis of the changes as the key to significant improvements in this field and propose the use of a Visual Analytics approach to achieve this. In this contribution, we present a concept for the development of a change analysis framework based on this idea.

2 Goals

The overall idea of the concept is to support the development of innovative methods for the detection and analysis of land cover change. To achieve this, the following goals have been defined:

1. Improve insight in change process

For a substantial improvement of change analysis results there is a need for a better understanding of the change process itself. Coppin et al state “[...] one of the challenges confronting the remote sensing research community is to develop an improved understanding of the change detection process on which to build an understanding of how to match applications and change detection methods.” (Coppin et al. 2004, p. 1589). Hence, the first goal is to enable a better description of the change process.

2. Leave the paradigm of complete automation behind

In the past, most research in the field of change detection focused on the development of completely automated change detection methods. From today’s experience it is likely that
this goal will never be reached. Nichol et al, e.g., state that “[...] the development of ‘black box’ algorithms, implying complete automation, is probably neither possible nor desirable and researchers are encouraged to work with end users to develop realistic local solutions” (Nichol et al. 2007). So the concept is focusing on the integration of automatic with non-automatic methods.

3. Communicate change uncertainties

Though a lot of work has been done on uncertainty modeling in geographic data (a comprehensive overview on uncertainty modeling in spatial data and analyses can be found in Shi 2010), common methods and GIS-tools rarely support their management and communication. Additionally, according to change uncertainty, there is still the need to define practical models and ways of communicating this information to the user.

4. More powerful filtering

For determining target changes, i.e., change areas of interest, they must be separated from artifacts (false-positive changes resulting from imperfections in the imagery and/or pre-processing steps) and non-target changes (real changes but not of interest for the questions posed). We see great potential in developing more powerful ways to filter detected changes for target changes.

3 Concept and Framework

We propose a user-centered concept for interactive visual tool support to reach the aforementioned goals. The basic idea of our approach is to overcome the drawbacks of existing (semi-) automatic methods by integrating them into a Visual Analytics environment. Visual Analytics is defined as “[...] the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook 2005, p. 4). The main idea is the combination and integration of analytical reasoning techniques and visual representations. According to our application, this means that existing change detection methods are combined with interactive visual methods to support the user with the analysis of changes in an iterative workflow. The framework consists of four parts which will be presented in the following: The description of change objects, of change analysis tasks, of change analysis workflows, and the recommendation of visualization and interaction methods.

3.1 Description of change

The first part of the framework is the systematic description of land cover change. We use the term ‘change object’ for a single change entity: A (geographically) connected area with the same change attributes which we describe in the following way:
Figure 1 Collection of attributes to describe a change object

*change sequence*: land cover classes in chronological order, e.g., “forest → arable land → arable land” meaning that in the first scene, the area was classified as “forest” and in the following two scenes, “arable land” was detected.

*spectral characteristics*: spectral information about the object derived from the remote sensing imagery

*geometry*: geometric attributes such as area and shape

*spatial characteristics*: location, orientation and relative position in geographical space

*temporal characteristics*: course of the change over time and its temporal scale

*uncertainty*: information about change uncertainty

### 3.2 Task definition

In order to improve the insight into the process of analyzing changes, we identify and document typical recurring questions and tasks. This is done in a way similar to van Elzakker who formulates questions and map use tasks for exploratory cartography (van Elzakker 2004). Analogous to this concept, we define questions from change analysis and the tasks associated to them. A small selection can be found in Table 1.

<table>
<thead>
<tr>
<th>Question</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>What changes are there?</td>
<td>to get an overview on change sequences</td>
</tr>
<tr>
<td>How often do certain changes occur?</td>
<td>to quantify the occurrence of change objects</td>
</tr>
<tr>
<td>What is the spatial distribution of certain changes?</td>
<td>to find spatial patterns / anomalies of change objects</td>
</tr>
<tr>
<td>How are the spectral characteristics of certain changes?</td>
<td>to find patterns / anomalies in spectral signature of change objects</td>
</tr>
</tbody>
</table>

*Table 1* Selected questions concerning change analysis and the associated task
3.3 Workflow description

For the systematic design of workflows, their documentation, and for reasons of interchangeability, we see a need for a formalization concept. For this, we created a workflow description concept based on a graph structure. It contains the following categories:

**Hypothesis**: A hypothesis about change

**Computation**: Computational steps in the workflow

**User**: User interaction involved in the workflow

**Visualization**: Visual communication to the user

**Data**: The data used in the workflow (e.g., remote sensing imagery, GIS layers etc.)

For each category, a number of tasks can be defined as nodes and directed edges indicate possible paths in the workflow graph.

Figure 2 shows an exemplary workflow to find artifacts in a set of detected changes. Artifacts are changes that have been falsely detected due to imperfections in the imagery or errors introduced during the processing of the data.

The workflow starts with the import of two or more remote sensing scenes as the basis for the change analysis. The change detection step is done using a standard post-classification algorithm. Next, an overview on all change sequences is presented to the user who can navigate to see the details about certain change sequences and their uncertainty. This is one of the three visualization tasks in the iterative loop, completed by the task to visualize the spatial distribution and spectral information of the change objects. On the basis of this information, the user has the opportunity to define hypotheses on which changes are artifacts by brushing change objects as candidates. With the information about the selected objects the user can revise the hypothesis by changing the candidate set in an iterative loop until he or she confirms it. Eventually, the changes marked as artifacts are deleted and the filtered change objects are exported as the result of the workflow.
3.4 Implementation recommendations

Another goal of the framework is the support for developing change analysis software tools. For the implementation of the workflows, developers will be provided with practical recommendations about suitable methods to fulfill the visualization and interaction tasks defined in the framework. This is still ongoing work but to convey the idea we present three visual components and the tasks they are supposed to fulfill. Interaction methods have not been included yet. Figures 3-5 show first designs of the respective components.

3.4.1 Map view component

A map-based view component (see Figure 3) is designed for tasks concerning the geometry and spatial characteristics of change objects including uncertainty:

- overview on location of change objects
- exploration of geometry (size, shape) of change objects
- exploration of change uncertainty distribution
Figure 3 A first design for the map view. A thematic map shows the spatial attributes of change objects. Change uncertainty is expressed by annotation lines using noise as a metaphor for uncertainty.

3.4.2 Change sequence view
This component (see Figure 4) has the purpose to support tasks dealing with the sequence of land cover classes over time and the associated uncertainty:

- overview on change sequences and their quantities
- comparison of change sequences
- exploration of change uncertainty
Figure 4 A first design for the change sequence view. Each row represents a change object. The circles are color-coded by land cover class and show the sequence of classes and the temporal intervals between them. The uncertainty of the change from one class to the next is represented by opacity.

3.4.3 Spectral signature view

The third visual component (see Figure 5) covers the tasks concerning spectral attributes of change objects:

- overview on spectral signatures of change objects
- comparison of spectral signatures
- exploration of change uncertainty in spectral space

Figure 5 A first design for the spectral signature view. A parallel coordinate plot is used to show the values per spectral band. The change uncertainty is represented by opacity.
4 Conclusion and Outlook

In this contribution we presented a basic concept for the improvement of change analysis in remotely sensed data.

We identified the major drawbacks of current methodology, mainly the relatively low quality of the detected changes. Several reasons for this were mentioned, i.e., the complexity of remotely sensed data, the lack of transferability of change detection parameters, the insufficient support of multi-temporal analyses and the missing integration of uncertainty information.

We proposed the integration of change analysis into the detection step as a promising way and suggested a Visual Analytics approach. Basically, this means that existing methods are combined with interactive visual tools to build iterative workflows. In order to support the creation of such methods and tools based on this concept, we presented a framework that is currently under development, including the description of change objects, the definition of change analysis tasks, the description of workflows and recommendations for visualization and interaction components.

The next step in our research will be the validation of workflows created with the framework. For this, we are currently developing three visual components: Map view, change sequence browser and spectral signature view. These will be implemented as prototypical, fully-functional software tools. On the basis of the software prototypes, we will run expert user tests using real-life change scenarios.

References


Modeling and Validation of Spatial Uncertainty as a Context to Provide Ambient Services

Anahid Bassiri
K.N.Toosi University, Teheran, Iran
E-Mail: anahid_bassiri1984@yahoo.co

Abstract
Ambient services are viewed as services that relate to the surrounding physical environment of the user, have geographical boundaries for their relevance and utility, and can be considered a form of location-based services. When a user enters in a predefined area, ambient services are provided but out of that area services are not available for him/her. The predefined geographic areas in which services are available for users within it, is called ambient service domain. Service domains have sharp and clear boundaries in current ambient services. However because of inaccurate positioning systems and some linguistic and fuzzy phrases used to define ambient service domain (e.g. around, here, far, near), it is necessary to define service domains in a framework which can handle their uncertainty. If an uncertain spatial object is modeled in a certain and crisp framework, some of data related to that object will be missed. There are other aspects in which uncertainty should be considered. When a user is located in several service domains at the same time, it is very important to provide most appropriate set of services. Sometimes all the services associated with all service domain can be provided but what if some of services have conflict with each other? In this regard it is possible to implement Uncertainty-handler Inference system to deduce best service sets. This paper is supposed to view uncertainty in ambient services from three points of view: 1) Inaccuracy in positioning system which causes uncertainty in user position. 2) Linguistic variable which are used to define service domain. 3) Policy conflict resolution to deduce which services should be delivered to a user within several service domains at the same time. In this regard, Rough set and Fuzzy set theory are considered to handle the first two kind of uncertainty. In order to resolve policy conflicts, Rough and Fuzzy inference engines are considered. Simplicity and powerfulness of these two theories are the reasons of picking up them and excluding the other possible devices. It is expected to have an evaluation and compare them from QoS parameters.

Keywords: Uncertainty, Ambient services, Service domain, Fuzzy set, Rough set.
1 Importance of Uncertainty Handling in Ambient Services

Developments of wireless networks and miniaturizing of handheld devices have given rise of mobile services. Context-aware services are one of the mobile services which considers context and situation of user and environment. Ambient services are context-aware services that are related to the surrounding physical environment of the user and can be considered a form of location-aware services (Loke 2006). Ambient services are available in a geographical area which is called service domain (Loke 2006). Whenever users enter into a predefined service domain, ambient services are available for him/her. Based on service domain’s boundaries and user’s location, ambient service provider decides whether user is located in the service domain and should receive the service, or not. Ambient service provider uses such a service boundaries as geographical filters to provide their services. Service domains, which have defined in current systems, have had sharp and clear boundaries. Since there is uncertainty in position of the user and area of service domains, it is better to define and model service domains in a framework which can consider their uncertainty. In addition most of service domains are defined based on linguistic phrases. For example if a new collection of dress around the shopping center is supposed to be advertised, as “around” has not been mathematically and sharply defined, it is better not to model “around” in a crisp framework (Bassiri 2012).

One of the contextual data that ambient services use is position of the user. The accuracy of location information might be within centimeters, meters, or greater. For example, global positioning system (GPS) determines the position of a GPS receiver within an accuracy of several meters (Ziejkolodziej 2006). Mobile phone networks can determine the location of an individual within a suburb or town (around 150 to 300-m accuracy) in the case of an emergency call. Consequently, position of the user, is inaccurate to some degrees. Another aspect of uncertainty is vagueness which related to clarity of description of phenomenon. As we usually describe a service domain by some terminologies of human knowledge, e.g. around this supermarket or not far from this shopping center, etc. Many terminologies express a general characteristic of an object, i.e., they cover a large extent of area. The areas corresponding to these notions are distributed continuously in space and have a characteristic in common – they have indeterminate boundaries. So if we assign a crisp boundary, we may add or lose some part of that area.

Another aspect of ambient services in which there is uncertainty is policy conflict resolution; as service domain is a logical area which is defined by service provider based on application, a user may be located in several service domains at the same time. In such a case, it is very important to decide which combination of services should be provided (Loke 2006). This combination of services may take a number of different forms. It is possible to provide all the services associated with all the service domains surrounding user or services in one area may have precedence over other services in other areas. (Loke 2004) proposed a language of operators for defining such a collection of services in terms of constituent services from the overlapping service domains e.g. union and intersection of two sets of ambient services.
These operators were defined on crisp and accurate service domain whose boundaries are sharp and clear but uncertain service domains are more sophisticated than classic one because of their boundaries. Consequently, operators which were defined by (Loke 2006) are not enough to decide which combination of services should be provided to a user located in several uncertain service domains. There are some other operators can be defined between two uncertain service domains, e.g. Rough and Fuzzy set theories can be considered as powerful frameworks which can consider uncertainty. It is possible to define different operators on two service domains defined based on fuzzy and rough set theory.

As it shown above, ambient services suffer from different aspect of uncertainty. Many theories have been proposed for solving different aspects of the uncertainty problems by addressing different facets. Rough set theory, introduced by Pawlak (1991), represents the uncertainty of an event by the approximation of sets using a collection of sets. It is widely adopted in the field of data mining since it is powerful for reasoning based on incomplete information. Rough set theory is one the most powerful device to deal with uncertainty while, in comparison with other theories, it is very simple to implement. Simplicity of rough set theory can be very helpful in ambient services because time and performance of the service is the most important criteria in mobile services. Most of the time ambient services are provided to moving uses whose position and service domain change over the time. Consequently, it is very important to calculate their service domains and provide services as fast as possible to be relevant to current request of users. Rough set theory is one of the simplest frameworks with less complication, so it can be a good candidate. Also many rough inferences have been developed and enable us to use them to deduce which area should be considered as service domain. These rough inference engines have got a higher performance in comparison with fuzzy inference engines and other inference engines especially fuzzy inference engines e.g. Mamdani and Sugeno. Another candidate of uncertainty handling device is Fuzzy set theory (Zadeh 1975). Fuzzy set theory is one of the most powerful devices which has been developed from many aspects, e.g. there are many different fuzzy inference engines or many software and systems are working based on fuzzy set theory. This paper is focused on uncertainty handling in ambient services based on fuzzy and rough set theory.

2 Methodology and Implementation

In order to Model ambient service domain based on rough set theory, it is very important to consider lower bound region as “in” the service domain, boundary region as “around” and outside region as “out” of the service domain. In addition if a service domain is supposed to be modeled based on fuzzy set theory, a membership function should be defined for service domain to show up to which extent each point belongs to a service domain. This part is focused on rough set-based service domain.

In order to find "in" and “around of the service domain”, some criteria, including distance, line of sight of the viewer and finally topological relationships between features around the service domain, were considered. Based on these criteria and using a rough inference
engine, boundary area or “around” of the service domain was determined. After defining three areas of lower bound, upper bound and outside sets of all the criteria, a rough inference system is developed to infer output area of “around the service domain”. Some of the rules defined in rough inference system are: If a feature is very near and completely visible, temporally connected and completely contained by a place, then it is located “in” the service domain (Bassiri 2012).

In order to test the developed inference system, developed software was installed on all devices of users. Users were supposed to click on one of the features shown on their cell phones' screen. Once they clicked, our service provider generated three areas called “in”, “around” and “out” of that feature, sent them back and also stored those areas in spatial database for further usage using DAC (Data Access Component Figure 1, in below, illustrates interaction between user device and service provider and displays components of system.

![Figure 1 Interaction between user device and service provider](image)

As it is illustrated in below, an area which has been patterned was selected by a user. Inference engine found three areas of "in", "around" and "outside" corresponding to the selected feature.

![Figure 2 A place which is selected by a user is shown in pattern and three corresponding areas of "in", "around" and "outside" deduced by inference system](image)
users were supposed to comment on these three areas after "in", "around" and "out" areas of the clicked feature displayed on the mobile screen. They fulfilled a questionnaire to express whether they think in the same way. In these questionnaires, users were asked which features should have been considered as features “in”, “around” and “out” of the selected feature from their point of view and to what extent result were close to their conceptions. They were supposed to assign a score in the range of 0 to 100 to express to what extent their perceptions were compatible with results. They also were supposed to draw two polygons presenting the area of “in” and “around” on the map based on their conception. Results of comparing these new polygons drawn by users and inferred polygons shows successfucness of developed rough inference engine.

Knowing “in”, “around” and “out” of the service domain makes it easy to provide relevant advertisement to a user located in one of these areas. In this step, based on user’s location, the application finds in which of area, “in”, "around" or "out", user is located the most appropriate service is provided. In order to find appropriate product to advertise, non spatial contexts are used (Bassiri 2012).

At last, we conducted a survey to find out if the service was successful or not. In order to evaluate, users were supposed to compare our service with a similar ambient service. As the difference between rough set-based ambient services and the traditional ones is model of service domain, users were supposed to compare rough set-based ambient services and classic set based ambient services. We provide both services to the same users, who were in different ages and had different level of education, and wanted them to comment on. Users were distributed in all around a specific place. Classic set based ambient services were provided. After receiving these messages on the screen of their cell phones, rough set-based ambient services were delivered. Then we wanted users to evaluate and compare these services with each other.

As it is obvious, users who were located "in" and "out" the service domain could not understand that much difference between two classic and rough-based ambient services provided to them Users because they either receive same message or they receive no message. On the other hand, users who were located “around” the service domain could understand difference between two sets of services. Because in classic set-based services, these users did not receive any message but in rough based services they receive some information. In general in classic services users are categorized into two groups “users in the service domain”, who are supposed to receive services, and “users out of the service domain” who cannot receive services. On the other hand, rough set based services categorized user into three groups; users located in the service domain, users around the service domain and users out of the service domain. First and second groups receive information but in different level of detail. Third group cannot get any message.
References


Towards Understanding The Factors of Integrating Spatial Data Infrastructures and Crowdsourced Geographic Data

Mark Iliffe
University of Nottingham, Nottingham Geospatial Institute / Horizon Doctoral Training Centre, United Kingdom
E-mail: psxmi@nottingham.ac.uk

Abstract
Currently within the geospatial world the two strands of research into Volunteered Geographic Information and Spatial Data Infrastructures are by and large distinct strands, joining rarely. This paper identifies a future area for research into the gap in-between these topics setting an agenda for future research.

Keywords: Spatial Data Quality, Crowdsourcing, Spatial Data Infrastructures, Community Mapping.

1 Introduction
This paper aims to review the current state of affairs within the research strands of Spatial Data Infrastructures (SDI) and Volunteered Geographic Information (VGI). Then positioning a future agenda for understanding the factors which will need to be researched for these strands to be tied together, and why the nature of these artefacts currently causes these research topics to be apart.

2 Spatial Data Infrastructures
Groot and McLaughlin (2000) defines a Spatial Data Infrastructure as being “the networked geospatial databases and data handling facilities, the complex of institutional, organisational, technological, human and economic resources...facilitating the sharing, access to, and responsible for the use of geospatial data and an affordable cost for a specific application domain or enterprise”. As such SDIs have evolved as part of a drive to facilitate the sharing and usage of spatial data this is identified as a key part of the global digital economy; United States have a National SDI policy derived from an executive order from President Bill Clinton - U S Federal Register (1994). Within the European Union the SDI is mandated by the INSPIRE directive - European Commission, (2007) - designed to help standardise data structures and facilitate data interoperability between European nations.

Coleman and McLaughlin (1997) overview the history and development of SDIs, linking them to the development and ubiquity of computing, especially the development of the Internet and ability of computing. They provide analysis of the early development of SDIs and NSDIs in particular setting principles for the development of NSDIs; Specifically they should be
widely available, easy to use, flexible and act as a foundation for other activities. They also present the components which constitute an SDI;

- **Sources of spatial data** - The groups and individuals acquiring, creating, and managing spatial data and databases in the private sector and at all levels of government.

- **Databases and metadata** - Sets of organised spatial data and information about those sets (e.g., directories), such as where it is located, how it was collected and is maintained, by whom, how it can be accessed, and what the characteristics are (e.g., scale, coverage).

- **Data Network** - The communication highways in various forms (e.g., telephone lines, local area networks, and new broadband integrated services networks or B-ISDNs) linking databases, sources, and users.

- **Technology** - The data conduits, but also the intelligent connections at each end, optimising the management of the databases at the source and maximising the potential application of the data by the user.

- **Institutional Arrangement** - The coordination of the many organisations involved in the NSDI development and maintenance.

- **Policies and standards** - The data communication rules, common conventions, and protocols, in addition to the broader and very critical policies addressing social and economic issues such as privacy and pricing.

- **Users** - The data communication rules, common conventions, and protocols, in addition to the broader and very critical policies addressing social and economic issues such as privacy and pricing.

The SDI is composed not just of these factors and systems but also encapsulates the interactions between systems; in effect an SDI delineates this system of systems. This list is not the only constraints on an SDI with Masser (2005) noting that “SDIs cannot be realised without coordinated action on the part of governments”. Williamson, (2003) makes a further case for the importance of SDI in the concept of the “Triple Bottom Line” namely the Economic, Social and Environmental components.

Because of these definitions and how SDIs are used currently throughout the world, they broadly are organised with a top down hierarchy with directives and oversight. Within this stack, the spatial data is generally collected/shared by an National Mapping Agencies (NMA). By their very nature these NMAs collect data based on their countries specific legislation; be it topological, thematic or cadastral. The process of data collection, regardless of its actual process and methodology will go through quality assurance and various processing stages to the point where it is released with an effective stamp of quality from the issuing NMA. This data would then feed into the SDI, be it at a national, international or more local level.
3 Volunteered Geographic Information

Volunteered Geographic Information is a relatively new term coined by Goodchild (2007) about the rise of citizen generated geographic content and citizen science. Even though this phenomena is a relatively new field a small but well formed body of research exists around it. Goodchild (2007) surveys the rise of this field placing it within other citizen driven projects and content like Wikipedia and blogging, expanding into other services. Location is a component in many services and the derived data from applications like Foursquare (Personal Location), Flickr (Geo-referenced Photographs), while this data is important, projects like Open Street Map - “Like Wikipedia For Maps”. Due to the openness of VGI datasets anyone can contribute therefore questions revolve around quality and quality assurance the dataset, exist.

Haklay et al. (2010) when testing Linus’ Law\textsuperscript{13} - in four areas of London, comments on the large variation of completeness between an authoritative dataset and VGI dataset; “This range of values is not surprising, because the information is provided by many participants, who are acting independently and with loose coordination” Haklay et al. (2010):316. The model within community mapping involves collocated, synchronous work to produce VGI therefore jars against this analysis.

Within the current literature of Volunteered Geographic Information (VGI), quality studies and assessments are generally performed as ground-truth comparisons using the buffer method described by Goodchild and Hunter (1997); Haklay et al. (2010), Zielstra and Zipf (2010), Girres and Touya, 2010. This involves comparative measurement of a given set of characteristics from VGI against some accepted and trusted ground-truth dataset. Girres and Touya (2010) and Mooney et al. (2011) further illustrate metrics that could be used to assess the historical quality of a VGI dataset, without comparison with a ground truthed dataset. Girres and Touya (2010) also provides an analysis of the attribute, logical consistency, semantics, temporal and usage qualities.

The factors which are tested in these quality assessments are by and large typified from van Oort (2006) namely Lineage, Positional Accuracy, Attribute Accuracy, Logical Consistency, Completeness, Semantic Accuracy, Usage Purpose and Constraints, Temporal Quality and Meta data quality.

In typifying these factors it opens up the opportunity for quality assessment of VGI datasets where a ground truth is not feasible or affordable. Kovacic and Iliffe (2010) illustrate this situation where a VGI dataset is the only data available from the Kiberan Slum, Nairobi, Kenya. This is part of a growing sphere of humanitarian mapping efforts, kickstarted by the Haitian Earthquake disaster response, Zook et al. (2010), recently brought as a subset of the VGI banner by Goodchild and Glennon (2010) - coining this wave of humanitarian mapping as Time-Critical Volunteered Geographic Information (TCVGI).

\textsuperscript{13} Given enough eyeballs, all bugs are shallow
4 Discussion

Within the literature and previous work there is scant understanding of how data from non-authoritative sources, for instance using crowd sourcing, could become authoritative or if it needs to. The nature of this data is such that it comes without large scale oversight, potentially from within communities, Perkins (2007), in effect collecting and presenting the data upwards using constructs like OpenStreetMap. This is in contrast to the top down methodology espoused by a SDI.

The direction of research is moving towards a more dynamic model for SDIs with Kobben et al. (2010) describing the ‘SDIlight’ concept as a way of teaching and exploring the theory and practice of SDI. They note that a SDI “usually denotes large, complex systems... it is of particular interest to our students that come from developing countries” (p.25). This is particularly important as students coming from development countries (most without an National Spatial Data Infrastructure and potentially a National Mapping Agency Masser (2005)). Therefore these students (and others) need to be familiar with techniques and methodologies of working with an SDI which due to economics and other concerns associated with the ‘triple bottom line’ are either deficient or missing. In turn allows for the potential of a crowd sourced VGI dataset which is cheaper than traditional, NMA espoused methods of data collection.

Continuing within this thread in setting a research agenda the nature of map data (thematic, cadastral and topographic would need to be considered). While the factors of spatial quality have already been discussed there is currently no literature linking spatial quality with the practice of surveying VGI data.

Within this the assumption that the data within SDI is from an authoritative source is made. The top down nature of SDI jars against the bottom up nature of VGI. Hence posing the following questions aims to bridge this gap, understanding the relationships in Fig. 1:

![Figure 1 Relationships between Quality, Practice and Requirements](image)
• What are requirements of countries which may not have legislation and policy around SDI and NMAs?
• What factors need to be considered to make the data from created from crowdsourced mapping authoritative?
• What is the relationship between authoritative and quality. Is ‘good quality’ equivalent to ‘authoritativeness’?

5 Conclusions
This paper reviews the current state of the literature within the topics of Volunteered Geographic Information and Spatial Data Infrastructures, illustrating a future research direction in exploring the space between these areas.

At first, while these areas may seem separated and abstracted, upon closer inspection this is a disingenuous statement. Though they come from different directions there is a clear agenda for research in combining the two.

References


I Williamson. SDIs: Setting The Scene. In I Williamson, A Rajabifard, and ME Feeney, editors, Developing Spatial Data Infrastructures: From Concept To Reality, chapter 1, pages 1–16. Taylor & Francis, 2003


Background

Soil science has conventionally focused on static soil properties with the intention of supplying secure and permanent basis for soil taxonomy (Pennock et al, 2006). In the past twenty years information technology has boomed and made available great computational capacity, powerful Geographic Information System’s tools (GIS), remote and proximal sensors and vast amounts of data such as Digital Elevation Models (DEM) indicating new ways ahead (McBratney et al, 2003). Scientists in the 21st century are now able to study and describe soils as dynamic entities in an interconnected landscape context.

Soil is an intricate system of interrelated physical, chemical and biological factors. In order to gain a better understanding of this complexity, scientists have utilized mathematical and statistical models to its quantification. This has created a movement of scientists “pedometricians” (Figure 1) exploiting knowledge from other scientific fields and developing quantitative techniques to predict soil properties from landscape attributes (Minasny et al, 2008).

This way of blending field observation methods with statistical spatial prediction techniques has created a new branch of research in soil science called digital soil mapping (DSM). DSM enables us to predict soil classes or specific soil properties (organic carbon, texture, bulk...
density, etc.) in areas with no information by spatially extending point observations of individual soil properties using mathematical and statistical techniques as well as indicating the uncertainty of such predictions. After more than 10 years of intensive research and applications DSM has emerged as a reliable alternative to traditional soil mapping (Carre et al, 2007). Despite its short history, the development of DSM has consisted in a series of issues that became apparent from the research work carried out and had to be tackled as soil scientist want to expand the scope and prediction power of their modeling. One of the concerns, since the foundation of this technique, that remains still unsolved is the issue of scale (Addiscott, 1998). Thomson et al (2001) have also suggested that this will become increasingly important with the fast development and implementation of regional soil-landscape models. The choice of scale frames the analysis and shapes the end result of our DSM analysis indicating that a better attention and quantitative knowledge of it will help to improve soil predictions.

Monitoring of natural resources is important to assess the effects of policy changes and to determine modifications due to climate and land use change. Information from soil surveys is then essential for successful natural resources management. Our ability to map soils depends on understanding the effect of the predominant natural and anthropogenic processes on soil formation. The objective of this project is to understand the effect that these processes have at different scales on our ability to map soils and to provide a guideline for the selection of the best covariates’ resolution enabling the change of support (up-scaling or down-scaling) of soil information to global, national and regional scale. This will permit to improve digital soil mapping ability to provide reliable soil data urgently demanded by the scientific community, practitioners and policymakers to address the global environmental issues that are threatening the planet.

1 Introduction

As stated by Levin (1992): “*scale represents the window of perception, the filter or the measuring tool through which a landscape may be viewed or perceived*”. The environment cannot be studied, modeled or visualized in its full complexity and details. Scale is then important because of its role on features selection and information’s generalization, essentially simplification.

Scale is a complex concept to grasp with many different and often divergent meanings; it is also highly dependent on the context of study (Marceau et al, 1999). The word scale may incorporate spatial, temporal or spatio-temporal aspects (Figure 2).
In soil science scale primarily involves space and this will be the focus of this research. A wide range of comprehensive reviews of the significance of spatial scale in geography and environmental science can be found in Lam (1992), Blöschl (1995) and Goodchild (2001) defining four spatial meanings of scale:

1. Representative fraction or cartographic scale (the ratio used to scale a feature).
2. Spatial extent or geographic scale (the extent or scope of a study).
3. Process scale or operational scale (the extent at which a phenomenon operates).
4. Spatial detail or spatial resolution (the shortest distance over which change is noted).

In GIS models the level of geographic detail cannot be represented in cartographic terms, for example the representative fraction loses significance because there is no set distance in the model to compare to the real world (Goodchild et al., 1997). In this context spatial resolution better encapsulates scale representing the level of spatial detail or the size of the smallest element in the dataset (in raster datasets pixel size).

The typical effects of spatial scale are:

- accuracy and precision in both data and modelling;
- the way to collapse and aggregate data in order to make them workable and relevant to the problem investigated;
- the process by which to extract measures of variation and correlation, enabling to make sense of the phenomena in question and to establish theories;
- the approach to use to communicate science through visualization.

In this paper scale will be regarded as the physical dimension of a phenomenon or process in space expressed in spatial units (pixel and neighbourhood sizes).
2 Scale in DSM

Spatial heterogeneity of soil parameters has proved to be a major problem in the representation of soil properties (Sinowsky et al., 1999) especially when spatial variability differs significantly between the scale of observations and the one in which processes are active. In the context of DSM studies, the scale diagram (Figure 3) shows the hierarchical levels of scale in soil science, with the pedon as the base unit and defines other levels as a reference to it.

Figure 3 Relationship between the level of soil units, scale, grid size, classification levels (Keys to Soil Taxonomy and World Reference Base) and auxiliary data (Hengl, 2003)

McBratney et al. (2003) reviewed pedometric methods for soil prediction in literature and suggested three main resolutions of interest (Table 1), which are:

- <20m (local extent);
- 20m–2km (catchment to landscape extent);
- >2km (national to global extent).
In DSM the determination of an optimal grid size for environmental factors to use in soil prediction is still an unsolved issue with only few empirical guidelines available. The success and accuracy of the prediction of soil attributes is highly dependent on finding the most suitable DEM resolution from which terrain attributes and the spatial resolution of the other pedogenetic covariates are derived. This problem is connected with the determination of the spatial scale of environmental phenomena or processes involved in soil formation. This is a critical problem because different pedogenetic laws and landscape processes operate at distinctive spatial scales (Florinsky et al, 2001) and thus analysis based on data of one scale may not apply to another scale.

The two main problems with DEM pixel sizes are that:

- at finer resolutions terrain variables hold an excess of details generating too much “noise” that inevitably invalidate the accuracy of the prediction;

- at coarser resolutions terrain variables behave irregularly and lose their predictive capacity.

In statistical terms observations made at a fine scale contain more variance than observations at a very coarse one. The greater assortment of observed processes and relationships at a detailed scale gives obviously more information but also more noise oppositely coarser scales can be too simplified and carry insufficient data (Figure 4).

<table>
<thead>
<tr>
<th>Pixel size and spacing</th>
<th>Cartographic scale</th>
<th>Resolution 'loi du quart'</th>
<th>Nominal spatial resolution</th>
<th>Extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;(5 x 5) m</td>
<td>&gt;1:5000</td>
<td>&lt;(25 x 25) m</td>
<td>&lt;(10 x 10) m</td>
<td>&lt;(50 x 50) km</td>
</tr>
<tr>
<td>(5 x 5) to 1:5000-</td>
<td>(25 x 25) to</td>
<td>(10 x 10) to</td>
<td>(500 x 500) to</td>
<td></td>
</tr>
<tr>
<td>(20 x 20) m</td>
<td>1:20,000</td>
<td>(100 x 100) m</td>
<td>(40 x 40) m</td>
<td>(200 x 200) km</td>
</tr>
<tr>
<td>(20 x 20) to 1:20,000-</td>
<td>(100 x 100) to</td>
<td>(40 x 40) to</td>
<td>(2 x 2) to</td>
<td></td>
</tr>
<tr>
<td>(200 x 200) m</td>
<td>1:200,000</td>
<td>(1 x 1) km</td>
<td>(400 x 400) m</td>
<td>(2000 x 2000) km</td>
</tr>
<tr>
<td>(200 x 200) to 1:200,000</td>
<td>(1 x 1) to</td>
<td>(400 x 400) to</td>
<td>(2 x 2) km</td>
<td></td>
</tr>
<tr>
<td>(2 x 2) km</td>
<td>1:2,00,000</td>
<td>(10 x 10) km</td>
<td>(4 x 4) km</td>
<td>(20,000 x 20,000) km</td>
</tr>
<tr>
<td>&gt;2 x 2 km</td>
<td>&lt;1:2,00,000</td>
<td>&gt;(10 x 10) km</td>
<td>&gt;4 x 4 km</td>
<td>&gt;200 x 200 km</td>
</tr>
</tbody>
</table>

Table 1 Suggested resolutions and extents of digital soil maps (from McBratney et al, 2003)
The level of detail represented by a raster dataset depends on the pixel size so that the cell is small enough to capture the necessary information without undermining the analysis. Fine resolution data have a smaller pixel size and a higher feature spatial accuracy that affect the processing time of models and require large storage capacity. The increase availability of high resolution DEM (1m LIDAR for example) has the consequence of increasing data storage and processing time exponentially and often confusing the modeller. A common approach is in fact to use the finest resolution available of DEM believing that this will improve the accuracy and precision of the prediction while on the contrary is increasing only the “noise”. For example, it is likely that a lowland area could be represented effectively with a DEM of 100m resolution or more, whereas an upland or hill areas would probably need a much finer grid size (Pain, 2005). Furthermore, Thompson et al (2001) have shown that higher-resolution DEM may not be necessary for generating useful soil-landscape models. Another concern is that most applications use algorithms running in small neighbourhoods (usually 3 X 3 rowing window) to perform terrain analysis, thus fixing the scale of resulting layers to the spatial resolution of the available DEM. This is expected to provoke mismatches between scale domains of terrain information and the environmental variable of interest (Smith et al, 2006; Dragut et al, 2009).

### 3 Land Surface Parameters

Covariates, the environmental factors recognized as governing soil formation, vary at different scales and this spatial variation at some scales may be more strongly correlated with soil than at others (Lark, 2007).
The relationship between soil properties and landscape attributes has been confirmed as a central concept in soil science (Hudson 1992). Terrain attributes are the most widely used predictors in DSM because of their primary role in soil formation and the broad availability of DEM (Behrens et al, 2010). DEMs are representations of the endlessly varying topographic surface of the Earth, and they are a widespread data source for terrain analysis and other spatial applications. Terrain analysis provide a number of high-resolution environmental information quantitatively derived from DEM including slope, aspect, plan curvature, etc. those topographic features are the core for a wide range of landscape-scale environmental models (Gallant et al, 1997).

The most suitable resolution of DEM to apply in soil prediction depends on the scale of the processes controlling pedogenesis and this is landscape dependent. As a consequence there is a real need for a more landscape-scale pedology that could offer the connection between soil processes involved in soil formation and soil surveys, creating the foundations for up scaling soil information to regional, national and global scale (Pennock et al, 2006).

However, the predictive power of any soil-landscape model that employs data derived from DEM is clearly dependent on their quality and scale (Figure 5). If these generalizations are within the scale threshold of the landscape processes that are operating in the environment under study there are no problems but if they are greater or finer than the spatial resolution of those processes, any result derived must be treated with caution.

![Figure 5](https://example.com/figure5.png)

**Figure 5** Schematic representation of surface roughness changing with pixel and neighbourhood sizes (from Grohmann et al, 2010)
4 Methodologies

The procedure for optimizing DSM prediction power of soil data with environmental covariates will focus on the selection of an optimum scale, represented as the interaction between pixel and neighbourhood sizes, correlated to the pedogenetic processes active at the landscape scale. The statistical relationship between soil series or properties and terrain derivatives will be used to select the scale at which terrain parameters correlate better with soil data and predict the most accurate soil information.

Previous studies (Smith et al, 2006; Roecker et al, 2008) have shown the influence of pixel size on DSM analysis. In this study a standard 20m resolution DEM was used to investigate the scale dependency of terrain attributes when up-scaled to coarser resolutions. A series of DEMs (Figure 6) were created from a 20m pixel size DEM by nearest neighbour resampling to 30, 40, 50, 60, 80, 100, 120, 140, 170, 200, 230, 260m pixel sizes in different window sizes from the standard 3 x 3 to 5 x 5, 7 x 7, 9 x 9, 11 x 11, 13 x 13, 15 x 15, 17 x 17, 19 x 19 and 21 x 21.

Figure 6 Original DEM at 20m re-sampled at 50, 100 and 500m pixel sizes and with a 3 x 3, 5 x 5 and 21 x 21 neighbourhood sizes.

All the different combinations of pixel and neighbourhood sizes were used to generate commonly used terrain parameters (slope, aspect, curvature, plan curvature, profile curvature, slope height, valley depth, normalized height, standardized height, mid-slope position and convergence index) from which 4 points per km² were randomly extracted and feed into data mining inference systems (both Artificial Neural Network and Random Forest) to find an optimum scale; starting from the finest resolution of DEM available in a 3x3 window size and incrementally increasing pixel and neighbourhood sizes. A quantitative comparison have been made calculating the validation performance of the two data miner
classifiers. This will enable to better understand the role of scale and experimentally select an optimum scale to use as a benchmark in the second and third stages of the research.

Based on the results of the empirical approach statistical diagnostic techniques will be tested and verified allowing, in the final stage of research, the development of a multi-scale methodology. A broad spectrum of techniques has been proposed and developed to overcome the issue of scale; the most promising ones (wavelet and geostatistical analysis) will be tested and verified against the experimental benchmark set in the first stage of research.

Wavelet analysis mathematically approximates a data series by a linear combination of functions with specific scales (resolutions) and locations. The scale-dependence of the correlations of soil properties is investigated by computing wavelet correlations for each scale of analysis (Lark et al., 2001). Gallant et al. (1997) also suggested that the positive wavelet decomposition is a powerful tool to study the effect of scale on terrain attributes avoiding the problem of introducing artifacts due to changes in grid resolution.

Geostatistics offers statistical tools for analyzing and predicting the structure of spatial data (Goovaerts, 1999). The variogram is a central tool of geostatistics, it can describe the scales of spatial variation and through regularization can be used to predict changes in the variable according to changes in the scale of measurements. (Atkinson et al., 2000). Lark (2005) has investigated scale-dependence in correlation to soil properties by using a nested sampling technique coupled with multivariate geostatistics.

The diagnostic techniques currently tested will be utilized to identify optimum resolutions for each environmental covariates; in essence a multi-scale approach to DSM. Multi-scale modelling is therefore a representation of the interactions among covariates, with different spatial scales of activity, at their own best spatial resolution. As Behrens et al. (2010) suggest, multi-scale approaches have not received the attention they deserve in soil science despite the fact they could offer the opportunity to both increase knowledge of soil formation and to improve the accuracy of DSM.

The final step of this research will be the development of “scaling” guidelines to be used in the first steps of data analysis in DSM. Guidelines are useful tools for making DSM more consistent and efficient and for closing the gap between what practitioners do and what scientific evidence support. This will take the form of statements or decision diagrams to assist practitioners in the selection of the optimum scale or scale thresholds to be used in the prediction of soil properties with DSM techniques.
5 Preliminary results

DEM is fundamental in assessing soil variability and regularly used in DSM analysis as a core parameter and scale benchmarks for all the other environmental covariates to predict patterns of soil properties. Eleven common attributes of topography – slope, aspect, curvature, plan curvature, profile curvature, slope height, valley depth, normalized height, standardized height, mid-slope position and convergence index – have been computed using SAGA GIS terrain analysis tools.

In this research the effect of DEM resolution in DSM was investigated by deriving the terrain attributes with different combinations of pixel and window sizes and empirically looking for their validation performance with data miner classifiers. As a first result one should mention that as expected the three morphologically different study areas are sensitive to scale and behave in unique ways showing two main patterns of behaviour:

- Flat homogeneous areas seem to prefer coarser resolutions (above 140m in this study) across all the tested neighbourhood sizes;
- Morphologically varied areas, with characteristic features as the drumlins or abrupt changes in topography as steep slopes, seem to prefer fine resolutions (30m in this study) with small neighbourhood sizes but also show opposite behaviour with coarse resolutions and big window sizes.

Both areas are sensitive to pixel variation even though in opposite ways but the interaction between pixel and neighbourhood sizes provides the key for the scaling of morphologically rough areas in DSM analysis. Neighbourhood size has a significant role only in interaction with pixel size, influencing which terrain features can be identified and hence, the scale of analysis.

References


Ad-hoc-Geoprocessing in Spatial Data Infrastructures – Formalizations for Geooperators

Johannes Brauner

Technische Universität Dresden, Geoinformation Systems, Germany
E-mail: johannes.brauner@tu-dresden.de

Abstract
Increased technical possibilities and availability of geodata raises demands for more sophisticated geoprocessing in Spatial Data Infrastructures. For this, syntactic and semantic interoperability are essential. Whereas syntactic interoperability is widely achieved, semantic interoperability is still missing. To support semantic interoperability in this context, three research challenges are described in this paper: (1) Definition of sets of elementary geooperators, (2) the development of formalized and semantically enabled descriptions of these geooperators and (3) an approach to create and validate workflows built from these geooperators.

As this paper is an extended position paper for a PhD thesis project, it represents the general work plan for the thesis and omits detailed results.

Keywords: Geoprocessing, geooperators, SDI, semantics, interoperability, workflows

1 Introduction
So far, Spatial Data Infrastructures (SDI) are mostly limited to provide mere visualization of and access to geospatially related data (geodata). On the other hand, according to the Next Generation Digital Earth position paper, as more and more geodata is available, the need for sophisticated processing and analysis models which create insight and support intelligent actions increases (Craglia et al. 2008). This development is also supported by increasing network bandwidth and computational power, facilitating further the usage of geodata in SDI (e.g. Foerster et al. 2011). For instance, conduction of geospatial analysis tasks to derive geoinformation from geodata via web-based geoprocessing technology is becoming more and more important.

The term geoprocessing stands for the processing of geodata. Using web-based geoprocessing technology allows for unsupervised, distributed and provider independent geospatial analysis. Additionally, it allows for the ad-hoc, on-the-fly and flexible task specific orchestration of geoprocessing workflows. (Foerster et al. 2011)

Such workflows are created by applying geooperators to geodata. Geooperators represent well-defined functionality for elementary geoprocessing tasks.
Various sources for geoprocessing are already available in SDI. However, existing solutions lack a clear basic definition and formalization of geooperators, including their relations and possible combinations. To enable the (machine supported) usage of geooperators, syntactic and semantic interoperability are required. Whereas syntactic interoperability is already widely achieved by applying matured standards (e.g. ISO 19119 and OGC Web Processing Services (WPS; Schut (2007)), semantic interoperability is still lacking. (Brauner et al. 2009; Foerster et al. 2011)

Semantic interoperability supports and is required for finding suitable geooperators (to ensure the proposed functionality), for the exchangeability of operators (for the comparability of results and to maintain provider independence) and for the flexible ad-hoc and (semi-) automatic creation and validation of meaningful workflows. Semantic interoperability is a required step leading towards semantic integration as defined by Kuhn (2005).

The main goal of the presented work is to support the achievement of semantic interoperability for geooperators in SDI in the three ways described above (find geooperators, exchange geooperators and create/validate workflows). According to Craglia et al. (2008), this approach tackles four of six current limitations of geoprocessing services: Missing or inadequate descriptions containing information about processes in general, relevance, limitations and expected outcome. Additionally, Craglia et al. (2008) recommend research for "universal elements and [a] language for dynamic modeling", in other words geooperators and an algebra supporting the space-time analysis and modeling by workflows. To enable their usage in SDI, research for "intelligent descriptions (automatic, user driven) of data, services, processes, models, searching and filtering" (Craglia et al. 2008) leading to better fitness-for-use assessment is required. Kuhn (2005) outlines two problem classes relevant for the presented work: (1) "Service Discovery and Evaluation" and (2) "Service Composition", which leads back to the three ways to support the achievement of semantic interoperability, as for service discovery (finding and exchanging geooperators) meaningful descriptions are required and service composition enables the chaining of Web Services to create workflows.

The presented work analyzes existing geoprocessing functionality to derive common geooperators and describes a methodology to create and expose these formalizations in SDI. Additionally, a formal language (e.g. algebra) is developed to enable a (semi-) automatic chaining and validation of geoprocesses.

To enable machines to deal with meaning and semantics and to ensure interoperability, formalizations are mandatory. Thus, structured unambiguously defined operator descriptions are necessary to be able to automatically create task specific geoprocessing workflows and to ensure the exchangeability of geoprocessing functionality from different providers. Standards and concepts from the Semantic Web (Berners-Lee et al. 2001) on the one hand and SDI on the other hand play a major role to achieve this goal. Their integration into each other provides a challenge (Janowicz et al. 2010). Nevertheless, spatial concepts
are of high importance in the Semantic Web (Berners-Lee et al. 2001). Egenhofer (2002), Janowicz et al. (2010) and Kuhn (2005) primarily focus on the semantics of data. Nevertheless Egenhofer (2002) also states that "the semantics of geospatial comparators is captured in a geospatial-relation ontology" which is not further defined. Janowicz et al. (2010) and Kuhn (2005) motivate to better understand the meaning of processes by comparing the output of a process to the semantics of the input. Taking these different findings into consideration is essential for the achievement of semantic interoperability in SDI.

As this paper is an extended position paper for a PhD thesis project, it represents the general work plan for the thesis and omits detailed results. Section 2 describes the approach to define a set of common geooperators, leading to a methodology to formalize descriptions of geooperators (Section 3). Based on the geooperators and its descriptions, ideas to define an algebra to support the ad-hoc generation and evaluation of more complex geoprocessing workflows are outlined (Section 4).

2 Defining Elementary Geooperators

The first step in achieving automatic workflow creation is to accomplish comparability and exchangeability among existing geooperators (formalized by e.g. taxonomies or ontologies) by including and bringing together state-of-the-art analysis about existing scientific concepts (e.g. Tomlin’s Map Algebra or Egenhofer’s and Albrecht’s operators) and best practices (by analyzing geoprocessing capabilities in various GIS backends). This poses challenges in terms of developing distinctive features, taking operator granularity into account (atomic geooperators vs. workflows), measuring semantic similarity, considering the semantics of input and output data, defining requirements for meaningful workflows and defining requirements for their usage in SDI.

The expected result is a definition of sets of elementary geooperators, formalized e.g. as taxonomy or ontology as suggested by Kavouras & Kokla (2008). A taxonomy is regarded as an ontology without semantics, whereas an ontology is a more formalized approach to describe semantics of concepts (Kavouras & Kokla 2008). Lemmens (2008) defines an ontology and also four requirements from an SDI perspective that lead to the ontology: (1) Hierarchical structure (to enable humans to read it), (2) Non-overlapping classes (as much as possible), (3) include most important geo-operations (without further specifying what ‘important’ means) and (4) extensibility.

Often lightweight approaches to formalize semantics are proposed. Kavouras & Kokla (2008) describe the necessity to introduce lighter ontologies focusing on common sense rather than scientific expertise. They describe e.g. Egenhofer’s approach “towards the Geospatial Semantic Web” (Egenhofer 2002) as such a lightweight solution.

It can be assumed that there is not the one taxonomy or ontology, but many different views on geooperators not just from a high level research perspective but also taking into account
the community and user perspective and the way GIS vendors organize their functionality. Even in the standardization domain where one would expect a strong motivator for explicitly defined and standardized geooperators, discussions are ongoing if there is only one way and one view on geoprocessing functionality (see e.g. Foerster & Schaeffer (2010)). Nevertheless, standardization also requires semantically enabled and described geooperators (Kiehle & Foerster 2010). The same authors further suggest to use the Reference Model for Open Distributed Processing (RM-ODP) to track the different viewpoints of Web Processing Service (WPS) Profiles (Kiehle & Foerster 2010). As WPS Profiles are one means to formalize geooperators into categories, this suggestion is highly relevant for this work.

Beside these different views on geooperators coming from science, users, user communities and GIS providers, the granularity of geooperators play a major role. Three different granularities can be distinguished: (1) Atomic operators as either building block for more (2) sophisticated operators, which in turn together with atomic operators provide the building blocks for (3) geoprocessing workflows addressing complex geospatial problems. Granularity of geooperators is especially important in regard to their usage in SDI, where special constraints like the amount of data to be transported and Web Service response times are crucial. On the one hand, invoking a huge amount of atomic operators produces a lot of overhead in terms of data transportation. On the other hand, a calculation of a climate model in an SDI context has no reasonable response times. See Müller et al. (2010) for a more detailed analysis on geoprocessing in SDI.

The following list represents a non-conclusive selection of possible categorization criteria for geooperators including prominent examples, taking into account the different perspectives on geooperators. Categories can be built according to:

- Literature in GI Science (mainly GIS textbooks and papers):
  - Albrecht’s Universal Analytical GIS Operations (Albrecht 1998),
  - Tomlins Map Algebra (Tomlin 1990),
  - Mennis’ multi-dimensional Map Algebra (Mennis 2010),
  - The Egenhofer operators (e.g. Egenhofer (1993)),
  - Goodchild’s six basic classes of spatial analysis (Goodchild 1987),
  - Chrisman describes a transformational view on classifying GIS operations (Chrisman 1999),
  - Six elements of geoprocessing (Ding & Densham 1996),

- The underlying geodata model (raster vs. vector vs. raster2vector and vice versa),

- Comparing and listing I/O parameter,
• Requirements from user communities (e.g. health-environment analysis or hydrology),

• Standardization efforts:
  o The service taxonomy in ISO 19119 (ISO/CEN 2006),
  o The dimensionally extended nine-intersection model (DE-9IM) in ISO 19125 (Herring 2011).

• Status quo in existing GIS implementations (keywords, manual topics, naming conventions; e.g. ArcGIS toolboxes, GRASS modules),

• Taking a look at the result (semantics): Distinction between plain data transformation (e.g. data format transformation) where no new information is created and ‘real’ GIS analysis where usually new information is created?

• Applicability in an SDI context (atomic geooperators vs. climate model; see above),

• Technical / mathematical / application background:
  o E.g. in High Performance Computing: Is parallelization possible? (Ding & Densham 1996)

These many heterogeneous categorization criteria support the above given assumption that there is neither one view on nor one taxonomy of geooperators. To integrate these different views into sets of elementary geooperators, a faceted classification or controlled vocabulary can be used.

Further challenges arise from requirements to technically publish these sets of elementary geooperators in an SDI. One way is described by Janowicz et al. (2010). They suggest a Semantic Enablement Layer to connect the Semantic Web concepts (e.g. ontologies, reasoner) and SDI services by profiling existing OGC Web Service interface standards.

3 Formalizing Geooperator Descriptions

As a second step, formal operator descriptions (for finding suitable geooperators) are proposed, including state-of-the-art analysis about existing concepts from Semantic Web formalizations and technologies (e.g. SKOS - the Simple Knowledge Organization System). Considering and deriving requirements for meaningful workflows and measuring semantic similarities are challenging.

The expected result is a formalization of geooperators as a foundation for distributed geoprocessing in SDI.

So far, only one means to formally describe geoprocesses in an SDI context exists: WPS DescribeProcess documents (Schut 2007). They focus on syntactic details like the data type and data format for the input and output parameters. Semantics can only be captured by providing a block of natural language text which cannot be further interpreted by the WPS. Thus, semantics (if present) can only be understood by humans.
This gap has been addressed by Lutz: He presents an approach to provide descriptions based on ontologies in addition to syntactical descriptions, keeping both descriptions separate (Lutz 2007). Lemmens et al. also suggest an ontology based approach for the semantics (the deep service descriptions) but integrate both descriptions into each other (Lemmens et al. 2007).

A third approach originates from standardization approaches. The concept of semantic annotations for OGC Web Services (Maué 2009) also describes means to integrate semantics into existing syntactic Web Service and process descriptions. It represents a more light-weight approach compared to Lemmens et al. (2007) and is directly targeting SDIs. In combination with an SKOS (Simple Knowledge Organization System) vocabulary, which is based on Resource Descriptions Framework (RDF) statements rather than being backed by a full ontology, this seems to be a promising approach for the presented work, following the pragmatic and light weight approach motivated above.

To support the finding of suitable geooperators, concepts like faceted browsing are taken into account. This can be based on the faceted classification described above.

Another approach is the Web Ontology Service (WOS), a profile to the OGC Catalogue Service for the Web (CS-W) accessing the ontology which is the basis for the semantic descriptions (Janowicz et al. 2010). Klopfer & Simonis (2009) also describe the concept of a semantic catalogue which is nevertheless designed for users who already know what to find. The work for a semantic catalogue in the project ENVISIONS is ongoing (Larizgoitia & Toma 2011).

To compare different operator implementations and their descriptions is challenging. The concept of semantic similarity measurement as described by Rodríguez et al. (1999) and Schwering (2008) focuses on geodata sets. To enable the measurement of semantic similarity for geooperators further research is required.

4 Algebra for Geoprocessing Workflows

Last but not least, an algebra is required to support the ad-hoc orchestration and validation of geoprocessing workflows based on the defined and formalized geooperators, including state-of-the-art analysis about existing concepts (e.g. using ideas from formal languages, existing Application Programming Interfaces and Scientific Workflows). Special challenges in terms of taking the semantics of input and output data into account, defining semantic references systems for the algebra and finding validation means, will occur.

The expected results are formalized combinations of elementary geooperators to assemble task specific geoprocessing workflows on-the-fly.

In general the problem is two folded. On the one hand, one needs to define formalisms to represent / encode an algebra and on the other hand the algebra needs to be interpreted (Kavouras & Kokla 2008). This work is focusing on the description of an algebra for the ad-ho
workflow creation and its evaluation/validation. The interpretation of the algebra is of minor importance and more of a technical challenge.

So far, multiple approaches to build workflows from geoprocessing functionality exist (mostly) in Desktop GIS. ArcGIS uses the model builder to create workflows, ensuring syntactic compatibility of its tools and allowing to save the workflow as an own tool on the user’s machine. The Gearscape project (http://gearscape.fergonco.es/) uses the GGL (Gearscape Geoprocessing Language) to build workflows. PCRaster uses concepts from cartographic and dynamic modeling and defines the PCRaster environmental modeling language to build complex workflows. The Web Coverage Processing Service (WCPS) Language (Baumann 2009) defines a language for extraction, processing, and analysis of multi-dimensional raster data in SDI. All this previous work from the geoinformatics domain needs to be considered to develop the algebra.

In computer science, various means exist to formalize an algebra. Among others, prominent examples are the Extended Backus-Naur Form (EBNF) and functional programming languages like Gofer or Haskell. From a pragmatic perspective, the concept of Scientific Workflows applied in a geospatial context (Moodley et al. 2008) seems promising. Scientific Workflow tools (e.g. Vistrails) allow for building workflows out of buildings blocks from various sources, be it Desktop GIS, WPS or further external tools. They take care of the syntactical correctness, enable a full tracking of provenance / lineage information in each workflow step, can possibly link to ontologies (the framework can easily be extended) and thus seem to be a promising environment to host the algebra.

5 Conclusions

Today, the required syntactic interoperability for ad-hoc geoprocessing in Spatial Data Infrastructures is widely achieved. Nevertheless, semantic interoperability is still missing. The meaning behind geoprocessing functionality - based on geooperators – cannot yet be formalized to enable the ad-hoc creation and validation of geoprocessing workflows. To seamlessly use geooperators from different providers as the building block of more complex workflows, a common understanding of the meaning of geooperators is necessary. To enable machines to deal with these sets of elementary geooperators, formalized and semantically enriched descriptions are required. To achieve semantic interoperability the presented work (1) defines a set of common geooperators, (2) provides formalized geooperators descriptions and (3) develops an algebra to create and validate geoprocessing workflows.

It can yet be concluded that there is not the one categorization or taxonomy of geooperators. Especially the user community and GIS providers have many different views on geooperators. A pragmatic approach based on common sense is preferable rather than developing a complex and complicated ontology. The same holds true for the formalized process descriptions and the algebra. Using scientific workflow tools to create and validate
more complex geoprocessing workflows and exchange them is widely accepted in the research community and holds potential to be the means to support and enable ad-hoc geoprocessing not only in SDI.

References


Towards a CityGML based data model for the GRB, the 2D-large scale topographic inventory of Flanders

Ine De Cubber
Katholieke Hoogeschool Sint-Lieven, Gent, Belgium

Jos Van Orshoven
Katholieke Universiteit Leuven, Belgium

Abstract
Although international standards for data modeling are available, like CityGML, the large scale topographic inventory of Flanders (GRB) uses its own vernacular data models, resulting in the lack of consistency in the information flow through the modeling process and the absence of certain modeling actions at each abstraction level. In this paper, a schema matching will be done, to describe the similarities and dissimilarities between the data model of the GRB and CityGML, with a view to standardise the GRB data model, to make the information flow more consistent and to make the GRB data model international more accessible. By this, the potential of CityGML to store 2D data from existing LSTI, like the GRB in a more standard environment, will be proven.

Keywords: data modeling, CityGML, GRB, schema matching

1 Introduction

1.1 Data modelling
As described by De Cubber & Van Orshoven (2011) data modelling is about abstracting a portion of the real world. For each of the three abstraction levels, modelling actions can be defined, describing the non-spatial information (semantics): at conceptual level, object classes and high level relationships are defined; for the logical level, attributes and relationships, together with primary keys and operations; in the physical level, tables are created and stored in files or databases. For the spatial information (geometry), the actions at conceptual level are the view, geometry representation and high level primitives; at logical level, the spatial data structure; at physical level, spatial tables are created, and stored in spatial files or databases.

1.2 (International) Standards for data modelling
For non-spatial data modelling, approaches like UML (ISO/IEC 19501:2005) and ERM (Chen 1976) are frequently used, while for spatial data modelling, the ISO standards for geographic information, described by Kresse & Fadaie (2010) are used. Inspire (2012) took the initiative to create an infrastructure for spatial information in Europe. As most of the existing standard modelling approaches focus only on the semantics or on the geometry, the OGC CityGML data model, (The Open Geospatial Consortium – CityGML 2012) combines both the semantics and the geometry.
1.3 Local large scale topographic inventories and their data models

In Belgium, the large scale topographic inventories (LSTI) are organised by each of the political regions, in three unique databases: GRB (Flanders region), PICC (Walloon region) and Urbis (Brussels region). (De Cubber, Kips & Van Orshoven 2009). Other European initiatives of LSTI are the OS mastermap in the UK (Ordnance Survey 2012), and the GBKN, BGT and IMGEO (respectively the current LSTI, the future LSTI and the extended LSTI) in the Netherlands. (Geonovum 2012)

2 Problem statement and objectives

The starting point of this paper is that although international standards for data modelling are available, each of the LSTI has its own vernacular documentation, describing their own data models and modelling actions at conceptual, logical and physical level. For the LSTI in Flanders, the GRB documentation was examined in detail by (De Cubber & Van Orshoven 2011), proving the lack of consistency in the information flow through the GRB modelling process and the absence of certain modelling actions at each abstraction level.

Against this background, the objectives of this paper are Finding similarities and dissimilarities between the data model of the GRB and CityGML, with a view to standardise the GRB data model, and to make the information flow more consistent, and make the GRB data model international more accessible and roving the potential of CityGML to store 2D data from existing LSTI, like the GRB in a more standard environment.

3 Methodology

A linguistic, element-based schema matching will be used to determine a set of correspondences that identify similar elements in the different abstraction levels of the data models. (Rahm & Bernstein 2001)

4 Materials

The GRB is a database holding large scale/high resolution topographic inventory data in a specified data model and data format. The GRB data model contains a detailed description of the topography in 2D, with information on buildings and building infrastructures, parcel separations and borders, roads and road infrastructure, railways and water infrastructures, ...

(AGIV 2012)

![Figure 1 Extract from the GRB dataset](image-url)
(Kolbe, Gröger & Plümer 2005) describe CityGML as a multi-purpose and multi-scale representation for the storage of and interoperable access to 3D city models, based on the standard GML3 of the OGC and covering the geometrical, topological, and semantic aspects of 3D city models. CityGML combines both spatial and semantic properties, and similar object classes to the GRB are modelled, (like buildings and other man-made artifacts, vegetation objects, water bodies, and transportation facilities like streets and railways.

![Figure 2 Virtual 3D model of Ettenheim, using CityGML](Research Center Karlsruhe, 2012)

## 5 Results

At the conceptual level, the logical level and the physical level, the data models of the GRB on the one hand and the CityGML data model on the other hand, were compared, to find similarities and dissimilarities.

### 5.1 Conceptual level

At this level, the real world entities are translated into the object classes and the high level relationships are defined, together with the choice of view, the geometry and the high level primitives. (The sign ≅ should be read as “matches with”)

Certain entities are modelled identical within the GRB data model and CityGML, like the WaterSurface and Railways, resulting in identical object classes. Other entities are modelled in much more detailed within CityGML, like buildings, resulting in some identical object classes (GRB – Building ≅ CityGML – Building, GRB – façade line ≅ CityGML – WallSurface), and some extra object classes, giving the possibility to store information on roof surfaces, on openings like doors and windows, and even on building furniture and building installations (CityGML). Some entities are represented in less detail within CityGML, resulting in some identical object classes (GRB – parcels ≅ CityGML – Landuse) and some extra object classes, like the GRB – parcel reconstruction points, showing the border points of a parcels. A last group of entities are not represented within CityGML at all, like manholes and façade points in GRB or are not represented within GRB, like relief or vegetation in CityGML.
### Table 1 Examples of entities, modelled in GRB and CityGML, resulting in identical, more detailed, less detailed or unavailable conceptual object classes.

<table>
<thead>
<tr>
<th>Comparison of entities...</th>
<th>GRB - conceptual model object class</th>
<th>CityGML - conceptual model object class</th>
</tr>
</thead>
<tbody>
<tr>
<td>... with identical object classes:</td>
<td><img src="image1" alt="GRB Water" /></td>
<td><img src="image2" alt="CityGML WaterSurface" /></td>
</tr>
<tr>
<td>Water</td>
<td></td>
<td></td>
</tr>
<tr>
<td>... in more detail in CityGML:</td>
<td><img src="image3" alt="CityGML Building" /></td>
<td><img src="image4" alt="CityGML WallSurface" /></td>
</tr>
<tr>
<td>Building</td>
<td><img src="image5" alt="CityGML Opening" /></td>
<td><img src="image6" alt="CityGML RoofSurface" /></td>
</tr>
<tr>
<td>/ /</td>
<td></td>
<td></td>
</tr>
<tr>
<td>... in less detail in CityGML:</td>
<td><img src="image7" alt="CityGML Parcels" /></td>
<td><img src="image8" alt="CityGML LandUse" /></td>
</tr>
<tr>
<td>Parcels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>... not in CityGML:</td>
<td><img src="image9" alt="GRB Manholes" /></td>
<td><img src="image10" alt="GRB Façade points" /></td>
</tr>
<tr>
<td>Manholes, façade points</td>
<td></td>
<td></td>
</tr>
<tr>
<td>... not in GRB:</td>
<td><img src="image11" alt="GRB Relief" /></td>
<td><img src="image12" alt="GRB Vegetation" /></td>
</tr>
<tr>
<td>Relief, vegetation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Within the GRB conceptual model, only three high level relationships are defined: GRB - Façade line bounds GRB - building, GRB - Façade points bound GRB - façade line and GRB – building has GRB - façade points. Within CityGML, the same high level relationships are defined, but sometimes with different syntax: CityGML - Abstract building is bounded by CityGML - boundary surface As CityGML does not have an object class for the façade points, this relationships cannot be translated.

Within CityGML, other relationships are defined, like CityGML - AbstractBuilding consists of CityGML - BuildingPart. Although the GRB has also the object classes GRB-building and GRB-building part, this relationship is not foreseen at conceptual level.
Table 2 Comparison of the relationships at conceptual level

The choice of view, geometry and high level primitives are not explicit mentioned in the GRB data model. The spatial modelling actions at conceptual level for both CityGML and GRB are the same, namely: discrete objects are represented by vectors, with geometric primitives. As CityGML uses the GML3 standard for spatial modelling, the possibility exists to store also continuous fields (represented as geometry), and to store the topology in an explicit way.

5.2 Logical level

At this level, the attributes, relationships, primary keys and operations are added to the conceptual model, and the spatial data structure is chosen.

The comparison of all the attributes between GRB and CityGML, has proven that each object class has its own specific attributes. Some of these attributes from the GRB, can be stored within CityGML, others can’t.

The high level relationships, found at conceptual level, disappeared in the GRB logical model, and translated into the logical level of the CityGML data model. Each object class in the GRB data model has a primary key, IDN. Within CityGML, no primary keys are defined. No operations are defined in the GRB data model or the CityGML data model. (Table 3)
Comparison of GRB logical model | CityGML logical model
--- | ---
Attributes, primary keys, Relationships and operations

| Table 3 | extract of the GRB logical model, with attributes and primary keys for each object class, without relationships an operations (left) and extract of the CityGML logical model with attributes and relationships for each object class, without primary keys an operations (right) |

The GRB uses the Simple features data structure (ISO 19125-1:2004), for storing points, lines and polygons. CityGML uses the GML 3.1.1 spatial data structure (The Open Geospatial Consortium-GML 2012): which is also an ISO standard (ISO 19136:2007) and provides possibilities to store points, linestrings and polygons.

5.3 Physical level
At physical level, the tables and spatial tables are created, to store the data into spatial files or databases.

For the GRB, the geometric data is stored in shapefiles and additional tables are stored in DBF format. For CityGML, XML schema definitions are available, describing how a CityGML file should be stored, resulting in a text-file, which can be read by specific CityGML viewers, like LandXplorer CityGML viewer (Autodesk 2012)

As the physical models of the GRB and CityGML are totally different, the comparison at physical level was not done, to search for similarities or dissimilarities.
6 Discussion

6.1 Schema matching at conceptual, logical or physical level

Within a data modelling process, starting from the requirements, it is extremely interesting to start with a conceptual model, which can be fine-tuned in a next phase into a logical model, and transformed into databases or files, with the physical model. During the process, schema matching at each of the three abstraction levels can be done, to see if the designed model can also be used for modelling other data.

In the case of standardising the GRB data model, where the data modelling process for both the GRB and CityGML was already done, resulting in data models at the three abstraction levels, there is no real advantage in doing the schema matching at the three levels anymore. As the logical level is an extension of the conceptual level, a matching at conceptual level only resulted in a comparison of the object classes, ignoring the importance of the attributes.

The physical data model, on the other hand, was already chosen for both the GRB and CityGML. A fundamental comparison between the OGC CityGML format and ESRI shapefiles (ESRI 2012) could be done, but a schema matching at physical level in this specific case gives no added value for creating a standardised data model for the GRB, as the ESRI shapefile is already a de facto standard for vector spatial data.

6.2 Matching cardinalities

Certain object classes of the GRB cannot always be easily translated into one common object class within CityGML, but will have to be modelled in different object classes. (Rahm & Bernheim 2001) describe this as “the matching cardinality”, where one element of the first data model may match with more elements of the other data model and vice versa. (in his definition, an element should be replaced by “an object class”)

Within the GRB, different types of objects can be stored in one object class. In this example, the different types of the object class “civil construction” are shown in figure 3. When trying to match this object class with CityGML, some of the types will be stored within the object class CityGML-BuildingFurniture, (with specific BuildingFurnitureFunctionType) like tanks (type 9) and water mills (type 11), others will be stored as CityGML-CityFurniture, (with specific CityFurnitureFunctionType) like columns (type 5) and chimneys (type 7) and again others cannot be stored at all within CityGML, like breakwaters (type 22). (table 4).

For the matching of multiple object classes at a time, complex expressions will be needed to specify how the object classes and attributes are related. A simple textual element based matching, as done before, will not be satisfying for this situation.
**Figure 3** Use of different types of objects of the object class “Civil construction”, found in the textual documentation of the conceptual GRB

<table>
<thead>
<tr>
<th>Comparison of the...</th>
<th>GRB logical model</th>
<th>CityGML logical model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity “civil construction”</td>
<td><img src="image1.png" alt="GRB logical model diagram" /></td>
<td><img src="image2.png" alt="CityGML logical model diagram" /></td>
</tr>
</tbody>
</table>

**Table 4** Matching of one GRB object class “civil construction” with multiple CityGML object classes “CityFurniture” and “BuildingFurniture”

### 6.3 Object classes with type attributes versus parent-child object classes

In the GRB, the use of different types of objects within one object class is already introduced in the paragraph above. The information on the specific type of an object is always stored as an attribute (TPC) in the GRB data model. Within CityGML, different types are sometimes modelled as an attribute, like for CityFurniture and BuildingFurniture, but sometimes result in the creation of new object classes, childs of the parent, which inherit all the characteristics of the parent. In this example, the object class GRB-road, has an attribute TPC, indicating the difference between squares and roads. Within CityGML, the object class transportation complex, is parent of the object classes track, railway, square and road.
The difference in modelling of the road types, (object classes with type attributes versus parent-child object classes) is not necessarily a problem when matching the data models. The advantage of a parent-child structure is that each of the characteristics (attributes) of the parent is inherited by the child, but each child can have its own specific extra characteristics. In the example above, each object class “road” (child) has inherited the function and the usage of the object class “TransportationComplex” (parent), but can have extra attributes, like for example width and material. When using the “type” attribute, adding additional information to one specific type of object, becomes hard.

7 Conclusion and outlook

The comparison of the GRB data model and the CityGML data model had proven that the CityGML provides the possibility to remodel most of the GRB object classes with their attributes, which would make the GRB data model more standardised and international more accessible.

With a case study in which a GRB dataset is transformed into CityGML, based on the GRB-CityGML schema matching results of this paper, the schema matching could be tested and fine-tuned.

For those object classes and attributes of the GRB, that do not fit in the CityGML data model, an answer should be found on which information is lost (overview of different types of information loss), how much information is lost (quantification), if the information is essential for the GRB users (requirements analysis) and if CityGML can give a solution for storing this information (possibility to use ADE (application domain extensions) or generic objects and attributes).
References


De Cubber, I & Van Orshoven, J 2011, ‘Unstandardized terminology complicates the communication about 2D and 3D spatial data models’ in Proceedings of the IAMG conference 2011, Salzburg, Austria.


108
The Open Geospatial Consortium 2012, CityGML. Available from:
<http://www.opengeospatial.org/standards/citygml> [February 2012].

The Open Geospatial Consortium 2012, GML. Available from:
<http://www.opengeospatial.org/standards/gml> [February 2012].
Modeling and analysis of migration movement of pre-colonial cultures in the Amazon basin

Carolin von Groote-Bidingmaier
University of Augsburg, Germany
E-Mail: carolin.vongroote-bidingmaier@geo.uni-augsburg.de

Sabine Timpf
University of Augsburg, Germany
E-Mail: sabine.timpf@geo.uni-augsburg.de

Abstract
This research performs an analysis and modeling of migration movement of cultural areas with respect to both spatial as well as temporal aspects. Past research focused mainly on moving point objects, which may be satisfactory in most cases, but does not answer the questions posed in our thought experiment in the Amazon region.

Archaeologists assume that pre-colonial cultures migrated from the western part of South America along the Amazon River to its delta. When representing these cultures (or their influence areas) as polygons a movement or migration pattern should become visible based on different ages of excavated artifacts and language shifts. Using this particular example, I aim to analyze the movement of polygons across the landscape. Therefore the topography and geographical conditions are taken into account in order to categorize the type of spatial change.

When analyzing moving polygons a change of shape versus movement needs to be differentiated. An area can change its shape between time $t_i$ and time $t_j$, i.e., shrinking and expanding, or it can change its location i.e., moving, or both. In an initial approach a comparison of polygons in different time steps is being made, to narrow down the number of not clearly defined changes. Additionally the geographical context is going to be factored in the analysis of shape changes.

Keywords: moving polygons, spatio-temporal analysis

1 Introduction
The focus of this research is on modeling and analysis of spatio-temporal data in an archaeological context.

Archaeologists assume that pre-colonial cultures migrated from the western part of South America along the Amazon River to its delta. Depending on the cultural habits different environmental conditions are required for the movement. To describe this the concept of affordances is used (Gibson, 1979) which characterizes the offers of the environment depending on the specific needs.
To gain a better understanding of how the migration processes took place a spatio-temporal analysis is reasonable. A more detailed view on “where” and “when” may lead to new knowledge concerning the movement behavior.

A lot of research focuses on moving points, which, in most cases may be satisfactory. “The majority of research into monitoring and analyzing movement behaviors focuses on moving point objects. Acknowledging that handling space-time is already difficult enough, moving entities are mostly treated as moving points even if they have spatial extent (as for example moving cyclones).” (Laube 2009, p2). For other approaches the use of points for representing moving objects is not adequate and further research has to be done.

Space-time statistics allow among other things the identification of clusters but “...provide no mechanism for exploring variations in the shape, orientation, size, or movement [...], or to map locations where there have been temporal changes in identified spatial patterns” (Robertson et al. 2007, p.208). Robertson et al. described spatial processes for polygons observed in several time steps. Therefore they present five new movement events to characterize the changes through time (see Figure 1).

![Figure 1](image)

**Figure 1** Types of movement based on polygon events involved in proximity relation. In each case, the unshaded polygons are T1 and shaded polygons are from T2 and the movement type is based on a polygon being within a threshold distance (d) of the start event. For other movement events, only the generation or disappearance events are reclassified as the new movement type (source: Robertson et al. 2007, p.212)
These moving events are based on homogeneous regions that don’t factor the geographical context.

Junghans and Gertz (2010) presented a method to model and forecast the development of moving regions. Their research is based on moving regions that can either be homogeneous or heterogeneous where isolines are used to describe the heterogeneity of region’s interior. They put emphasize on the fact that research focusing on moving regions is has a certain relevance for broad range of applications.

Factoring the affordances and identifying from a change of the polygon weather or not a change also signifies a movement or if the change is due to the vagueness of the underlying data, or an actual shrinking/expansion of the influence area of a people is a topic this PhD thesis is dealing with.

### 2 Movement or Shape change?

The areas of settlement can shrink or expand, move, or both. Assuming an isotropic and homogenous space, the distinction between movement and changing shape may be difficult – or even impossible. Incorporating geographical base data into the spatio-temporal analysis means that potential movement becomes constrained with respect to features in the base data, thus allowing to distinguish between potential movement and shape change.

A first approach for analyzing moving polygons is to describe a polygon for two or more time steps. Based on predefined assumptions a first characterization of the polygon’s transformation is possible (see Figure 3). Let \( P \) be a set of polygons of the same entity in several time steps. If set \( P_i \) is equal to \( P_j \) (D1) then neither change of shape nor movement has taken place.

\[
\text{D1: } P_i = P_j; \quad \forall x (x \in P_i, \exists y \in P_j)
\]

If the intersection of \( P_i \) and \( P_j \) is an empty set (D2) a movement process has taken place. A definite statement about the change of shape is not possible by only identifying that there is no overlap.

\[
\text{D2: } P_i \cap P_j = \emptyset
\]

In many cases the intersection of \( P_i \) and \( P_j \) returns a non empty set (D3) that makes it difficult to detect movement or change of shape.

\[
\text{D3: } P_i \cap P_j = \neg \emptyset
\]
To provide more information about the moving polygons it is necessary to additionally compare some other characteristics of the polygons (i.e. center, standard distance, extent, or rotation). If at least one of these aspect changes during two time steps \( t_i \) and \( t_j \) and the intersection returns a non empty set, a change of shape can be identified but a statement about movement is not possible. E.g. if a landowner purchases additional land properties the shape of the estate changes but does not imply movement.

By examining more than two time steps (if available), some of the ambiguous cases can be categorized. For instance let the intersection of \( P_t \) and \( P_{t_j} \) be a non empty set but the intersection of \( P_t \) and \( P_{t_{i+1}} \) is an empty set (see Figure 3).
In this case there has been movement but it’s not clear if there has been a continuous movement process or not. When \( P_t \) is a subset of a sequence of time steps \( P_t, P_{t+1} \ldots P_{t+x} \) the polygon is probably expanding (D4). 

D4: \( P_t \subseteq (P_t, P_{t+1} \ldots P_{t+x}) \)

One current hypothesis is, that by clustering the different properties of polygon change, some more concrete statements are possible. The previous considerations are based on an isotropic and homogenous space and still don’t cover every possible combination of change. Thus, the first goal of the thesis is to enumerate and categorize all possible polygon changes building on Robertson et al.

3 Enter Geography

A solution to further differentiate between moving and changing shape could be to factor the geographical context in the analysis of the shape change. For this, semantic categories need to be introduced such as stationary objects functioning as barriers (e.g. mountains), or facilitators of movement (e.g. rivers). In addition, field-like qualities such as soil condition or population density might be necessary. Topography is a critical determining factor for movement because it may have a restricting or a facilitating influence. The Amazon River e.g. can support as well as constrain movement. The described spatial influences on migration movement are also time dependent. For instance, the technical capabilities to overcome spatial obstacles depend, among other things, on the respective time period in which the culture lived. In order to identify e.g. the potential path area (PPA), knowledge about migration behavior (barriers and facilitators, speed, technical capabilities) is necessary. Therefore a combination of temporal as well as spatial knowledge is required to explore characteristics of the migration movement.

To clarify the idea explained above it shall be described using two examples:

3.1 Oil spill

Observing the extent of oil spill is another example for the spatio-temporal analysis of moving polygons (see Figure 4). The affordances oil requires to spread in water are different from those needed by cultures to move. E.g. the ocean current and wind speed and direction influence the change of the extent. Only comparing the characteristics of polygons in an isotropic and homogenous
space refuses the identification of whether a change of shape or a movement in several cases. This means the context in which the movement is going to be analyzed is necessarily needed.

3.2 Language groups

Curt Unkel Nimuendajú created a map containing the South American tribes, their spread and the language group they belonged to (see Figure 5). Some of the cultures seem to move along the Amazon River or its tributaries while others stayed at the coastline and therefore had to cross rivers several times. Nimuendajú’s map intensifies the presumption that at least the topography is an important factor for movement of indigenous tribes in South America. But the influence of the topography (and maybe other affordances) varies. To identify the potential facilitators or barriers for movement/migration is a difficult task because they are depending on the cultural habits, their technical capabilities, etc. The affordances of space (e.g., quality of soil, height of the land above floodplain, nearness to drinking water, potential for trade, accessibility, etc.) for the specific properties of the peoples (e.g., nomadic, sedentary) have to be taken into account.

It seems that the context (the affordance) is necessary to identify movement. It might be conceivable that the comparison of several polygons in two different time steps returns the same results but is movement in one case and no movement in the other.
4 Additional considerations

Challenge: no culture-chronological framework

Our knowledge of the pre-colonial period of the immense Amazon basin is disproportionate to its size. Traditionally, archaeological research in South-America still is dominated by the so-called High-cultures of the Andean region. While, based on a high quantity of detailed archaeological information and collected data, it is possible to make statements about, for example, culture-change, social systems, nutritional and subsistence habits of the Andean region, in the tropical lowlands archaeologists are still working on a culture-chronological framework for the few known pre-colonial cultures.

Challenge: different preservation rates of archeological artifacts

The difficulties of systematic archaeological research in the Amazon are not only related to theoretical questions and to geographical factors but also to preservation of material culture. The variety of archaeological artifacts is drastically reduced by the tropical climate conditions. In the pre-colonial sites of the more arid regions of Peru, besides ceramics, metal, human relics, wood, leather, feather, cloth and fibers are preserved. In the Amazon area in the contrary, due to highly moist combined with high temperatures, only un-organic materials, such as ceramic and some stones and charred materials, can be found by archaeologists. (Hilbert & Hilbert 1979)

Challenge: Small number of excavations, potentially non-representative

This research is based on archaeological data obtained from excavations. For each excavation site, the data contains not only spatial information but also additional facts about the different cultures (horizons, traditions and phases) that had settled there as well as their chronological order. Due to the data source, the available information does not contain every settlement but only those that were excavated. That leads to the problem of data availability and reliability. The excavations up to date have been mostly done in so called Terra Preta terrain, above the floodplain along the river in relatively easily accessible places. However, several excavation sites in the middle of the forest have proven that there might
be many more former settlements in the jungle than hitherto known. In addition it is not clear if the floodplain has ever been settled or if the settlement places were restricted to places outside of the floodplain. Another problem related to the data is that the materials used in earlier cultures cannot be preserved and thus found after centuries, if they are of an organic nature such as feathers and wood. This can lead to misinterpretation or over-estimation of the influence of certain culture-traditions that used stone and ceramics instead of bones, wood and feathers.

Challenge: inherently vague data

Furthermore, archaeological data is inherently vague (geographically and temporally). The radiocarbon dating method to discover the age of artifacts is imprecise and produces a range of potential dates or an error margin. In addition, radiocarbon dating is only done for some artifacts because of its high costs. An additional problem lies in the fact that the dated artifact might even not be of that culture because of trade or marital customs.

The goal of this PhD thesis is to develop a procedure to analyze spatio-temporal movement behavior considering the archaeological data focusing on the characteristics of moving polygons.

References


Crowdsourcing Building Interior Models

Julian Rosser
University of Nottingham, Geospatial Institute, United Kingdom
E-Mail: psxjr@nottingham.ac.uk

Abstract
Indoor spatial data forms an important requirement underlying many ubiquitous computing applications. It gives context to users operating location-based applications, provides an important source of documentation of buildings and can be of value to computer systems where an understanding of environment is required. Unlike external geographic spaces, no centralised body or agency is charged with collecting or maintaining such information. This paper takes the position that models of building interiors can be volunteered by users of these spaces. The widespread deployment of mobile devices provides a tool that would allow rapid model capture and update.

Keywords: building modelling, indoor, VGI, 3D

1 Introduction
Digital modelling of buildings is required for a wide variety of purposes. As such, the topic has been approached in different ways from a variety of disciplines. Architecture, engineering and construction (AEC), navigation, positioning, robotics and emergency management are just some of the disciplines and application areas that have a stake in improving the digital capture, update, management and utilisation of building models. The term “building model” refers to digitally encoding useful information on both the geometric structure and the semantics of buildings. Included within the model is information pertaining to the interior of the building which is focused upon here.

The capture and management of building interior data presents a different set of technical requirements and social challenges to those found when working with outdoor environments. To derive 3D virtual cities which model building exteriors on a wide scale, data from air- and space-borne sensors are often used in a semi-automated or fully automated workflow (Haala & Kada 2010). In an industrial context, 3D city model generation may be carried out by National Mapping Agencies (NMA’s) or companies with expertise in remotely sensed data processing. But, unlike exterior geographic spaces, no centralised body or agency is charged with the responsibility of collecting or maintaining the building interior details. This paper discusses the importance of building models to ubiquitous computing and volunteering of this data by the building’s users.
2 Background

2.1 Applications of building models

Building models are useful in a wide variety of scenarios. An exhaustive list is outside the scope of this review; however, the key domains set to benefit from indoor building models are discussed.

One motivating application for deriving interior models of buildings is their use in different positioning and navigation scenarios. Different contexts for indoor navigation are clear. For example, this information would be useful for general navigation and routing, first-responders in emergency situations and individuals with visual or physical disabilities. Locating people and objects with reference to an interior map is not an application restricted to large environments. Explicit spatial modelling of home environments is required if smart buildings and software agents are to understand a user’s context and activity. Indoor spatial models that hybridise geometric and symbolic information offer a promising approach for fulfilling the range of uses required for context-aware applications (Afyouni et al. 2012). These models can be used for improving Augmented Reality (AR) systems which can benefit from provision of a prior initial scene model (Reitmayr & Drummond 2006).

The importance of 3D spatial data in disaster and emergency management contexts has received much discussion. Access to building information has clear advantages in search and rescue (S&R) scenarios providing appropriate modelling (Lee & Zlatanova 2008). Efficient S&R operations are vital to lowering loss of life in disasters and when reviewing information needed in earthquake events. Guven & Ergen 2011 note the importance of building layouts and floor data when identifying information needs for earthquake S&R. Likewise, in 2008 NIST conducted a workshop to identify building information needs of emergency responders in fire, medical and policing contexts (Jones et al 2008). As well as noting that fire services survey large or special buildings, important additional information may be made available to them such as entry and exit points; hazards and obstructions. Workshop participants liked the idea of 3D information for certain features (e.g. pipes and stairwells) and 2D (floor plans) for illustrating incident location.

Building information can provide information for response and recovery planning across smaller geographic scales. Airborne and satellite-based remote sensing is often used for assessing damage resulting from earthquakes. Certain types of building damage can be better identified with additional information such as building height, footprint, floor plan inventory and ground photos (Saito et al. 2004). Knowing the pre-event state of the building helps detect changes resulting from the earthquake. The importance of supplementing remote sensing with ground survey is noted as important in helping quantify uncertainty in final damage assessment results (IWRSDR 2011).
3 Crowdsourcing models

Crowdsourcing interior data for building models enables a building’s users to capture and manage data pertaining to these public and private spaces. The widespread availability of mobile computing devices incorporating cameras and other sensors offers a possible tool for initially producing and maintaining interior building data. The successful application of crowdsourcing to geographic data capture has already been shown in projects such as OpenStreetMap and is in keeping with the notion that GIS should be resident-generated (Talen 1999).

3.1 In situ, bottom-up sensing

The complexity of urban environments presents a need for integration of data sources covering different spatial and temporal scales. Whilst traditional remote sensing can be used for modelling the building’s rooftops and facades, additional in-situ and collective sensing is required to provide a complete view of a city including indoor environments (Blashke et al. 2011). The need for digital encoding of indoor spaces has been identified as an important development for the infrastructure underlying many ubiquitous computing scenarios. The wide-range of applications demonstrates a need for standardised datasets which maybe generalized across the range of uses. But currently there is no ready source of this information. Google has recently launched a service for businesses to upload their own floor-plans for display in Google Maps. Meanwhile ongoing research is looking at how to extend OpenStreetMap to indoor environments through extending the existing tagging scheme with a 3D ontology of building exteriors and interiors (Goetz & Zipf 2010).

Imagery forms an important source of information in addition to geometric and semantic information on interior spaces and which can easily be provided by users. Not only can the imagery be used to texture building models but it can also provide relevant landmarks to users navigating a particular route. Utilisation of landmark information within electronic navigation aids has been found to have significant advantages over paper maps, enabling shorter navigation times and the lowering of mental and physical demand, and is particularly effective with older users (Goodman et al. 2005). Design guidelines on navigational services for pedestrians that state that landmarks should be used as the primary means of routing information (May et al. 2003). It is also important to note that navigation and wayfinding present tasks are required not only for human users but also robotic agents operating in the environment.

3.2 Semantics and temporal resolution of indoor spaces

Both public and private indoor spaces require digital modelling to fulfil the needs of applications identified in Section 2.1. Providing that appropriate management tools for model data can be designed, the users of a building are in many ways the best sources of the information about it. They have access to the building’s rooms, know about different levels of accessibility and understand roles fulfilled by particular areas. As the use of space within
buildings can change, timely updates to the information can be provided. Furthermore, previous model versions may be retained. This update of the information provides a strong advantage over traditional top-down acquisition methods, such as laser scanners. Putting the user in control of the interior data is a primary motivator for this research.

3.3 Model management

It is acknowledged that in some circumstances users may provide incomplete or inaccurate information and this may be accidental or even deliberate. Thus any system to curate crowdsourced building models must be designed to cope with disparate and dynamic user input. Logical checks, quality control and optimization of submitted data can be performed to help maximize the probability of useful contributions. As the technologies for positioning indoors with associated location-based services mature, the number of users capable of active or passive contribution will obviously increase. So we can already recognise that users of an interior modelling system might fulfil two roles: either as a contributor or an application user. However, an application user may also be fulfilling a secondary role of active/passive data update.

Issues of privacy are inherent in the modelling of interior spaces. Firstly, the use of video or image capture methods requires careful consideration. The intention is that 3D geometric features are rendered with textures provided by the imagery; however, these may contain personal details that should not be made publicly accessible. For example, the presence of particular people or valuable objects may be identifiable from the imagery. Furthermore, the model information itself can also be considered personal information.

4 Capture of building geometry

In this section, we outline what we consider to be important requirements for geometric model construction in a crowd-sourcing context.

4.1 Common devices

For high levels of metric accuracy, reconstructing the 3D geometry of indoor environments can be achieved using specialist data capture mechanisms. Terrestrial laser scanners are typically used, providing dense point clouds which can then be modelled in 3D. Research on fully-automatic construction with these sensors is ongoing (Budroni 2010; Okorn et al. 2010). There are also efforts to make laser scanners more portable and more appropriate for indoor model building (Liu et al. 2010). However, the specialist nature of this type of equipment means its use is restricted and updates may be infrequent. Consumer-grade depth-sensing cameras (such as Microsoft’s Kinect) can be used for room modelling (Izadi et al. 2011), but are not yet ubiquitous. For consumer modelling, high accuracy is of less importance and accessibility is critical.
4.2 Feature-less environments

The topic of 3D reconstruction is a long-standing topic within the computer vision research community. A common approach to generating a 3D point cloud from a sequence of image of video frames is to use a Structure-from-motion pipeline. This style of approach has been demonstrated on user-generated imagery (Snavely et al. 2006) and indoor environments (Furukawa et al. 2009). Similarly, Visual Simultaneous Localisation and Mapping (SLAM) techniques generally aim for online estimation of camera pose estimation and key feature locations. The PTAM framework (Klein & Murray 2007), for example, is designed for AR usage and has been shown to work on a mobile phone (Klein & Murray 2009). However, both SLAM and structure-from-motion approaches are not yet robust for all indoor scenarios. Large sections of interior walls either lack the texture required for feature detection and matching or require long model generation times.

4.3 Semi-automatic modelling and clutter

The criteria detailed in 4.1 and 4.2 indicate a need for a semi-automatic modelling process. And furthermore, indoor environments tend to be cluttered by furniture or other objects which occlude wall surfaces – the most desirable features for reconstruction. It could be argued that objects such as furniture should be included within interior models. Recording the location of furniture would be beneficial for certain applications, such as navigation, particularly when visual senses are impaired as in an emergency response scenario. However, it is better to determine the true wall position and then populate the model with further information on objects contained within it. Although making the entire reconstruction process automatic is appealing, it is accepted that the final interaction to identify structure must have user input at some stage.

5 Mobile data capture application

In this section we briefly outline ongoing work on the development of a smart-phone application for modelling building interiors. The aim of the application is to incorporate a simple semi-automatic modelling process which may be used as a starting point for building interior data capture. The application has been implemented for Google Android and tested using the Samsung Galaxy S phone.

Using the orientation sensors and the camera preview of the phone, the positions of key features denoted by the user are estimated (see Figure 1). Trigonometry is used to estimate the distance of the features from the phone assuming a user-specified device height. Ceiling height can also be estimated to produce a 2.5D polygon (see Figure 2). This simple modelling approach contains a variety of sources of error based on several key assumptions: an accurate height measurement for the device, an unobstructed view of key features, a single-level floor and ceiling and a reasonably accurate orientation estimate. These issues contribute to irregularities in the resultant polygon and are regularized to a Manhattan-
world (using constrained least-squares) where all key points are either orthogonal to or parallel with each other.

**Figure 1 (left)** Mobile application interface.

**Figure 2 (right)** Extruded floor plan polygon

---

### 6 Conclusions and outlook

Volunteering information about buildings offers clearly identifiable benefits over traditional top-down methods of indoor survey. While much research effort is directed on 3D reconstruction less attention has been given towards the management, conflation and update of these models. For example, in our data capture application, the initial model scale is dependent on an accurate user selection of device height. The model scale is also changed during the process of geo-referencing and hence the overall representation will vary with different users and levels of base mapping available. Absolute scaling is a common problem with many reconstruction methods that rely on geometric cues or arbitrary estimations of movement.

Furthermore, volunteered building models will likely to grow in an incremental and disparate fashion, but still require efficient storage with appropriate structural and topological constraints. Users may want to model spaces where there are few existing geographic features to refer to. We envisage employment of a rule-based engine to help optimise and enforce constraints, leaving the possibility of sourcing data from a variety of sources and methods. It is also important to note that information capture is required not only initially - when no pre-existing model is available - but also when change to the geometry, texture or semantics of the space is apparent. Therefore, consideration of the infrastructure requirements for the combining and editing of these models is needed and will form an important part of further work.
References


Taking VGI Seriously: The Portuguese Administrative Modernization Agency (AMA) Case Study

António José Silva
Universidade Nova de Lisboa, Instituto de Superior de Estatística e Gestão de Informação, Portugal
E-Mail: d2011090@isegi.unl.pt

Jorge Gustavo Rocha
Universidade do Minho, Braga, Portugal
E-Mail: jgr@di.uminho.pt

Abstract
Under names like neogeography, volunteered geographic information (VGI), wiki-mapping, GIS 2.0, crowdsourcing, citizens as sensors, collaborative mapping, community mapping among others, innovations and new applications of on-line geospatial technologies and crowdsourced spatial data continue their rapid expansion [2]. Users and therefore companies were the first to understand this power in the Web 2.0 age. Complex geographic information started to be produced by completely amateurs, using low cost GPS enabled phones. In less a decade, several millions of roads were traced in OpenStreetMap (OSM). This has revolutionized the way how maps are made. It is thus possible to obtain updated data over the territory, made by whom knows it best: the citizen. However, even with the budget and human resources constraints that many governments are facing, they were not able to take advantage of the huge potential of VGI. No alternative opportunities to collect, update, analyze and use geographic information are being explored.

Using VGI in government is not a technological limitation. We know from our experience in Public Participation Geographic Information Systems (PPGIS) projects, that the use of sophisticated interfaces does not guarantee a massive public participation. By the contrary, the lack of sophisticated data models and tools does not prevent for example OSM users to create detailed maps.

In this PhD work, we have the opportunity to study how VGI can be used in government. In the context of a collaboration with the AMA (Administrative Modernization Agency) public agency, we were challenged to connect the world of VGI and government. But these are two different worlds and we can point some evident difficulties:

i. although public authorities could benefit from geographic information provided by citizens, these public bodies still gather data from traditional sources, and it is legally difficult to accept data from non trusted sources,

ii. although the legal framework for participation in all democratic countries has enshrined the right to participate, the public participation is almost zero in practice,

iii. although the public participation is very low, we never had so many
people participating in social networks and in open data projects. The use of VGI in the local and central administration is a new challenge and the initial attention is almost dedicated to content creation. We want to fully explore the new and updated data from citizens, in order to facilitate, speed up and improve the government decision-making processes.

**Keywords:** VGI, Geographic Information; Government, Citizens.

## 1 Introduction

In recent years there has been a large growing interest in using the Web to create, assemble, and disseminate geographic information provided voluntarily by individuals [5]. With 20 million articles written collaboratively by volunteers around the world, 100,000 regularly active contributors, and with editions in 283 languages, launched in January 2001, the Wikipedia (the free encyclopedia) has become the largest and most popular, general reference work on the Internet (Wikipedia, 2011) and perhaps the best-known user-generated project. There are others, also well-known and interesting, user-generated projects like: Flickr, an image and video hosting website, with geotagging capabilities, and with over 51 million registered members, Youtube, the most used video sharing website, among many others. The amount of information produced by users every day is huge, and it would be a pity to waste it.

Geographical information is not an exception to this trend, and the success of projects like OSM, an worldwide volunteer effort to create a map of the world that will contain as much detailed information as possible, and this information is being collected by volunteers [13] is probably the best example of this. OSM had a significant boost of success and popularity around the world when in January 2010 occurred the catastrophic earthquake in Port-au-Prince, Haiti’s capital. In a few days people around the world have strongly improved and updated the the few available maps for the area using post-earthquake satellite photography provided by GeoEye and aided with immediate reports provided by volunteers on the ground. The OSM map from Haiti was the best map available, constantly updated and made in a record time, it would be impossible to beat if done with traditional methods. This was the first time that the United States Army used maps that weren’t made by them. This case, has demonstrated the power of data created by citizens.

Other projects like Wikimapia, a combination of Google Maps and Wikipedia where anyone with an Internet connection can select an area on the Earth’s surface and provide it with a description [5], and Google Map Maker, in which contributors add to the Google base mapping dataset, are very good examples that attest the growing interest in this topic. Mobile applications as Weddar, a people powered weather mobile service, or Waze, a free community-base traffic and navigation application are examples of a new trend that combines social networking with user generated content. All these sites and services are empowering citizens with tools that enhance the creation of a global patchwork of geographic information, while sites like GeoCommons, Google Earth and others are
encouraging volunteers to develop interesting visualizations and applications using their own free data.

To better understand why this phenomenon is growing so fast lately, it might be interesting to first understand the technologies that make it possible. The proliferation of the Web (broadband communications, and easy access) and consequently the emerging role of Web 2.0 and wikis; the mobile technologies as laptops, notebooks, and more recently smartphones and tablets, digital photography and standards based authentication processes to contribute information to the Web; the widespread use of Global Positioning System (GPS), people now view their own position in real time, that can be easily accessed by a wide range of consumer products; image-based mapping technologies by professionals and amateurs; and the growth of social networking tools, practices and culture [1,5] this whole set of technology is increasingly accessible to a larger number of people.

As expected, geographic data produced by individuals is growing at an incredible pace, faster than the development of work flows able to take full advantage of it. Nowadays, anyone equipped with a basic smartphone with GPS is able to go to the street and start to collect their own geographical data, and easily share location information describing points of interest, places visited, recent construction, and corrections to out-of-date feature attributes. In addition, this is a cheapest source of geographic information, while compared to the traditional ways, and sometimes the only source, particularly in areas where access to geographic information is regarded as an issue of national security [5]. This capability to “view and provide contributions in context” is fundamental to the vision of a spatially enabled society [1].

1.1 VGI

Early on, researchers realized the need to find a name that clearly identifies this concept. Several names have been appearing at the speed with which the various works were being published, each one emphasizing, in most cases, their area of interest, or research. Since 2006 various papers and examples of applications have presented us many terms that tried to describe and define this phenomenon of spatial or geographic user-generated content.

In 2006 Andrew Turner have introduced the concept of neogeography [12], which might be defined as geography for the masses, even as mapping for the masses [8]. One year later, Michael Goodchild termed volunteer geographic information (VGI), a special case of user-generated content where participants are private citizens, mostly untrained, and their actions are almost always voluntary [5]. Song et al, defines VGI as a human side of the sensor revolution, while points some paths to the future saying that “it could be useful to enhance the information used by professionals as well as decision makers” with the advantage that some information cannot be sensed remotely such as local knowledge [11]. This term has been largely adopted by many researchers since then. Both VGI and neogeography are terms
that draw our attention to a new way to produce geographic information in opposition to the classical methods. In fact, this could be a new paradigm for the production of maps, with great impact on Geographic Information Systems (GIS), and a very important manner to narrow the relationship of the citizen with geographical information.

However, other terms have appeared with the same basic idea even though they differ slightly in the focus. Ostlaender presented in 2010 the term Geowikis, “the confluence two social computing trends: user-contributed content and map-based interactive Web applications” [9] very similar to wikification of GIS concept presented by Sui in 2008 and GIS 2.0 by McHaffie in the same year. These names suggest the Web 2.0 modes of interactivity, user-generated content, the iterative mutability of wikis, and their implications for the processes and relationships of knowledge making. Terms like collaborative mapping, community mapping or ubiquitous cartography are terms where the emphasis is on the technological side and in its influence on a interactive way to create geographic information and knowledge [2].

1.1 VGI Issues

Despite the already discussed advantages, and the enormous potential, recognized by all, there are still doubts as to use VGI data, particularly by government, who have always been accustomed to use geographic information collected through traditional methods, with standards and specifications that are usually assured by experienced and qualified technicians in cartography. Due to the amateur nature of VGI, one of the major issues of using it in this organizations is how to assess trust of contributors and the confidence of their actions [1,4]. This is not a new issue, or exclusive to VGI, in many other projects related to the Web 2.0 movement, including Wikipedia among others, questions were being raised about the accuracy and correctness of the information provided [13]. Due to this, several researchers have been studying in the last years the quality of VGI. Between 2008 and 2010 period, Zielstra et al in Germany, Haklay et al in England and François-Girres et al in France, performed exhaustive tests to assess the quality of OSM data in their respective countries. The overall results were quite similar, pointing to a significantly higher quality (positional and completeness accuracy) in larger cities than in inner-cities and rural areas, which can be explained by the presence of more active members on the project in the larger cities [3,6,13].

Beside quality issues, there are other issues that can arise regarding the use of VGI. These issues can be related with:

- the lack of will that people to contribute to government while compared with social networks;
- how do organizations attract new volunteer contributors and how do they keep existing volunteers “engaged”;
- what are the motivations of volunteers;
- the recognition of the contributors;
- will individuals remain interested in making contributions.

These are some issues to which attention is required for the adoption of VGI in an organization (private or public) in order to better understand it.

1.2 VGI Applications

Already exist throughout the world several examples of use of VGI data in various fields. Below are some successful case studies illustrating the use of VGI data. For an easiest understanding this use cases can be grouped by the nature of the users: Government, Businesses and Users themselves.

VGI data in government

Unfortunately, there aren't many use cases of VGI data in government besides severe catastrophe cases, where without an official updated alternative, the authorities have been seen obliged to use VGI data. This reveals that the government still did not understand the information assets it has at its disposal, when it does, it will see many benefits. Below are an illustrative case of how VGI was used in an extreme situation.

Apart from the Haiti earthquake example, discussed earlier, the experience of a series wildfire events in Santa Barbara, 2007-2009, is often used to examine the use of VGI data. In these catastrophes, the severity of the threat meant that approaches used to inform the public during were no longer adequate, so were used instead numerous postings of VGI, using services such as Flickr, provided an alternative to official sources. Citizens were also able to access and interpret streams of data from satellites such as MODI, and several volunteers realized that by searching and compiling this flow of information and synthesizing it in map form, using services such as Google Maps, they could provide easily accessed and readily understood situation reports that were in many cases more current than maps from official sources [4].

VGI data in businesses

There are several good examples of use of VGI data on businesses. Google have understood the potential that lies in VGI and have created Google Map Maker that now provides tools to citizens with the ability to help populate and update Google Maps geographical and attribute data [1].

As google, TomTom’s online MapShare Service is another of the best examples of how one large commercial data supplier uses VGI data. The Company employs a graduated approach to sharing, assessing, and using the volunteer provided updates. MapShare contributors have choice of only using their updates themselves, within their own group, or with the general TomTom community [1]. With the adoption of VGI, both companies have been able
to update their maps quickly and inexpensively, saving valuable resources while providing better service to its customers.

VGI data used by the community

Also the users themselves have been taking advantage of VGI for the resolution of particular themes, organizing themselves in communities where people gather around a common interest. An interesting example is OurWay, a project that enables pedestrians to grade road segments with regards to accessibility, for subsequent use in route planning. Using this mechanism, users with local knowledge, can create feedback and essentially map out their neighborhood. Using OurWay, users can draw upon each other’s knowledge to quickly find the better paths through town [7]. Other interesting example is Cyclopath3, a web-based mapping application serving the route finding needs of bicyclists in the cities of Minneapolis and St. Paul, an area of roughly 8000 square kilometers and 2.3 million people. Cyclopath is a geographic wiki, or geowiki, unifying interactive web-based mapping with the open editing of wikis [10].

2 Hypotheses

We already stated the problem under investigation: how VGI can be used in government?

The answer to this question is, accordingly to our hypothesis, is not focused in how to incorporate VGI in the government's work flows. We argue that the relations between all stakeholders (government, companies and citizens) must change, and equal access and opportunities to the information must be guaranteed to these stakeholders. Each part must be able to take advantage of the information shared and produced by the others. If just one of these parts tries to win and get more advantages than the others, all will lose the opportunity.

Since we are introducing VGI usage in a public agency, our focus is more related with changes in the administration policies related with data and public participation.

3 Methodologies

This work will involve the establishment of a theoretical framework to research and study of literature that allows us to understand VGI, the use of VGI data, their adoption by government and private organizations and what motivates the citizen to participate and contribute in VGI projects. Besides literature, many initiatives are taking place all around the globe related with open data, and promoting news ways of citizens' participation.

Beyond this theoretical framework, our collaboration in AMA will allow us to develop and evaluate several VGI initiatives with national impact.

We will need a comprehensive spatial-temporal data model that will allow the storage and the analysis of large VGI data-sets, resulting from initiatives under study. This data model
must capture and register all observations and measurements that will be taking before and after each change introduced in the VGI initiatives.

Three initiatives are already planned and will be used in the first experiments.

“Fix my street” (“A minha rua” in Portuguese), is already running on the web, allowing citizens to report problems in their neighborhood. All reports are routed to the corresponding local or central authority in charge. But since many requirements were not initially identified, and some major limitations were present in this initial release, we are redesigning the all service. We have the possibility to completely redesign the service, and to change and enhance the communication and between the citizen and the administration. Besides web, we also introduce interfaces for smart phones able to capture georreferenced snapshots of the problems reported. We will also support an innovative interface, on the TV power box. The test and development cycles of this “Fix my street” service will provide us valuable feedback about public participation, rewarding mechanisms, etc.

“Services near me” (“Serviço + perto”) is another platform that we will design from scratch. It will take advantage of many statistics available about administration services, to provide guidance about where and when someone should go to get something from the administration.

The third initiative “Let's map Portugal” (in Portuguese “Vamos mapear Portugal”) was already announced, but still on it’s early stage. We will developed and implemented monitoring mechanisms for understanding the adoption of the "Vamos Mapear Portugal" by government, companies and the citizens. By carefully implementing those mechanisms in advance, we will be able to measure and evaluate its acceptance.

From the beginning, open and service oriented architectures will be adopted. This might enable private companies to take advantage of VGI. For example, a missing one way restriction reported by a citizen can be used simultaneously by the local municipality, the OSM community and by a transportation company.

4 Results

The AMA agency is willing to make data available and to provide useful services to the citizen. With the experiments already scheduled and with others to develop during this PhD, with good measurement instruments, we will be able to get very interesting temporal data about VGI and public participation. From such large observations, we will be able to identify the best practices and less successful initiatives. We be able to prove our initial hypothesis if the identified best practices are promoting equal access and opportunities to both the government, companies and citizens.
References


2. Elwood, S. Geographic information science: emerging research on the societal implications of the geospatial web. Progress in Human Geography 34, 3 (2009), 349-357.


Context-based geoinformation technologies in e-Tourism

Thomas Spangenberg
University of Applied Sciences Harz, Wernigerode, Germany
E-Mail: tspangenberg@hs-harz.de

Abstract
The paper presents challenges and some results of our research project and should be understood as “work in progress”. The main focus is on personalized location-based services in tourism, travel planning, points of interest descriptions and context-awareness on mobile systems. In order to verify the formal concepts and data models, partial results were evaluated as prototypes in different scenarios at the Harz region.

Keywords: tourism, location-based services, points of interest, context-awareness

1 Introduction
Tourism is an important global industry and a strong growing market, also in the Harz region in Germany. But the trend towards individualization changed the requirements concerning Tourism. More and more people are using powerful smartphones including geo-sensors for satellite positioning and Internet connection, therefore a boom of mobile, especially location-based applications is noted in Tourism (Goell et al. 2010). Certain touristic target groups have a particularly high demand for location-based information. The identification of specific user requirements is therefore an important issue (Nivala et al. 2009). With the currently existing tools and services it is not easy to identify the expected information. An intelligent and content-based information filtering requires a high degree of detailed information.

Geoinformation technologies provide the opportunity to respond to the specific interests of tourists with adequate information and personalized recommendation systems. In addition to presenting spatial data on the Web, mobile devices offer the ability to respond to context-based user interests. However, easily accessible and platform independent, individualized location-based services (LBS) are not yet established (Angerer 2010). Regarding the architecture and the content there are lots of pros and cons of existing LBS (Goossen et al. 2010).

Most tourism activities are composed by several integrated services like transportation, accommodation and other external factors. However, common standards are needed to allow interoperability for data exchange between different systems.
2 Approach

The goal of our project “Communication and Information Technologies for sustainable, regional Development (KOGITON)” is to support tourism in the Harz region using geoinformation technologies. The focus is on tourist information systems, travel planning and intelligent location-based services to actively assist tourists by giving access to relevant context-based information during their trip (Pundt et al. 2011). Service-oriented architectures distribute this information through standardized web services in heterogeneous contexts. The integration of various information sources requires reliable interoperability of the different system components, achieved through the application of open standards.

In a first step, relevant spatial and temporal data have been defined for tourism purposes in the Harz region. The focus lies on points of interest from various sources, especially Volunteered Geographic Information (VGI), and the presentation by OGC-compliant services for web and mobile systems.

2.1 Exchange of Points of Interest

Points of Interest (POI) are specific locations which are used in many navigation applications, maps or location-based services. For the storage and exchange of POI data, there are still no uniform standards. The practical implementation is done in different proprietary formats, for example:

- Destinator Points of Interest (*.dat)
- Garmin Points of Interest (*.gpi)
- iGO2008 Points of Interest (*.upoi)
- Maptech Waypoint File / Maptech Exchange Format (*.mxf)
- Nokia Landmarks (*.lmx)
- TomTom POI File (*.ov2, *.asc)
- Text-based (*.asc, *.csv, *.txt)
- …

The exchange of POI data for the use in different applications inevitably leads to a large transformation effort.

The "Point of Interest Exchange Language" (Kanemitsu & Kamada 1999) defines a general-purpose specification language for describing POI location information, without taking into account context-awareness. Based on existing approaches, such as the xPOI concept (Kroesche & Boll 2005) for modeling, processing and visualisation of context-aware POI and a description of geo-referenced multimedia objects of interest (Haid et al. 2005), the research will be continued in future. De facto standards like the Topografix GPS Exchange Format (*.gpx) show a positive trend towards interoperability, but have to be extended for
the exchange of POI data and adapted to the individual users needs and his/her special situation.

In the current project stage, we use the mapping of POI in a PostgreSQL database and SOAP web services (example query in 1) for the platform-independent exchange of POI. As basic data about 700 POI from the Harz region are used, provided by a project partner (brain-SCC GmbH Merseburg) and combined with POI from the OpenStreetMap Project.

![Soap-Request:](image)

**Soap-Response:**

```xml
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <SOAP-ENV:Body>
    <ns1:getByCategoryRequestCompleteLevel="2" language="de">
      <category id="1"/>
      <category id="2"/>
      <category id="3"/>
    </ns1:getByCategoryRequestCompleteLevel>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

**Soap-Response:**

```xml
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <SOAP-ENV:Body>
    <ns1:getByCategoryResponseCompleteLevel="2" language="de">
      <category id="1"/>
      <category id="2"/>
      <category id="3"/>
    </ns1:getByCategoryResponseCompleteLevel>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

**Figure 1** SOAP web service with public POI data

### 2.2 Content adaption

The success of location-based services, in addition to the technical implementation, largely depends on the content. One project aim is a more comprehensive description of POI, which means among other things, that user-specific issues are included in their specification. The user expects that information is provided in a highly personalized way, adapted to his profile and preferences (Hoepken et al. 2010). This can be ensured if an application takes only such POI into account, which are actually of interest of the user; however this is not the case in many existing applications that show simply “all” POI in a defined region – but “all” POI must not be of interest necessarily for “all” users. The inclusion of such an approach in a prototypical travel planner, which is aimed at an improved “individualization” of the mobile application, is therefore a goal of the project mentioned above. The app, however, should react flexibly on such individual user requirements. The opening times of public institutions may serve as simple example (table 1).
<table>
<thead>
<tr>
<th>core elements</th>
<th>title, category, location</th>
</tr>
</thead>
<tbody>
<tr>
<td>extended</td>
<td>description, time related factors (e.g. opening hours, average duration of stay), language, admission charge, media information, address, contact, ...</td>
</tr>
</tbody>
</table>

**Table 1** POI elements

The data used for the travel planning app include, among others, the POI opening hours as part of the database (see figure 2).

<table>
<thead>
<tr>
<th>id</th>
<th>fac_id</th>
<th>time_start</th>
<th>time_end</th>
<th>day_mo</th>
<th>day_tu</th>
<th>day_we</th>
<th>day_th</th>
<th>day_fr</th>
<th>day_sa</th>
<th>day_su</th>
<th>day_holyday</th>
</tr>
</thead>
<tbody>
<tr>
<td>1250</td>
<td>2904</td>
<td>13:00:00</td>
<td>18:00:00</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1262</td>
<td>2910</td>
<td>10:00:00</td>
<td>16:00:00</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1263</td>
<td>2910</td>
<td>10:00:00</td>
<td>17:00:00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1265</td>
<td>2915</td>
<td>09:00:00</td>
<td>17:30:00</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1266</td>
<td>2915</td>
<td>10:00:00</td>
<td>16:30:00</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 2** POI opening hours

The inclusion of such information for each POI ensures that only those POIs are displayed by the application that are open in the current situation. If the user is aiming at visiting the POI just an hour before closing time, the travel app could warn him and make a suggestion to visit another point. Another exemplary factor is the “nearness” of the user. Pedestrians, for instance, have other requirements compared with bikers, or car drivers, so, the meaning of “near” depends from the situation of the user:

- Pedestrian → 500 meters ?
- Biker → 2000 meters ?
- Car drivers → 10 kilometers ?

In general, the consideration of interoperability issues, technically and semantically, is a crucial point. One aim defined in the project work plan is therefore the development of a formal ontology related to the POI database (Pundt & Spangenberg 2010). This ontology should lead to semantically enriched information provision aimed at improving context-based, individual trip planning. In a GML-based approach, points of interest are enhanced with semantic meta information for the use of intelligent LBS.
3 Use Cases and Scenarios

The development of a reliable geoportal for travel planning using smart web mapping technology and advanced spatial functionality is a central goal in KOGITON. The Travel Planner (Spangenberg & Pundt 2010) - based on a Content-Management-System - can provide a mathematically optimized tour with the help of select POI for day or weekend trips. Users can furthermore integrate individual requirements and personal wishes (as shown in 3) into the planning procedure.

![Figure 3 Personalization in the Travel Planner](image)

A distinctive feature of the Travel-Planner is a modified algorithm for solving the classical Traveling Salesman Problem. It is not only about calculating the shortest path, but the best in terms of individual user needs. As a solution, approaches for multi-criteria optimization algorithms (Neumann et al. 2006) were used.

An extended use case is the integration of weather data like probability of precipitation, temperature, wind chill factor and visual range as well as the daily sunrise and sunset times as planning-determining factors (Spangenberg & Pundt 2011). Currently they require the user to investigate manually in different media sources and then draw conclusions to match the personal needs with the travel route. For a recommendation of a route, the existing travel planning prototype can be used for analysis to discover patterns that describe associations among features that occur frequently in the data.
In the project context, applications for basic city information, sightseeing, travel planning and hiking guides were created to get a more comprehensive knowledge of different user groups and their specific requirements. With the assistance of external companies, who show great interest in this research, these applications were tested on different platforms like Android and Apple iOS.

Acknowledgments

The research project Klik-KOGITON is promoted by the European Commission under the European Regional Development Fund (Fkz: 1211080007 KAT) and at the Centre of Competence for Information and Communication Technologies, Tourism and Services at the University of Applied Sciences Harz.

References


Pundt H., Spangenberg T.: Individualized Travel Planning through the Integration of different Information Sources including a POI Ontology, in: Zipf A., Behncke K., Hillen F., Schaefermeyer J.
Pundt H., Spangenberg T., Weinkauf R.: Individualisierung webbasierter und mobiler GI-Dienste zur

Spangenberg T., Pundt H.: Tourenplanung als Geo-Extension in einem Open Source Content-

Spangenberg T., Pundt H.: Integration of Dynamic Environmental Data in the Process of Travel