Geo-Ontology and Geographical Information System Extended by a first-order logic language: Application to malaria control

Guillaume Koum[1] Josiane Etang[2],[3]

[1] Département d’Informatique. Ecole Nationale Supérieure Polytechnique. B.P. 8390 Yaoundé, Cameroun
g_koum@yahoo.fr

[2] Organisation de Coordination pour la lutte contre les Endémies en Afrique Centrale B.P. 288 Yaoundé, Cameroun

[3] Faculté de Médecine et de Sciences pharmaceutiques. Université de Douala, B.P. 2701 Douala, Cameroun
josyet@yahoo.fr

Abstract

A geo-ontology can be built around a Geographic Information System (GIS) and enriched with a first-order logic language under the closed-world assumption (CWA). This extended system provides knowledge representation formalisms that are aimed to describe general conceptual information. They can also be used in the construction of the knowledge base of a reasoning tool. In this paper an environment is defined where a geo-ontology offers possibilities with a system that allows users to represent and reason about some aspects of the real world. With respect to the added intelligence; knowledge is preserved and enhanced. Knowledge enhancement is performed under the CWA in a subset of the first-order logic language. A geo-reasoning task is undertaken so that a GIS and a geo-ontology are integrated in what can correspond to a natural language processing system. The framework is implemented to demonstrate geo-ontology’s ability to enhance knowledge in malaria control domain.

Keywords: geographic information system, geo-ontology; closed-world assumption, geo-reasoning, malaria control.

1. INTRODUCTION

In order for the semantic web to function, computers need access to structured information and reasoning mechanisms [1]. Key to this is the use of ontology. Ontology is considered as specification of a conceptualization, which provides the structured vocabulary and semantics which can be used in the markup of web resources to provide machine understanding [2]. When formalized, ontology can be interpreted by computer easily, and the information contained in the ontology can be processed on semantic level effectively.

A geo-ontology is ontology with geographic features [3]. When defined, a geo-ontology model must be capable of:

1. Representing relationships between concepts.
2. Representing constraints on relationships.
3. Expressing integrity rules between individuals belonging to different concepts.
4. Representing advanced composition hierarchies.
5. Etc
A geo-ontology can be built with capabilities of extending a Geographic Information System (GIS). In malaria control for instance, as far as the geographic distribution and ecological requirements species are concerned, using a geo-ontology approach can be as efficient as predictive species distribution modeling approach based only on presence records.

Using computer modeling and data on climate and human populations, the complex landscape of malaria across a country can be revealed through malaria maps. These maps are aimed at everyone involved in the battle against the disease, particularly those actually doing the disease control work on the ground. However, accurate maps detail where the disease is most intense and where the largest concentrations of people at risk are found [4].

This paper is organized as followings. After the introduction in chapter one, the second chapter talks about ontology. The intelligence meaning and the Closed-World Assumption (CWA) are highlighted in this part. In chapter three, the construction of the geo-ontology is undertaken and the corresponding map of Cameroon is extracted from the GIS. The map is generalized through a change not of scale, but of layer. Generalization is one of the most important elements in effective representation of spatial data, especially for GIS data. A logic-based language is formulated. It extends the geo-ontology and contributes to evaluate predicates occurring in the system while complete knowledge is assumed. Chapter four is related to the conclusion where we project the work on the future.

2. ONTOLOGY

2.1 Geo-ontology

Geo-ontology results from analysis and modeling of ontology in geo-spatial application that is concepts and the relationships between concepts which is abstracted from real geographic space. A geo-ontology can be defined as this: Geo-ontology is the formalization of concepts sharing among GIS field. Sharing concepts refers to the concept models of geographic information, which are the abstract models generalized from cognition of geographic phenomenon added on the type of concept being used as well as some restrictions Geographic is dedicated to describe the physical structure of reality and ontology the concept explicitly. After formalization, geo-ontology can be readable both to human and computer [5]. Therefore, geo-ontology is used to describe the characteristic of data and resource and data acquiring mode, and thus to provide a uniform expression for data integration and sharing.

2.2 The meaning of intelligence

In the geo-ontology application field, geo-ontology reasoning and the query based on it are essential to GIS data sharing and interoperating. The function of reasoning is to discover potential relationship with the known relationship and acquire connotative knowledge with given knowledge by certain logic and rule. Ontology reasoning is a typical instance that the computer comprehends the ontology knowledge. Computer can interpret the information described by ontology and complete task intelligently [6]. The capability of ontology’s reasoning proves the fact that ontology is a formal specification of shared concept model. Ontology reasoning is able to judge if certain relation existing among instances.

2.3 Closed-world assumption

2.3.1. Definitions

The Closed-World Assumption (CWA) holds that anything that cannot be shown to be true is false; no explicit declaration of falsehood is needed [7]. Consequently, any query (which terminates) will either return true or false; there is no possibility of "unknown".
Technically, one can ask the system to prove that something is Unknown, or to prove that something is Known, so long as there is a means to represent such a query.

The real difficulty with the CWA is that one cannot add information. Everything is known a priori.

Most programming systems which use logic or predicate calculus use the CWA, including:

- Prolog Language
- Relational Databases (these can be viewed as predicate systems)

The concept of proof that underlines the Prolog algorithm is relied under the non-monotonic inference rule called Negation as Failure (NAF). For relational databases, the CWA says that all information not true in the database is considered as false.

2.3.2. Ontology domains

In the knowledge management arena, the closed world assumption is used in at least two situations: 1) when the knowledge base is known to be complete (for instance a mosquitoes database containing records for every vector), and 2) when the knowledge base is known to be incomplete but a "best" definite answer must be derived from incomplete information.

The geo-ontology may suggest more than one domain concerning malaria control (ecological component, biological process, resistance function, etc…), each represented by a particular ontology. All terms in a domain can trace their parentage to the root term of the domain, but the ontology itself has not a root. The root nodes are unrelated and do not have a common parent node, and hence the geo-ontology is referred to both as ontologies and as a single ontology consisting of many sub-ontologies. Some graph-based software may require a single root node; in these cases.

3. ELEMENTS OF PRACTICE

3.1. Geographical information system Map.

The map in figure 1 is a map of Cameroon. Cameroon is a country of Central Africa, located between 2-12° latitude North and 8-16° longitude East. It is commonly referred to as a “miniature Africa”, owing to the diversity of the geographical and climatic environments it presents. Three members of the An. gambiae s.l. complex exist in Cameroon and are distributed in five ecological regions from North to South [8].
An. gambiae is comparatively more associated with conditions characterized by higher rainfall and humidity, which are characteristics of the equatorial rainforest. Anopheles gambiae is in fact an assemblage of populations belonging to two molecular forms. Anopheles arabiensis is mainly distributed in the most xeric habitats of northern Cameroon that are characterized by high values of evapo-transpiration and sunlight exposure.

The presence of a highly differentiated malaria vector system occurring in a given geographical area, as observed in Cameroon, can clearly have a profound impact on the nature and intensity of transmission [9]. In this context, fine-grained mapping of the vectors’ distribution together with the identification, characterization and ranking of their ecological requirements, as well as of the ecological determinants to which mosquitoes respond, is of great interest to assess and predict disease transmission risk.

3.2. Term structure

The geo-ontology (GO) we construct is a set of standard terms—words and predicates—used for refining information. The structure of GO can be described by a two-level structure as followings:

1) the first level is a graph where nodes are general terms and the is-a relationship characterizes arcs between terms

2) - the second level is also a graph where each node is a specific term and the relationships between the terms may have different meanings.

The relationships used in GO are directed and the graph is acyclic, meaning that cycles are not allowed in the graph. The ontology resembles a hierarchy, where the two levels are such that the instance relationship traduces the arc between a child term and a parent term. Parent terms in the first level are general terms and child terms in the second level are specific. The diagram

Figure 1. Map of Cameroon with some malaria vectors ([8])
presents the pattern of the combination of two graphs, one per level and the duality of is-a relationship and instance relationship is expressed by a virtual and horizontal arc for the is-a association and by a physical and vertical arc for the instance association.

3.2.1. Unique identifier and term name
Every term has a term name—e.g. sudana savanna, forest, mangrove, etc.—and a unique identifier (often called the term). The suffix may be used to precise the nature of term. One example is the suffix pop which indicates the population of the locality.

3.2.2. Namespace
Denotes which of the sub-ontologies—ecologic component, biological process, resistance function, etc.—the term belongs to. The namespaces as in XML schema are nodes and are associated to a root.

3.2.3. Definition
A textual description of what the term represents, plus reference(s) to the source of the information is available in the ontology. The documentation aspect of the system is built through this way. All new terms added to the ontology must have a definition.

3.2.4. Relationships between terms
Two physical associations capture how the term relates to other terms in the ontology. Spatial features of the geo-ontology are visible by the fact that at each level of the structure; geographical information is modeled. In the first level for example, a region is considered as a set of spatial data. In the second level, despite the fact that a locality is suffixed by its population, it is an instance of spatial data term. The geo ontology in its second level employs a number of other relations, comprising include or part of (e.g. north, part of sudana savana) and locates (e.g. ngaoundere locates adamawa) as shown in figure 2.

![Figure 2. Geo-ontology limited to sudan savanna ecological region](image-url)
3.2.5. Border conflicts between ecological and administrative regions

The mapping between ecological and administrative regions leads to border conflicts. At this point, two kinds of border conflicts are identified:

- Two ecological regions may share at least one single administrative region. For instance sudan savanna and sahel ecological regions cover the Extreme North administrative region; Namely only the south of the Extreme North region is relevant to the sudan savanna ecological region.

- Two administrative regions are involved in at least one ecological region. This the case of the North West and West administrative regions which can be characterized by the sudan savanna ecological region.

3.2.6. Incremental ontology building

The complexity may arise when assembling geo-ontologies related to all ecological regions. Border conflicts can be intensified in that occasion. Areas of conflicts must be handled as exceptions in a programming language. As far as logic-based languages are concerned, the existential quantifier is able to consider the particularity of those borders.

Figure 3 shows a map of Cameroon corresponding to the administrative layer.

Figure 3. Map showing the 10 Administrative regions of Cameroon
3.3. Spatial data generalization versus multi-layering

Spatial data generalization in GIS has a close relationship to traditional map generalization, but
the differences exist at the same time. Spatial data generalization in GIS is driven by analysis
and query of geographical information [10].

By the analysis and comprehension of map generalization, CWA should entail the
generalization of spatial data and should determine their effective organization. Realization and
establishment of efficient and seamless geographic database is the foundation of multi-scale
spatial data representation and processing. Furthermore, natural language processing can be an
important key of generalization. The close relationships between terms in the ontology
emphasize the conceptual models of spatial data generalization that a natural language is able to
express.

In general; map generalization aims at creating a map, which is easy to understand and looks
pleasing to the eye. Besides, the complexity of spatial terrain features is simplified when the
scale changes smaller, in this way the primary and essential terrain features can be reflected
while the secondary and nonessential aspects can be abandoned.

Generalization is no longer the precondition of geographical information transmission between
multi-scale or multi-resolution data. We promote the multi-layer approach for data
generalization in GIS with the help of the geo-ontology.

A layer is a set of similar features representing a class of features that exists in the world. Here
two particular layers are concerned:

- an ecological layer sampling some malaria vectors in Cameroon,
- an administrative layer comprising regions of Cameroon.

A layer is not actually a data source, but is an object within the GIS that represents a data
source that may be present on a local or networked drive, or the layer data source may exist on
an internet mapping server.

A layer should not simultaneously represent more than one class of features, although it may
represent several subclasses.

A map document can contain many data frames, and each data frame can contain many layers.
Generally, the layers within a single data frame represent data for a common area of the earth.
Since these two layers represent the same country, they will overlap. The result is presented in
figure 4.
3.4. First-order logic language

We recall what follows [11]. Let \( \Sigma \) be a signature. The first order language \( \text{FO}(\Sigma) \) on \( \Sigma \) contains the following:

1. the set \( S(\Sigma) \) of symbols of \( \text{FO}(\Sigma) \), which is the disjoint union of the following sets:
   1. \( \Sigma \) (the non-logical symbols),
   2. a countably infinite set \( V \) of variables,
   3. the set of logical symbols \( \{\land, \lor, \neg, \rightarrow, \leftarrow, \forall, \exists\} \),
   4. the singleton consisting of the equality symbol \( \{=\} \), and
   5. the set of parentheses (left and right) \( \{ (, ) \} \);
2. the set \( T(\Sigma) \) of terms of \( \text{FO}(\Sigma) \), which is built inductively from \( S(\Sigma) \), as follows:
   1. Any variable \( v \in V \) is a term;
   2. Any constant symbol in \( \Sigma \) is a term;
   3. If \( f \) is an \( n \)-ary function symbol in \( \Sigma \) and \( t_1, \ldots, t_n \) are terms, then \( f(t_1, \ldots, t_n) \) is a term.
3. the set $F(\Sigma)$ of formulas of $FO(\Sigma)$, which is built inductively from $T(\Sigma)$, as follows:

1. If $t_1$ and $t_2$ are terms, then $(t_1=t_2)$ is a formula;
2. If $R$ is an $n$-ary relation symbol and $t_1, \ldots, tn$ are terms, then $(R(t_1, \ldots, tn))$ is a formula;
3. If $\varphi$ is a formula, then so is $(\neg \varphi)$;
4. If $\varphi$ and $\psi$ are formulas, then so is $(\varphi \lor \psi)$;
5. If $\varphi$ is a formula, and $x$ is a variable, then $(\exists x(\varphi))$ is a formula.

In other words, $T(\Sigma)$ and $F(\Sigma)$ are the smallest sets, among all sets satisfying the conditions given for terms and formulas, respectively.

Formulas in 3(a) and 3(b), which do not contain any logical connectives, are called the atomic formulas.

In our system, the language extending the geo-ontology contains a first-order logic core with generalized quantifiers, and plural reference expressions. It comprises the Geo predicate.

The mosquitoes database indicates that malaria vectors are either *An. arabiensis* or *An. gambiae*. We consider that $\text{vect}_1$ is a *An. gambiae*, and $\text{vect}_2$ is *An. arabiensis*. This can be represented in the language in the following way.

$$\forall x \text{ vector} (\text{inst: } x) \iff \text{An. gambiae} (\text{inst: } x) \land \text{An. arabiensis} (\text{inst: } x)$$

$$\text{isFound (inst: locality, theme: adamawa)}$$

Here $\text{vect}_1$ and $\text{vect}_2$ are constants, vector, *An. gambiae*, and *An. arabiensis* are predicate symbols. In the language, arguments of predicates are identified via keywords, e.g. any identifier preceded by a question mark, e.g. $?x$, is a variable.

In declaration we use the boolean operators:

iff ($\iff$ meaning equivalence),
and ($\land$ meaning conjunction),
or ($\lor$ meaning disjunction),
the universal quantifier forall ($\forall$).

The question whether all the vectors are in the Adamawa administrative region can also be formulated in the language. To answer this question we must try to infer from the database:

$$\forall x \text{ vectors (inst: } x) \iff \text{An. arabiensis (inst: } x) \land \text{An. gambiae (inst: } x)$$

$$\text{isFound (inst: locality, theme: adamawa)}$$

On the basis of the CWA, the answer has to be positive. Although other vectors may be present in the field, but since they are not recorded in the database the CWA entails a positive response. Users are more familiar to administrative data than ecological data.

At the border of the Adamawa and North administrative regions, the sudan savanna ecological region disappears and another ecological region is observed. The existential quantifier $\exists$ must be introduced in queries when there are areas affected by border conflicts.

Now let us assume that the *An. gambiae* S-form from vect1 is found nationwide. This can be represented in the following way.
Is-a(inst: vect1, theme: locality)
Geo (inst: locality)
isFound (inst: locality, theme: nationwide)
To answer the question whether all An. gambiae from vect1 are found nationwide we have to try to infer the following language expression.

(forall ?x
  isFound (inst: ?x, theme: nationwide)
  if Is-a (inst: vect1, theme: ?x)
  and Geo (inst: ?x)).

Under the CWA, the answer to this question is negative.

Geo predicate is what introduce the geo-ontology within the first-order logic framework. By taking into consideration all pop-suffixed terms, the aboutWhere association in the geo-ontology brings information at the view of the (administrative) region. As it is virtually defined in the first level of the geo-ontology, an ecological region is a set of (administrative) regions. This operation can be repeated for all localities of a region. By appreciating the presence or absence of given vectors in different localities, then we can determine a spatial generalization in the region not directly from the GIS but from the geo-ontology.

This is perhaps not surprising considering that marginality values are related to the extent of the spatial reference set, which in this case was constituted by the whole of Cameroon, a highly diversified country covering several different bio-geographic domains.

3.5. Related work

3.5.1. Former situation

Simard [12] has studied the biology, ecology, importance in the transmission of human pathogens, resistance to insecticide and population genetics of the five main human malaria vectors, all present in Cameroon. An. arabiensis extends from the dry savannas in the North (southern border of lake Chad) down to the evergreen forest edge (around 5°N) and An. gambiae s.s. is widespread throughout the country. The species, An. gambiae, is found virtually everywhere in Cameroon and transmits malaria to humans in humid, forested environments in the South as well as in dry savannas and ricefields in the North; in rural, periurban and urban settings; at low (Douala, 12 m a.s.l.) or high (Dschang, 1400m a.s.l.) altitude.

3.5.2. Comparison

According to Simard’s work, the M-form of An.gambiae was identified in Tibati (6°28’N; 12°37’E). For us, it is known that Tibati is a locality of Adamawa administrative region. However Tibati does not belong to the sudana savanna ecological region. If this was true we would have said that the An.gambiae M-form is not found in Tibati because this locality is a not under the influence of the sudana savanna ecological region.

Consequently we have to take into consideration the exception raised by the Tibati case as a border area.

To answer the question whether An.gambiae M-form from vect1 is found in Tibati we must try to infer the following language expression.

(∃?x if
  An.gambiae M-form (inst:?x)
  isFound (inst: ?x, theme: Tibati)
  if Is-a (inst: vect1, theme: ?x)
  and Geo (inst: ?x)).
Therefore the CWA proves in our study that the M-form of *An.gambiae* is found in Tibati. At the end, we have a common position with Simard’s work. Nevertheless, the difference between us is determined by the fact that our solution is derived from the database. In Simard’s point of view the presence is recorded in the database.

4. CONCLUSION

Analysis, query and visualization of GIS always come down to multi-resolution data, representation, multi-scale data integration, and so on, which are all supported by spatial data generalization. The most common solution is to establish and save different scale database in GIS and set the scale range of different elements and layers in display. When utilize GIS to process data sets of different scales, especially from different sources, some conflicts usually come forth A geo ontology enriched with the logical proof containing the CWA constitutes another method of spatial data generalization. Spatial data generalization satisfies not only map display and information transmission but also spatial analysis and data integration. This exercise can be performed in the malaria vectors distribution. *An.arabiensis* and *An.gambiae* are the ones concerned by the ecological partitioning in a geo-ontology and a logical point of view. Analyses may differ according to eco-geographical variables induced for instance in the Ecological Niche Factor Analysis. However as we demonstrated in this paper, the divergence depends only on the quality of data recorded in the field or derived.

In the future, we would call upon a more complex logic language to sustain the geo-ontology by introducing namely temporal and modal operations. The temporal aspect should permit the declaration of data collection time and snapshot issues. The modal operation should be the key for the reasoning in the system to be non-monotone.

Bibliography

1. Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann, and Susie Stephens (2007). “The Semantic Web in Action”. *Scientific American*, 297(6), pp. 90-97. December.

2. Tim Berners-Lee, James Hendler and Ora Lassila (2001). “The Semantic Web”. *Scientific American*, May.

3. Yandong Wang, Jingjingf Dai, Jizhen Sheng, Kai Zhou Jianya Gong (2004). “Geo-ontology design and its logic reasoning”. *Proc. Of SPIE* Vol. 6753 675309-09

4. Fontenille D, Simard F (2004): “Unravelling complexities in human malaria transmission dynamics in Africa through a comprehensive knowledge of vector populations”. *Comp Immunol Microbiol Infect Dis*, 27:357-375.

5. Joshua Lieberman, Raj Singh, Chris Goad (2007). “W3C Incubator Group Report”. 23 October 2007. [http://www.w3.org/2005/Incubator/geo/XGR-geo-ont/](http://www.w3.org/2005/Incubator/geo/XGR-geo-ont/)

6. T. R. Grüber (1995). “Toward principles for the design of ontologies used for knowledge sharing”. Padua workshop on Formal Ontology, March 1993, later published in *International Journal of Human-Computer Studies*, Vol. 43, Issues 4-5, November, pp. 907-928.

7. U. Hustadt (1994): “Do we need the closed-world assumption in knowledge representation?” *In J. Kunze and H. Stoyan, editors, KI-94 Workshops: Extended Abstracts*, pp. 123--124. Gesellschaft für Informatik.

8. Simard F, Ayala D, Kamdem G, Pombi M, Etouna J, Ose K, Fotsing J-M, Fontenille D, Besansky N, Costantini C (2009): “Ecological niche partitioning between *Anopheles*
9. Pombi M, Caputo B, Simard F, Di Deco MA, Coluzzi M, Della Torre A, Costantini C, Besansky NJ, Petrarca V (2008): “Chromosomal plasticity and evolutionary potential in the malaria vector Anopheles gambiae sensu stricto: insights from three decades of rare paracentric inversions”. *BMC Evol Biol*, 8;309.

10. LI Aiqin, DENG Yi, LI Mingfeng, SHI Xinxing (2008). “Conceptual modeling of spatial data generalization in GIS”. *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Geo-Simulation and Virtual GIS Environments*. Lin Liu, Xia Li, Kai Liu, Xinchang Zhang, Aijun Chen, Eds., Proc. of SPIE Vol. 7143 714336-1. 2008 SPIE Digital Library D.

11. Marker (2002). “Model Theory, an Introduction”. *Springer*.

12. Frederic Simard (2008). “Bionomics and Genetics of Anopheles gambiae in Cameroon”. [www.anobase.org/embo_meeting/.../Simard.pdf](http://www.anobase.org/embo_meeting/.../Simard.pdf) 2008.

AUTHORS

**Dr. Guillaume KOUM** is Associate senior Lecturer at University of Yaoundé 1 (Computer Department). His research domains cover a variety of topics relative to Database Systems, Software engineering, Artificial Intelligence, Information Systems. He first studied Mathematics in University of Yaoundé (Cameroon), where he obtained a Licence (Bachelor) in 1977. He after obtained a Doctorate in Computer Science from University Claude Bernard, Lyon1 (France) in 1982. He has been the Chief of Division in charge of Computer Technologies at the Ministry of Finance in Cameroon, before being appointed as Director of Centre National de Développement de l’Informatique (CENADI). He is currently a Senior Consultant and an International Expert in Information Management Systems. He has published more than forty papers in international conferences, journals and books. He is also a reviewer in some international journals.

**Dr. Josiane ETANG** is a Senior Malariologist and Medical Entomologist with extensive experience in research and training. She is member of several professional bodies: Central African Roll Back Malaria Network, African Network on Vector Resistance to insecticides, West and Central African Network on Anopheles Biology and Control, Cameroonian Society of Parasitology, International Society for Health Research and Training. She joined The Organisation de Coordination pour la lutte contre les Endémies en Afique Centrale (OCEAC) as a researcher in 2006 and subsequently she received an appointment as the sub regional Malaria Programme Coordinator in 2011. Furthermore, she has been appointed in the Faculty of Medecine and Pharmaceutical Sciences (University of Douala) as assistant lecturer in 2007 and after as senior lecturer in 2010. She holds PHD and Master’s of Science degrees in Parasitology, from the University of Yaounde1, Cameroon. She has published 27 articles in international peer-reviewed scientific journal.