

Semantic search engine for geographic data

Claudia Pegoraro¹, Mauro Velluto²

¹ CSI Piemonte, Sistemi Territoriali e Ambientali,
Corso Unione Sovietica 216, 10134 Torino, ITALY
tel. +39 011.316.9093 - Fax +39 011.316.8977 e-mail: claudia.pegoraro@csi.it

² CSI Piemonte, Sistemi Territoriali e Ambientali,
corso Tazzoli 215/12B, 10137 Torino, ITALY
tel. +39 011.432.6721 – Fax +39 011.740001 e-mail: mauro.velluto@csi.it

Abstract. A ‘useful’ knowledge of a data set, referred to a generic domain, allows finding, organizing and integrating data, without considering the specific context that justified their acquisition and making them available for transversal applications. The paper refers to the great amount of geographic data managed by CSI Piedmont and it describes the conceptual model proposed to improve their usefulness. It also examines in details the reasons that lead to the choice of grounding the model on a foundational ontology, providing both the description of the geometrical-geographical feature of territorial elements and their role in specific disciplines. Moreover it represents the base to develop a Semantic engine for geographical data (some operational examples are schematically described).

Keywords: Geographic Data, Foundational Ontology, Semantic, Search Engine, Civil Protection, Transports.

1.Introduction

A ‘useful’ knowledge of a fraction of reality should allow to gather, organize and integrate its data through its content without taking in account a specific discipline or the context that brought to the acquisition of the data.

CSI Piemonte (Consortium for Information System), is a consortium, gathering local governments, that supplies information services to its members. Among these services a major duty is the production and the management of a great amount of geographical data that represent a territorial knowledge which is much more detailed than ‘useful’. Using these data for purposes different from the ones they were acquired for, turns out to be quite difficult due to many factors: non homogeneous technical features, different production aims, last update and quality level. Besides all this, data are often stored in database accessible only by specific applications.

Our goal is therefore to make more useful our territorial knowledge by making it accessible to all those applications that are different from the ones the data was

originally acquired for. The path we have chosen to achieve this goal goes through the definition and the building of a rigorous and validated ontology, the core of which is a computational ontology that describes the geometric-geographical nature of territorial elements. It is possible to integrate this ontology with other ontological schemas that specify the meaning of elements in the different territorial disciplines.

A semantic search engine is the final result, but it's also, somehow, the very beginning of the work, since its functional requirements partially affect the decision-making over the upper ontology.

Therefore the main objective of this work is to spot a practical solution to optimize the use of this large amount of geographical information managed by Csi Piemonte.

The title of the article shows the final result that we want to achieve, a result that requires a preliminary work to define a strong ontology, validated and as universal as possible.

We have said that our final goal is also a starting point, meaning that, without considering the practical purpose of all this and the specific requirements it is asked to address, it would be impossible to build such a transversal and multi-thematic ontology. There is also another reason for such an approach: the will to check all the benefits coming from the ontological approach in comparison to more traditional techniques of conceptual modelling (ER, UML) that are used and well known by Csi Piemonte [1]

In paragraph 2 we will describe some aspects of our domain of knowledge: the geographical data regarding the Piedmont territory. Starting from the state of the art, we will pinpoint our aims in developing a semantic search engine, and the way to achieve them: the development of a conceptual model. In paragraph 3 we will introduce the four fundamental elements of the model, that will be discussed further in paragraph 4, taking in account both our vision and the theoretical reference models adopted. In paragraph 5 we will give the contents, to the four foundational elements introduced in the previous part. This is the decisional core of the work, since all the subsequent steps and results will be conditioned by these assumptions. The following sections provide some examples and applications and then the final considerations and the new aims we start to imagine at the present work stage.

2. Working within GIS data domain

As previously said, CSI Piemonte manages since many years, on behalf of public administration, a great amount of geographic data that, considered as a whole, provide a detailed territorial description of Piedmont region. Nowadays the instrument that allows users to orientate among these data is SITAD (Sistema Informativo Territoriale ed Ambientale Diffuso), the web GIS catalogue that allows data research through categories or lessical queries. The search can be refined if the users already know information like the owner of the data, the geographical extension and so on.

Therefore we have a worthy amount of information organized in a catalogue that allows users to do two operations:

- to get a panoramic view of the type of available data through categories exploration;
- to find a specific data layer (this operation can take a long time depending on the amount of details concerning the categorization known by the user).

For example: CSI manages data concerning the regional road network, (e.g. stored in shapefiles or SDE coverages): within SITAD we can find the related information, summarized in metadata forms providing a data type properties list (accomplishing ISO standard 19115):

- Title
- Description
- Issue
- Keywords
- Date (creation, update)
- Responsible
- Origin-description
- Distribution
- Coordinate Reference System
- Geographic coverage
- Temporal coverage (date of beginning and finishing of validity)
- Info on metadata themselves (metametadata)

Let's suppose we have two users, **A** and **B**:

- **A** uses SITAD because he needs the road network for a certain purpose; the time required to recover the related information will depend on the number of information he knows before starting the search (eg: part of the title, issue, keywords, etc.).
- **B** finds out, exploring the catalogue, the existence of the road network data and realizes he might use it for a certain purpose.

Simplifying:

- **A** knows which data he needs → he looks for it in SITAD → he finds the data.
- **B** enters SITAD → he finds out by chance that there are interesting data → he realizes he could use the data for his purpose.

Now let's try to change our perspective. Let's consider a **C** user that is not looking for a particular data, but has a complex need: let's imagine he has to produce risk indexes for a programming document for Civil Protection activities. If it was possible he would ask SITAD: "Which data, considering all available information, could be useful to calculate risk indexes?".

If SITAD was able to understand **C**'s question, and had some knowledge on the issue "Civil Protection" (and, in general, on GIS) he would probably answer that there is a calculation formula, used in many studies, that says that the Risk (**R**) is the product of the Probability that a certain event occurs (**P**), of the susceptibility (**S**), meaning the predisposition of a certain territory to suffer a damage, and of the suffered damage itself (**D**)

$$\mathbf{R} = \mathbf{P} * \mathbf{S} * \mathbf{D}$$

Considering this, almost all the available data could be useful to define **R**, and among all these SITAD would certainly extract the road network, specifying that roads may be, in consideration of the particular territorial context:

- a target element, if is located in the impact area of a disaster: it makes the **D** factor grow;

- a resource, if it is out of the impact area (eg: useful to transport supplies in a town interested by a landslide): it makes the **S** factor decrease.

Finally which instruments does SITAD need to answer **C**'s question? Common sense shows that SITAD isn't able to understand the question at the moment, but what does this exactly mean?

To make it simple, **C** user and SITAD have to speak the same language in order to understand each other. This means that, considered all the geographic data managed by CSI as the common domain of knowledge, the two ones should be essentially able to share the lexicon and the semantics:

- the **lexicon**: the set of terms to define the overall domain made of several thematic areas;
- the **semantics**: the meaning (or meanings) given to each term.

On the technical side we would also need to define a multi-disciplinary conceptual model within the SITAD, to be formalized and codified in order to be shared with **C**. Once this is done **C** will be able to ask complex questions getting back meaningful answers (exactly what happens with two people speaking the same language). Within the considered knowledge domain, made up by the overall set of territorial data, we suppose to need an integrated conceptual model, to formalise the meaning of data both if we consider the general cartographic aspects to represent them in a digital map, but also their roles within every specific thematic discipline. [3] [4] [5] [6] [8]

Getting back and focusing on our road network example, a robust conceptual model, formalized through an ontology, should be able to define:

A. the geometric/cartographic meaning of a road element, that can be represented by one or a set of lines or curves, with its own geometric and geographic attributes. This data will be related with other territorial objects through:

- Semantic relationships
- Spatial & topological relationships

B. The meaning of a road element according to one or more specific thematic disciplines and to one or more validated lexical sources.

As we said before, as far as Civil Protection is concerned, a road element can represent a risk source, the target of a calamitous event or a resource and in each case it provides an input to calculate the risk index. If we consider the same road element from the point of view of another discipline, such as territorial or urban planning, it may assume different roles.

3. Designing the ontology

To cover the conceptual modelling requirements and describe the common knowledge domain efficiently, we will consider four basic constituents [9]

- The **lexicon**: the complete set of terms to define the overall domain
- The **semantics**: the meaning (or meanings) given to each term depending on different sources and different thematic contexts.
- The **ontology schemas** or **ontological maps**: if we consider the two first elements as the content reference, the ontology itself represents the content for the conceptual model and it maps the objects and the set of hierarchical and logical relationships that describe connections among them.

- The **computational ontologies**: introduced to reach a reliable ontology: they complete the whole conceptual architecture through primitive definitions mapping abstract lexical meanings in the upper ontology level.

So our purpose is the building of a foundational Geographic Information (GI) ontology: the modelling rules will be set apart of every specific thematic context and furthermore will be inclined to avoid the building of a new, original set of computational ontologies; on the contrary they will adopt a validated one, widely acknowledged by the scientific community.

This is also going to save the principle saying that operative ontological models may exist in many applicative situations and be interoperable if a high level general, universal ontology is supporting them.

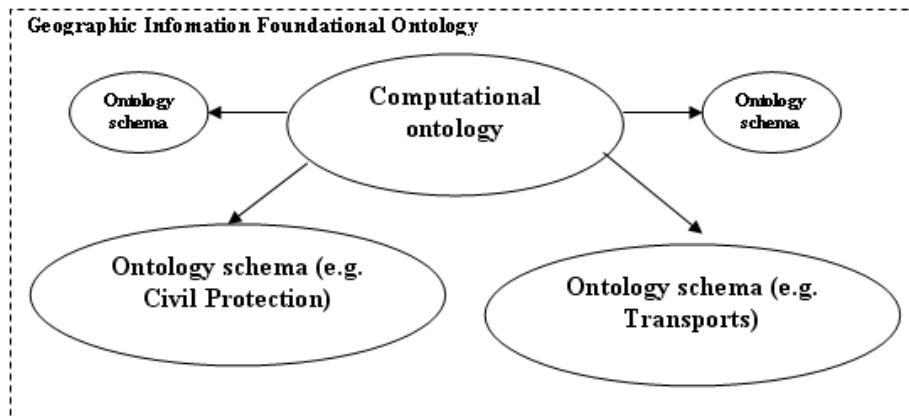


Fig.1 Geographic Information foundational Ontology schema

4. Building the ontology

Building each one of the fundamental elements of our conceptual model, we've dedicated a great care in choosing references that were recognized at least at national level:

- For the **lexicon** and the **semantics** of the overall and generic terminology it has been chosen the Treccani Italian Language Dictionary; it has been also used to check and validate input coming from the GI specifications adopted (see below). The national laws and regulations in every single thematic area were put beside the Dictionary to improve detail level of definitions (e.g. for the Transportation field: the Italian Transportation Code and many other technical and scientific texts).
- For the **ontology schemas** the GI specifications of INTESA GIS [13] [14] for the developing of our national Spatial Data Infrastructure were considered as the basis to build the ontological map. For the preliminary experimental activity the focus was initially on the stratum dedicated to the Transportation field, the first to be translated as ontology schema. A recall was also made to European projects involving the Transportation topics again (see EURORoads and the guidelines of the European Directive INSPIRE [7] [11][12]), to furthermore enlarge the validation of the modelling activity through models comparison. As far as Civil

Protection is concerned, we've also considered an ontological schema defined by the Joint Research Center (JRC) of Ispra, ordered by Piedmont Region as part of the activities to elaborate the Regional Program for Risk Forecast and Prevention: the schema gives a structure and an organization of the territorial data organic and functional to Civil Protection activities.

- For the **computational ontologies**, it was chosen the ISO international standards (ISO TC 211 series – Geomatics and other series dedicated to many thematic issues) in order to define the primitives of the upper level of the ontology: it is coherent with INTESA GIS metamodel, due to the common adoption of a GeoUML structure.

4.1 Computational Ontology structure: GeoUML

GeoUML is an UML based language, since it includes all the constructs of UML for class diagram specification and the OCL language (“Object Constraint Language”) for integrity constraints specification; but, most of all, it is based on the ISO TC 211 standards, since it specializes the classes for geometry representation proposed by the ISO 19107 document (“Spatial Schema”) and adopts the General Feature Model approach and the rules for conformant schemas as suggested by the ISO 19109 document (“Rules for application schema”). [2]

The GeoUML model contains the following components:

- A set of predefined UML classes for the representation of the spatial component of geographical information. These classes are a specialization of the UML classes proposed by ISO 19107 document “Spatial Schema”, called Geometric Types.
- A constraint template for the specification of spatial integrity constraints based on a reference set of topological relations called Topological Constraints. The template is based on a OCL formula skeleton and use the Relate functions of the Spatial Schema.
- A constraint template for the specification of spatial integrity constraints among objects with common structure (aggregate, complex, subcomplex and primitive represented in the geometric classes GM_Aggregate, GM_Complex and GM_Primitive of the Spatial Schema). These constraints are called Structural Constraints.
- A set of predefined schema skeletons representing widely used structures in the description of geographical information like: segmented attributes, layers, partitions.

4.2 Ontology language: OWL-FULL

Few words on the ontology language and the tool adopted for the experimentation. OWL-Full is the most expressive OWL sub-language. It is used in situations where very high expressiveness is more important than being able to guarantee computational completeness of the language. It was adopted and used for this study within the tool Protégé, one of the most popular free ontology tools. Automated reasoning on OWL-Full ontologies will be provided by an inference engine (e.g. Jena and Racer are two of the most popular inference engines).

Since a new (created) ontology has always to deal with real data, in our case we must check one last condition: the mapping of the INTESA GIS objects (presented in the ontological map) with the SITAD metadata forms actually available in the catalogue .

5. A practical example

Let us try to examine an ontology schema extract, including the upper semantics for a generic geometric object (GM_Object in Figure 2) derived from the GeoUML metamodel and the specific ones for two different disciplines, (both related to the object “road network”): Transports and Civil Protection.

N.b.: this approach does not exclude the future enlargement to any kind of objects: in fact the geometric one was only one of the possible branches to be developed within this model.

5.1 Ontology schemas

In figure 2 we can see a simplified schema of logical relationships among classes of elements belonging to our domain. The schema allows to identify a concrete data (the road network) navigating through a model that, starting from the most general and abstract class (“object”), bounds classes that are on a lower hierarchic level through relationships known by everyone familiar with territorial information systems and formalized with GeoUML. For example the class “GM_Points” belongs to, and inherits characteristics and semantics from the class “primitive” that, on its side, belongs to the class “geometrical object”. The model goes from general to particular until it reaches the element “road network”.

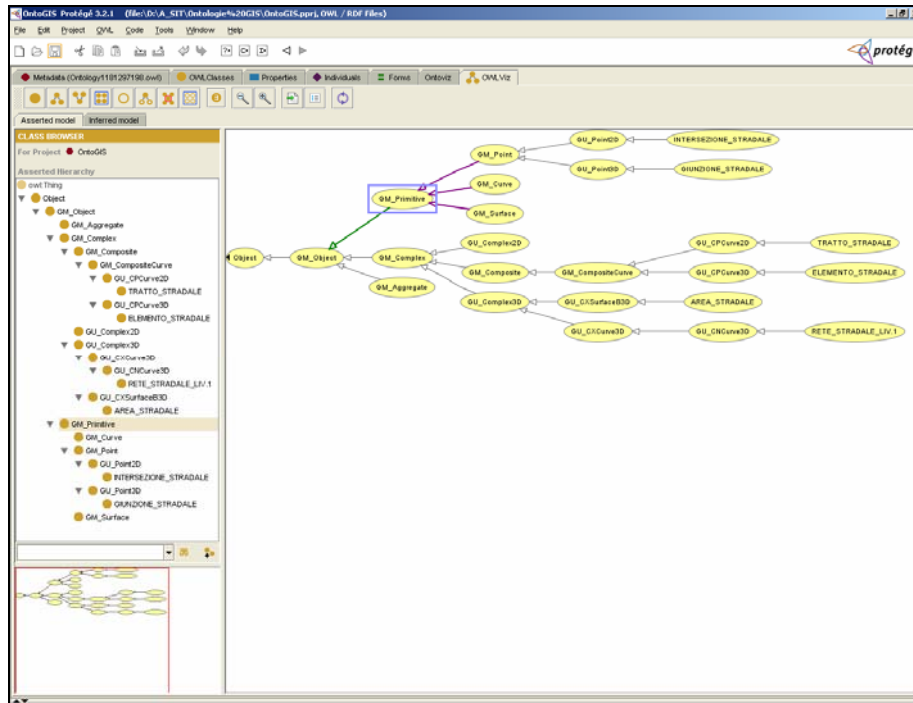


Fig.2 Partial view of the upper ontology scheme

In figure 3 (extracted from INTESA GIS specifications, document 1n 1007_4 - “Specifiche per la realizzazione dei Data Base Topografici di interesse generale, Specifiche di contenuto: Lo Schema concettuale delle Specifiche di contenuto in UML”

alias “Specifications for producing Topographic Data Bases of general concern, content Specifications: UML Conceptual schema”) we can see a more detailed UML schema containing logical relationships among a few classes from the Transportation domain; this schema (and some other ones) is going to be translated in the ontology formalization in the next future during our experimentation activity.

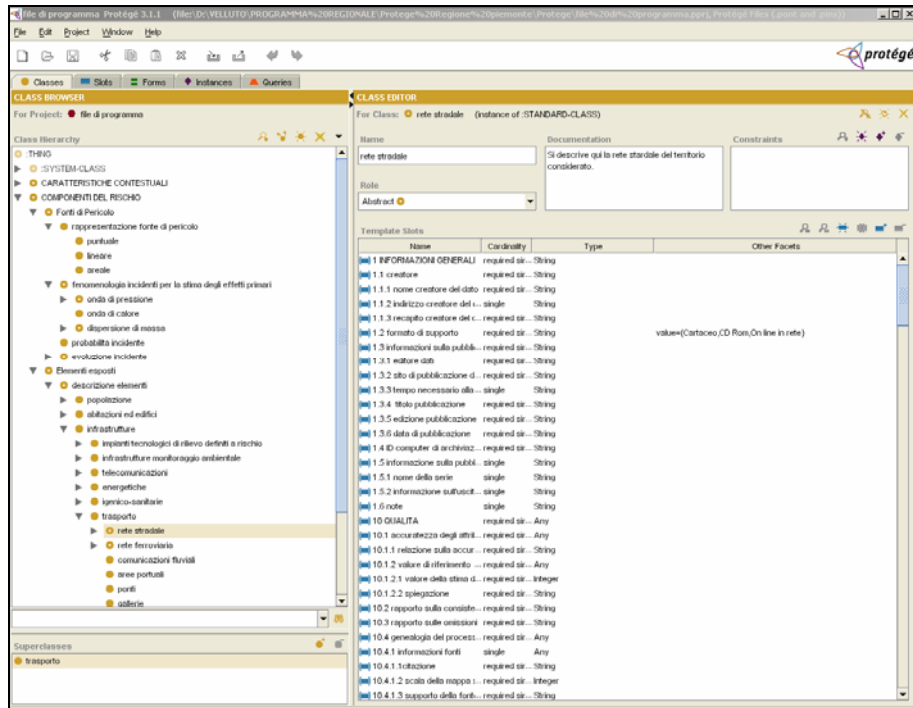




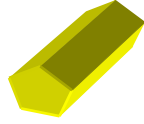



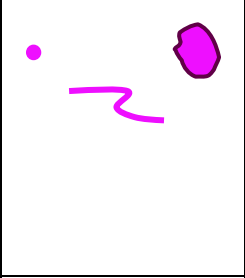
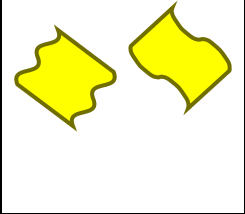
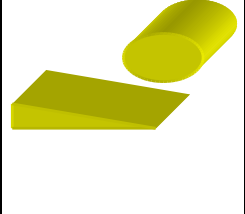


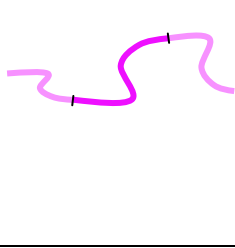
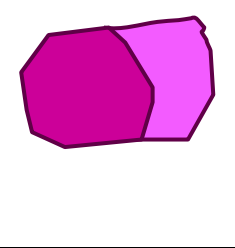
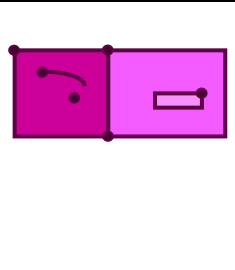
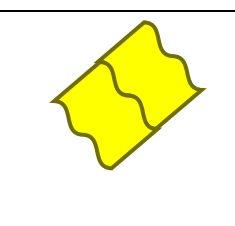
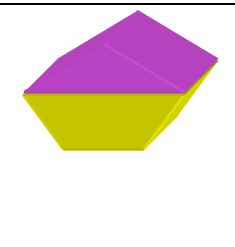
Fig.4 Class structure (from the JRC ontology)

5.2 Semantics examples

As said before, apart from defining logical relationships, semantics specifies the meaning (or meanings) of classes or of an attribute or a single element of the domain. In our conceptual model, upper level classes (the more general ones, the computational ontologies) were characterized by a precise semantics, related to their geometric nature, included in ISO 19107 document “Spatial Schema” that fix them in a unambiguous way. GeoUML took those definitions and built on them the simplified hierarchy we can see in figure 2. The simplified semantic interpretation of the ISO 19107 “Spatial Schema” definitions was given in Chart 1. So we have to deal with semantics with different level of detail and coming from different sources, but always on the same issues. An idea to develop may be to keep all these thoughtful semantics in our catalogue and to let them anyway accessible.

		Graphic example	Name	Description (SEMANTICS)	
GM_Objects	GM_PRIMITIVE	2D Euclidean space		Point	Basic geographic object made of a single isolated point
				Curve	Basic geographic object made of a single line or curve tract
				Polygon	Basic geographic object made of a single free shape polygon
		3D Euclidean space		Surface	Basic geographic object made of a single free shape surface; the surface may be a plane surface or not, positioned in any way in 3D space
				Solid	Basic geographic object made of a single free shape solid, positioned in any way in 3D space
	GM_AGGREGATE	2D Euclidean space		MultiPoint	Geographic object made of many (more than one) isolated points. The individual components are allowed to be elementary objects of type Point (note this is not mandatory: only if necessary!)
				MultiCurve	Geographic object made of many (more than one) separated lines or curves: no contiguity allowed. The individual components are allowed to be elementary objects of type Curve (note this is not mandatory: only if necessary!)
				MultiPolygon	Geographic object made of many (more than one) separated polygons: no contiguity allowed. The individual components are allowed to be elementary objects of type Polygon (note this is not mandatory: only if necessary!)

3D Euclidean space		Planar Mixed Aggregation	Geographic object made of many (more than one) separated 2D basic objects: Points, Curves, Polygons with no contiguity allowed. The individual components are allowed to be elementary objects of type Point or Curve or Polygon (note this is not mandatory: only if necessary!)
		MultiSurface	Geographic object made of many (more than one) separated surfaces: no contiguity allowed. The individual components are allowed to be elementary objects of type Surface (note this is not mandatory: only if necessary!)
		MultiSolid	Geographic object made of many (more than one) separated solid objects: no contiguity allowed. The individual components are allowed to be elementary objects of type Solid (note this is not mandatory: only if necessary!)
		Free Mixed Aggregation	Geographic object made of many (more than one) separated 2D and 3D basic objects: Points, Curves, Polygons, Surfaces, Solids with no contiguity allowed. The individual components are allowed to be elementary objects of type Point or Curve or Polygon or Surface or Solid (note this is not mandatory: only if necessary!)

		Graphic example	Name	Description (SEMANTICS)	
GEOMETRY & TOPOLOGY aspects	GM_COMPOSITE	2D Euclidean space		Composite Curve	Geographic object made of many (more than one) individual lines or curves which contiguity is requested. The individual components are allowed to be elementary objects of type Curve (note this is not mandatory: only if necessary!)
				Composite Polygon	Geographic object made of many (more than one) individual polygons which contiguity is requested. The individual components are allowed to be elementary objects of type Polygon (note this is not mandatory: only if necessary!)
				Planar Mixed Composition	Geographic object made of many (more than one) individual 2D basic objects: Points, Curves, Polygons which contiguity is requested. The individual components are allowed to be elementary objects of type Point or Curve or Polygon (note this is not mandatory: only if necessary!)
	3D Euclidean space		Composite Surface	Geographic object made of many (more than one) individual surfaces which contiguity is requested. The individual components are allowed to be elementary objects of type Surface (note this is not mandatory: only if necessary!)	
			Composite Solid	Geographic object made of many (more than one) individual solid objects which contiguity is requested. The individual components are allowed to be elementary objects of type Solid (note this is not mandatory: only if necessary!)	

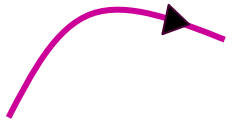
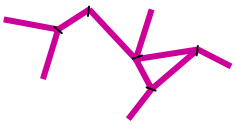

			Free Mixed Composition	Geographic object made of many (more than one) individual 2D and 3D basic objects: Points, Curves, Polygons, Surfaces, Solids which contiguity is requested. The individual components are allowed to be elementary objects of type Point or Curve or Polygon or Surface or Solid (note this is not mandatory: only if necessary!)
			Orientable Curve	Geographic object made of a single line or curve tract with an orientation verse declared on it
			Graph	Geographic object made of many (more than one) individual Curves or Composite Curves which contiguity is requested and multiple connection is allowed. The individual components are requested to be elementary objects of type Curve or Composite Curve.
			Edge-Node Graph	Geographic object made of many (more than one) individual Curves or Composite Curves which contiguity is requested, multiple connection is allowed and points function as nodes. The individual components are requested to be elementary objects of type Points, Curve or Composite Curve

Chart 1 Semantics for the upper level classes of the conceptual model

5.3 A real application

Let's consider once again our foundational ontology and let's represent part of it in a simplified way in order to make clear how, beginning with the class 'road network' is possible to integrate and use information using all the three ontological schemes presented in paragraph 5.1.

through the proposed conceptual model, user C will be able to gather two kinds of simple information:

- **Explicit:** looking for data about crossroads he will be able to find the right metadata form even if it is called “road intersection”
- **Implicit:** he can find out which ones, among the existent data over the Piedmont, could be useful to calculate risk indexes for Civil Protection.. As an example, road network might be useful to calculate the hydro-geological/landslides risk index: depending on the topological relationship that bind a single element of the road network to the impact area (also this informative level is managed), it can be considered as a target or as a resource.

6. Conclusions

In the exposed work we have focused on the need of building a rigorous conceptual model binding together validated and structured ontologies: starting from the definition of the final aim we have later justified the choice of the reference standards. The example in the previous paragraph shows how such a model can be strategic in order to build a transversal application: a semantic search engine.

In order to realize the implementation we'll have to import (for example using Protege) ontologies mentioned within a common schema expressed in OWL-FULL. Such a schema will be later refined through the definition of all semantics and inference rules. The use of an inference engine (such as Jena) will then allow us to check the overall value of the model before moving on to the development of an interface for data search.

6.1 Perspectives

From the application point of view, the perspective is to develop other transversal applications such as:

- A **‘reasoning’ vocabulary** that, exploring the model, provides information, about GIS and specific thematic disciplines, characterized by different complexity and level of detail. It could be useful both as a support to experienced users and as a basic didactical tool.
- One or more **GIS applications** that will no longer query a Geographic database through static and predetermined queries, but through dynamic and “deductive” ones interacting with the conceptual model. In this case the role of user C would be accomplished by the GIS application and the answer to complex questions would be a pleasant map and not only a list of data and information.

Under the conceptual model profile, one interesting perspective is to use an approach quite similar to other knowledge domains even if not directly connected to territorial disciplines. There will be the need of checking whether it's possible to integrate new ontologies with the one proposed in the article, but, even if that will turn out to be not possible, we believe that the definition of any whatsoever new conceptual model strongly needs a rigorous approach and must accomplish validation requirements.

References

1. Batini C., Ceri S., Navathe S., Conceptual Database Design: an Entity Relationship Approach, Redwood City, The Benjamin/Cummings Publishings Company, 1992
2. Belussi A., Negri M., Pelagatti G., GeoUML: an ISO TC 211 compatible data model for the conceptual design of geographical databases - Internal Report n.2004.21 Dipartimento di Elettronica e Informazione, Politecnico di Milano, 2004
3. Caglioni M., Rabino G. A., Theoretical approach to urban ontology: a contribution from urban system analysis, Workshop of COST Action C21 Towntology, 2006
4. Gruber T. R., A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993
5. Gruber T. R., 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 1995, pp. 907-928
6. Gruber T. R., What is an Ontology, <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>
7. INSPIRE, Infrastructure for SPatial InfoRmation in Europe - <http://inspire.jrc.it/home.html>
8. Keita A., Laurini R., Roussey C., Zimmerman M., Towards an Ontology for Urban Planning: The Towntology Project – 24th UDMS Symposium, Chioggia, October 2004
9. Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., WonderWeb Deliverable D18, Ontology Library (final) - from IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web
10. V.V.A.A., Deliverable D6.3 - Final specification of Road Network Information Model – from EURORoads Specifications, 2006
11. V.V.A.A., Geographic information — Standards, specifications, technical reports and guidelines, required to implement Spatial Data Infrastructure – Joint Research Centre, Institute for Environment and Sustainability , 2005
12. V.V.A.A., Requirements for the Definition of the INSPIRE Implementing Rules for Spatial Data Specifications and Harmonisation – Joint Research Centre, Institute for Environment and Sustainability , 2005
13. V.V.A.A., Specifiche per la realizzazione dei Data Base Topografici di interesse generale: Il Catalogo degli Oggetti – from INTESA GIS Specifications, 2006
14. V.V.A.A., Specifiche per la realizzazione dei Data Base Topografici di interesse generale: Lo schema in GeoUML delle Specifiche di Contenuto – from INTESA GIS Specifications, 2006