Ontologies Models Cohesiveness: A First Assessment of Integration

Mercedes Vitturini[1] and Pablo R. Fillottrani[1,2]

[1]LISSI - Laboratorio de I+D en Ingeniería de Software y Sistemas de Información, Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur
[2]Comisión de Investigaciones Científicas de la Provnicia de Buenos Aires

***Abstract*—Multiple ontologies over shared domain has been produced in the geographic ontology-design field for GIS environments. In geneal, each GIS solution has its own data model or application ontology. Ontology model mapping could provide a common language from which several systems could exchange information in a semantic form. In this paper we present a novel technique that collaborates with the task of analyzing the degree of cohesion between ontologies and servers in order to anticipate the quality resulting from the integration process.**
Keywords: Ontology Integration, Interoperability, GIS, Ontology

I. Introduction

Geographic Information (GI) accessible on line covers a range of systems types, such as Geographic Information Systems (GIS), Spatial Data Infrastructures (SDI), and Location Based Mobile Systems (LBMS). Private companies and national and international research and education organizations are some of their major customers. Generally, GI is unwieldy, has a complex structure and usually is distributed by *theme* over different servers. Many new users consider attractive the possibility of use GI services. If a new GI service is needed, an extra cost to consider is the acquisition of geographic data when unavailable. Sometimes it occurs that the GI data on a topic already exists in some previous development, for instance, information about "routes and roads". The possibility of sharing information and services means lower costs and start-up times, as well as improvement of information reliability. In the last few years, several software development problems have been faced with the need of sharing and reusing knowledge acquired for a specific domain. This is accomplished by the Semantic Web and it is linked to the notion of interoperability [Gua98], [ISO94], [OGC94]. The goal is to have an unambiguous knowledge of the Web that can be interpreted by automated agents. In particular, there is a need of a Geospatial Semantic Web [Ege02] on a framework comprising various thematic spatial ontologies. The design of some kind of solutions to GI heterogeneity problems is needed in order to share GI. In this way, data could be processed and interpreted by remote systems. Ontology semantic integration is based on several ontologies working as mediators in the communication between systems. For a successful mediation, a semantic mapping between ontology models is required. This task will be effective as long as the concepts of different ontologies are really comparable. In this paper we propose a simple technique based on relationships of concept sets that could help to anticipate the effectiveness of this communication process.

II. Heterogeneity in Geographic Information

For years, each new GIS development defined its own models of storage and visualization for spatial data, in addition to their conceptual data models. GI format diversity involves interoperability problems between GIS's.

A. Context

The GI concept encompasses information including spatially referenced data, i.e. linked to one or several points on the surface. GI is characterized by its inherently complex structure and volume. A *geographic data* is an abstraction that represents a real world object, such as a route, a building, an agricultural area, etc., which has a digital representation. Each object is called geographic *feature* [LGMR05]. A geographic feature is unique and distinguishable. A *feature type* is the abstraction that represents sets of geographic features of the same class. A feature type encompasses attributes and relations that model real phenomena. Attributes of a feature type are arranged into *thematic attributes* and *spatial attributes*. The spatial component keeps reference to the Earth's surface. Thematic components maintain the description characterizing each entity. Eventually, a geographic model also includes the definition of geometrical and/or topological relationships between features. In a geographic object, metric properties include length and area -depending on the dimension of the object- and metric relations between objects

Heterogeneity	Application A_1	Application A_2
Syntactic	A_1 represents the areas according to Bahía Blanca population density under the spatial vector model.	A_2 represents the areas according to Bahía Blanca population density under the spatial raster model.
Structural	A_1 represents the areas according to Bahía Blanca population with details about the distribution of public services.	A_2 represents the areas according to Bahía Blanca population with details about types of constructions -buildings, private neighborhoods, etc.-.
Semantic	A_1 represents the areas according to Bahía Blanca population density. The unit of measurement used is number of inhabitants.	A_2 represents the areas according to Bahía Blanca population density. The unit of measurement used is number of family groups.

TABLE I
EXAMPLES OF SYNTACTIC, STRUCTURAL, AND SEMANTIC HETEROGENEITY

such as distance and orientation. Topology refers to properties like proximity, adjacency, inclusion, and connectivity that remain invariant to morphological changes of scale or projection.

B. Heterogeneity Levels in Geographic Data

In any two given representations of a geographic problem, we will distinguish the following types of heterogeneities [LGMR05], [WVV+01]:

- *Syntactic Heterogeneity*: for a single phenomenon each solution provides different formats and space representation models -vector or mosaic-, and/or different coordinate representation systems.
- *Structural Heterogeneity*: refers to the "form" that each solution chooses in order to represent the same phenomenon. Many differences are expected to exist in terms of structure between models.
- *Semantic Heterogeneity*: it occurs when distinct solutions interpret different meanings for the same phenomenon.

Table I illustrates all these types of heterogeneities. The solution to GI heterogeneity problems encourages research in the field of Computer Sciences.

III. PROPOSALS ON GEOGRAPHIC INFORMATION INTEGRATION

Research work to make progress towards GI integration is addressed in two different ways. On the one hand, there is research that defines standards that normalize representation models for spatial data. On the other hand, there is research on semantic difference solutions that are generally linked to the definition of ontologies that provide formal specifications and tools for automatic integration. Defining integration rules is only possible if the meaning of data is known.

A. Standards in Geographic Information

The international standards for geographic data and services are primarily concerned with the Open Geospatial Consortium (OGC) [OGC94] and the Technical Committee of Standardization on Geomatics and Geographic Information ISO/TC211 [ISO94]. OGC is an international consortium. Its participants represent business companies, government agencies, and universities. It has a consensus process to develop interface specifications applicable to open source geo processing systems. OGC solutions are referred to as *Open GIS Specifications* and provide interoperable solutions to make the GI "geo-available". OGC mission is to lead to development promoting the use of architectures that allow for the integration of geographic applications.

Meanwhile, the International Standards Organization (ISO) established the Technical Committee for Standardization in Geomatics and Geographic Information ISO/TC211 to be responsible for defining reference standards for digital GI and for the transfer of data and services. The ISO 19100 family is a set of standards related to geographic features. These regulations deal with methods, tools, and services for managing, acquiring, processing, analyzing, accessing, presenting, and transferring digital GI among different users, systems, and locations.

OGC members also participate in ISO/TC211 through the Joint Consultative Council ISO/TC211-OGC. Its mission is to coordinate the efforts of both organizations and to ensure a single standard reference.

B. Ontology Models

Ontologies unify the interpretation of concepts and terms so that such interpretation can be unique [FEA+02]. This is true among people and also when automatic agents are involved in machine communication. Personal communications can solve semantic heterogeneity caused by different conceptualizations, terminology, context or incomplete information. For example, the generalization/specialization relationship between elements is clearly understood by most of people. However, this relationship is not trivial for many search algorithms based on finding matching terms in schemas and data. The mission of ontology is to provide the formal specification of concepts and their relationships. Figure 1 shows a simple case of ontological concept specification

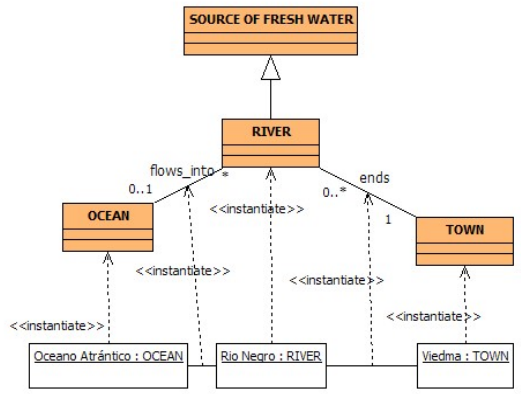

Fig. 1. Instantiation Feature-Ontology

and its relationships. The notation used is proposed by UML. In the example there are classes (color) which define the common properties of the elements of the same type, and instances (white) representing a particular concept, occurrence or instance. You can see a hierarchical relationship between classes SOURCE OF FRESH WATER and RIVER specified by *is_a* distinguished relation. Other relations in the example are: *flow_into*, which specifies that instances of RIVER lead to instances of OCEAN and *ends* which represents a TOWN where a RIVER ends. The link between instances and their respective class or relation is represented by the stereotyped dependency $<< \ instantiate \ >>$. These definitions make it possible to achieve the following basic conclusions:

1) 'Río Negro' is a RIVER.
2) 'Atlántico' is an OCEAN.
3) 'Viedma' is a TOWN.
4) 'Río Negro' flows into 'Atlántico'.
5) 'Río Negro' ends in 'Viedma'.

And these more elaborated implicit conclusions:

1) 'Río Negro' is a SOURCE OF FRESH WATER.
2) 'Viedma' is near the 'Océano Atlántico' OCEAN.
3) 'Viedma' has a SOURCE OF FRESH WATER.

A benefit of having this explicit representation for instances and their binding model is that an automated agent could reach to these same conclusions, as if it could understand or reason.

IV. ONTOLOGY-BASED ARCHITECTURE INTEGRATION MODEL

As stated, to find and recover efficiently distributed heterogeneous GI is a key factor. Standards promote interoperability and classification for GI by catalogs. However, difficulties caused by semantic heterogeneities are still a challenge in integrating distributed open environment GI.

In the field of ontology, there are different ontologies built for various application domains. They vary in the level of detail they express. Ontologies can be organized according to their degree of generality as follows [Buc09], [WVV+01]:

- *Generic Ontology (Top-Level)*: captures the general purpose knowledge, regardless of any particular domain, such as space, time, event, action, etc. It is expected that these ontologies will be adopted by a large community of users.
- *Domain Ontology and Task Ontology*: define the particular knowledge of a domain (for example, medicine, geography, etc.) or a specific activity (for example, trade), describing their vocabulary through the specialization of the terms introduced in the high-level ontology.
- *Application Ontology (Low-Level)*: captures the knowledge needed from an individual system or application. It describes concepts that depend on both the domain and activity ontology, that are often specializations of the two previous kinds of ontology.

The options for ontology-based semantic integration systems are arranged into different styles [CSH06], [WVV+01]. One style considers a single ontology shared by all applications. Another defines multiple ontologies along with integration functions between pairs. A more flexible option is to combine the two previous styles. The latter proposal of integration, based on *hybrid ontology*, establishes a *domain ontology* (DO) shared by a community of use that provides the definition of its basic terms (primitive). Hybrid ontology assumes that the common semantics of its primitives is known and understood by the community. Independently, each supplier is free to define his or her own $OGIS_i$ *application ontology*. Furthermore, the data model of each GIS solution plays the role of application ontology. Besides, the communication interface or "mapping" between DO ontology and $OGIS_i$ ontology should be established. This kind of semantic integration provides a flexible framework that respects application ontology and complies with every need, keeping the various $OGIS_i$ ontologies comparable [Buc09], something crucial when making semantic searches or requiring information integration services.

V. COHESION BETWEEN ONTOLOGICAL MODELS

In compliance with the integration style based on hybrid ontology semantics, each application is free to use its own application ontology or $OGIS_i$ data model. An $OGIS_i$ could eventually be shared by more than one application, as it is the case of distributed GIS that share the data model. In the following generality level, DO ontology is defined. In the particular case of geographic applications, a DO ontology corresponds to a theme such as "land use". This proposal for integration assumes the existence of consensus ODs. Thus, each GIS is responsible for formalizing its $OGIS_i$ application

scheme aligned to the GFM standard [ISO05] and to define the $m(OGIS_i) \rightarrow DO$ mapping function.

Thus, the problem of solving semantic heterogeneity among different GIS solutions turns into defining the correct $m(OGIS_i) \rightarrow DO$ mapping function, with the added advantage of having concepts formalized by an ontology. However, the effectiveness of the mapping, and thus the result of integration, depends on the degree of cohesion between the world shaped by DO and the world shaped by $OGIS_i$.

This work presents a technique used to measure the degree of interrelation or cohesiveness between DO ontology and $OGIS_i$ application ontology. In particular, we want to measure the level of cohesiveness between concepts defined in $OGIS_i$ and concepts defined by DO. In order to progress with rigor, we present the following definition [Vit09]: let C_{GIS} be a set of concepts defined by $OGIS_i$ ontology and C_{DO} the ontology concepts defined by DO ontology. It is possible to approximate the cohesiveness between the application model and the domain ontology by formalizing the following membership relations between sets:

1) $C_{GIS} \subset C_{DO} \wedge |C_{GIS}| \approx |C_{DO}|$, presents the situation of domain ontology with maximum coverage and high accuracy. This is the optimal relation between DO and $OGIS_i$ concepts. Domain ontology concepts cover the concepts required by the application. We can say that DO contains definitions and semantics close to the application problem. For example, suppose that C_{DO} defines the concept LOCATION while C_{GIS} provides the definition for a concept named CITY. In all instances, city in C_{GIS} is covered by the concepts LOCATION in C_{DO} and is hoped that its semantic definition is close.
2) $C_{GIS} \subset C_{DO} \wedge \neg(|C_{GIS}| \approx |C_{DO}|)$, represents a situation with high coverage but low precision. The relationship between concepts in domain ontology and the concepts in the GIS application can be defined as "good". In this case, DO also covers the concepts required by the application. However, the semantic content in DO is not as close to the semantic content required by the GIS, and the mapping shall be potentially less accurate. For example, C_{DO} has a definition for a concept SPECIES while C_{GIS} considers a definition for a concept NATIVE SPECIES. It is expected that the description for NATIVE SPECIES in $OGIS_i$ will be more refined than the characterization of the concept SPECIES provided by DO ontology.
3) $C_{DO} \subset C_{GIS}$. Concepts in DO do not cover the universe of concepts required by the $OGIS_i$ application. There are concepts defined by the GIS application for which there are no concepts to map in DO. The semantic relationship between $OGIS_i$ and DO is "not good". The greater the number of concepts excluded, the worse the cohesion ratio relation is. For example C_{GIS} contains a definition for SITE whereas C_{DO} defines LOCATION. There are site elements in C_{GIS} which do not map any location of C_{DO} and, thus, will be lost in the mapping process.
4) $\neg(C_{GIS} \subset C_{DO} \wedge C_{DO} \subset C_{GIS}) \wedge (C_{GIS} \cap C_{DO})$. Both ontologies contain concepts that do not have a correspondence in the other universe, although they share a subset of concepts. This is the worst relationship scenario among ontologies. For example, consider a C_{GIS} set which defines concepts such as SOURCE OF FRESH WATER, while C_{DO} defines concepts such as STREAM OF WATER. There are elements in C_{GIS} such as "reservoir" which do not map any concept in STREAM OF WATER in C_{DO}. Conversely, the semantic definition for the elements in DO, such as "seas" are not represented in the C_{GIS} of the GIS application.

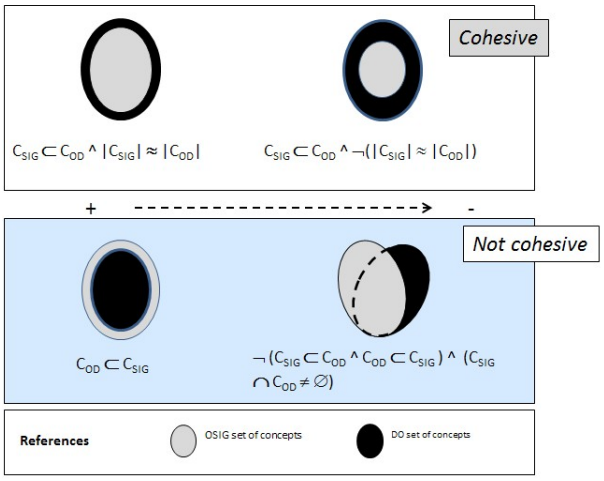

Fig. 2. Relationship between Domains and GIS Ontology Concepts

Figure 2 shows a graphic for the above items using traditional set representations. The kind of relationship between sets of concepts has an impact on the effectiveness of the mapping process from $OGIS_i$ to OD. We can state that the mapping function looses accuracy as we move away from the situation in the states in 1, being 4 the least desirable alternative.

An example that applies the above items is shown in Figure 3, which contain two partial views of possible solutions for theme "Tourist Information" and its ontological conceptual models. Model a) represents a DO about *Tourist Information Center*, model b) contains the $OGIS$ ontology for a Municipality which reports *Visitors Information*. Before

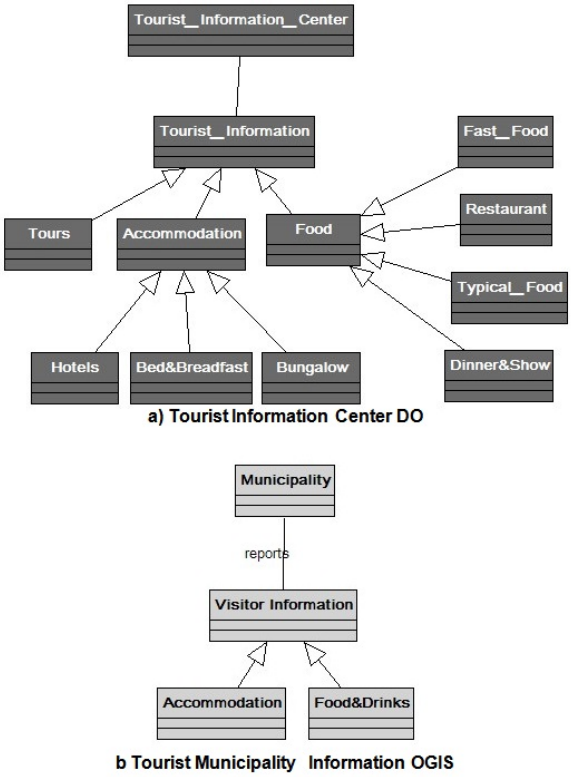

Fig. 3. DO & $OGIS$ Conceptual Model

$OGIS$ **Concept**	DO **Concept**	**Map Relation**
Municipality	Tourist Information Center	**(.4.) Bad**
reports	offers	**(.1.) Very Good**
Visitor Information	Tourist Information	**(.1.) Very Good**
Accommodation	Accommodation and its specializations: Hotels, Bungalows Bed&Bredfast	**(.2.) Good** *the map depends uppon the existence of a discriminator*
Food&Drinks	Food and its specializations: Restaurant, Typical Dinner&Show Fast Food	**(.3.) Not Good** *the map is partial and depends uppon the existence of a discriminator*

TABLE II
ANALYSIS OF RELATIONSHIP BETWEEN CONCEPTS

an automatic ontology mapping, we could apply the analysis of relationship between concepts present in both ontologies under our proposal. Results are shown in Table II.

As shown, in this case is difficult to ensure a good mapping relation between both models. Some possible solutions are: enrich de $OGIS$, if it is possible, and/or define manually mapping function. Surely the $OGIS$ and DO ontology were defined independently over a shared domain and requirements. If the case of new GIS application solution, the best option is to define an application model according to existing DO ontologies which ensure best results [Vit09].

VI. ANALYSIS OF THE PROPOSAL

In order to *semantic queries* and *ontologies reuse*, is needed to align application ontologies and domain ontologies contents. Generally, a domain expert must determine the correlation between concepts of these ontologies. Doing matching task manually is time-consuming, tedious, and error-risk process. To reduce this effort, many approaches to semi-automatically determine element correspondences were developed. As example, COMA++ (Combining Matchers) platform [Do06], present a generic and customizable systems for semi-automatic schema matching, allowing flexible combination of different match algorithms and a way for solving the match task in stages. On the same idea GEN-MAPPER (Generic Mapper) [Do06], present an approach for integrating heterogeneous web data sources using correspondences between objects. GEN-MAPPER explicitly captures existing relationships between objects to drive data integration and combine annotation knowledge from different sources. A generic schema is used to uniformly represent object data and correspondences.

Other research is iPrompt and AnchorPrompt [NM03], a suite of tools for managing multiple ontologies. This suite provides users with a framework for comparing, aligning, and merging ontologies, maintaining versions, translating between different formalisms with support to semi-automatic ontology merging: iPrompt is an interactive ontology merging tool that guides the user through the merging process, presenting suggestions for next steps and identifying inconsistencies and potential problems. AnchorPrompt uses a graph-based algorithm for finding correlations between concepts in different ontologies which provide additional information for iPrompt. It takes as input a set of pairs of related terms or anchors from the source ontologies. The requirement for “align concepts” between diverse sources, may mean align the schema or metadata, translating information from one structure to another, or may mean align data o content, when data gathering of diverse sources is needed.

A survey and a taxonomy that covers many of existing approaches of match partially automation is found in [RB01]. The work distinguish between schema-level and instance-level (instance data o schema level information), element-level and structure-level (simple attributes or complex schema structures), and language-based and constraint-based matchers (based on names or based on keys and relationships).

In general, all researches agree that it is not possible to fully automate the integration process and, at least in the phase of mapping definitions, the participation of domain experts is required. This work proposes a novel and simple technique based on the relations among sets of concepts to assessment the level of integration of two ontologies at

conceptual level. As described in the example developed in Figure 3 and Table II, the method includes identifying concepts in the ontology source -concepts and relationships- and classifying the mapping relationship (1 to 4) with regard to the concepts of the target ontology. As far as the target ontology covers the concepts of source ontology, it is expected that mapping to shared ontology shall be possible without loosing information. In our example, we see that between the two models there is a "very good" relationship with the *DO* ontology for the concepts "Visitors Information" and"reports"; a "good" relationship between "Accommodation" concepts, with a different and even impossible way of mapping the system used to classify them; and a "bad" and even impossible way of mapping for instances of the remaining concepts.

VII. Conclusions

Continuous progress in Information and Communications Technologies (ICTs) offers the possibility of having a great amount of heterogeneous GI available. In current research in the field of Computer Sciences on the topic of ontology, researchers are looking for ways of representing and accessing knowledge in digital GI towards promoting interoperable systems. Meanwhile, researchers in geography ontology-design field have developed ontologies in many domain areas. The distributed nature of ontology development has led to a large number of ontologies covering overlapping domains.

Ontology is a kind of tool that gains importance when looking for interoperability among heterogeneous GIS and web applications. The style of hybrid ontology-based integration provides customers with a unified abstraction layer that allows for independence from the conceptual models of each service provider. In order to make this possible, we need to define the mappings between the various implementation models and shared ontology.

Much of the effort of the research community in Semantic Integration aims at developing techniques and automated tools to ensure successful results. In this paper we propose a novel technique based on set relations in order to measure the interrelationship between different data models and to foresee the result of the integration process. As long as the relevant concepts of the application conceptual model maintain a good cohesive relationship with domain ontology concepts, it is expected that the result of integration shall be acceptable and that no information shall be lost. In a future work, we intend to break down this first measure into specific and measurable sub-measures and accompany the proposal with automated tools that shall give a result that will allow us to measure its applicability.

References

[Buc09] Agustina Buccella. *Integración de Sistemas de Información Geográfica*. PhD thesis, Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur. Bahía Blanca. Argentina, 2009.

[CSH06] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *SIGMOD Rec.*, 35(3):34–41, September 2006.

[Do06] Hong-Hai Do. Schema matching and mapping-based data integration. 2006.

[Ege02] M. Egenhofer. Toward the semantic geospatial web, 2002.

[FEA+02] F. Fonseca, Max J. Egenhofer, P. Agouris, C. Cmara, Frederico Fonseca, Max J. Egenhofer, Peggy Agouris, and Gilberto Câmara. Using ontologies for integrated geographic information systems, 2002.

[Gua98] Nicola Guarino. Formal ontology and information systems. pages 3–15, 1998.

[ISO94] ISO/TC 211 Geographic information/Geomatics. Website, 1994. http://www.isotc211.org/.

[ISO05] Geographic information – rules for application schemas. International Organization for Standardization (ISO) 19109. Technical report, ISO/TC 211, 2005.

[LGMR05] P. Longley, M. Goodchild, D. Maguire, and D. Rhind. *Geographic information systems and science*. Wiley, John and Sons, Incorporated, 2005.

[NM03] Natalya F Noy and Mark A Musen. The prompt suite: interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983 – 1024, 2003.

[OGC94] Open Geospatial Consortium, Inc. (OGC). Website, 1994. http://www.opengeospatial.org/.

[RB01] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10:334–350, 2001. 10.1007/s007780100057.

[Vit09] M. Mercedes Vitturini. Tesis de magister: Modelos de representación de información geográfica, 2009.

[WVV+01] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. Ontology-based integration of information - a survey of existing approaches. pages 108–117, 2001.

Received: June 2012. Accepted: July 2012.