GIS Application Improvement with Multilingual Lexical and Terminological Resources

Ranka Stanković, Ivan Obradović, Olivera Kitanović

GIS Application Improvement with Multilingual Lexical and Terminological Resources

Ranka Stanković[1], Ivan Obradović[2], Olivera Kitanović[3]
[1] assistant professor, [2] professor, [3] researcher
University of Belgrade, Faculty of Mining and Geology, Faculty of Mining and Geology
E-mail: {ranka, ivano, olja}@rgf.bg.ac.rs

Abstract

This paper introduces the results of integration of lexical and terminological resources, most of them developed within the Human Language Technology (HLT) Group at the University of Belgrade, with the Geological information system of Serbia (GeolISS) developed at the Faculty of Mining and Geology and funded by the Ministry of the Environmental protection. The approach to GeolISS development, which is aimed at the integration of existing geologic archives, data from published maps on different scales, newly acquired field data, and intranet and internet publishing of geologic is given, followed by the description of the geologic multilingual vocabulary and other lexical and terminological resources used. Two basic results are outlined: multilingual map annotation and improvement of queries for the GeolISS geodatabase. Multilingual labelling and annotation of maps for their graphic display and printing have been tested with Serbian, which describes regional information in the local language, and English, used for sharing geographic information with the world, although the geological vocabulary offers the possibility for integration of other languages as well. The resources also enable semantic and morphological expansion of queries, the latter being very important in highly inflective languages, such as Serbian.

1. Introduction

Internationalization is one of the major factors currently influencing the application of geographic information systems (GIS) throughout the world. The research described in this paper is based on an integration of lexical and terminological resources, most of them developed within the Human Language Technology (HLT) Group at the University of Belgrade, and the Geological information system of Serbia (GeolISS), developed at the Faculty of Mining and Geology and funded by the Ministry of the Environmental protection. Two basic results of this integration were multilingual map annotation and improvement of queries for the GeolISS geodatabase.

The paper introduces the main features and concepts of GeolISS, as well as of the geologic multilingual vocabulary and other lexical and terminological resources used. It also offers examples for both multilingual map annotation and query expansion. Multilingual labelling and annotation of maps for their graphic display and printing have been tested with Serbian, which describes regional information in the local language, and English, used for sharing geographic information with the world. However, other languages could also have been used, available resources permitting.

The first section of the paper introduces the approach to GeolISS development, which is aimed at the integration of existing geologic archives, data from published maps on different scales, newly acquired field data, and intranet and internet publishing of geologic information. The second section introduces various lexical and terminological resources used in the GeolISS environment. The third section presents the use of these resources for multilingual map annotation with an example. The resources also enable semantic and morphological expansion of queries, the latter being very important in highly inflective languages, such as Serbian. Multilingual map annotation is outlined in Section 4 and Query expansion in section 5.

2. GeolISS

This system represents a repository for digital archiving, query, retrieving, analysis and visualization of geological data, and allows users to create interactive queries, analyze spatial information, edit data and maps, and present the results of all these operations. Data modeling is inspired by different geological models (Richard, Matti, Soller, 2003), interchange schemes (SEE Grid, 2009), and standards proposed by ISO TC211 (ISO, 2009). The design is also significantly influenced by the Ontology Web Language (OWL). GeolISS is implemented using ESRI ArcGIS technology[1], and designed to function as a personal geodatabase and SDE enterprise geodatabase on MS SQL server.

The logical framework of GeolISS implementation is based on five packages of classes: *concept, observation*, *spatial entity, description* and *metadata* (Blagojević et al., 2008). *Concept* represents the core of GeolISS, and is implemented as an aggregation of geological vocabularies, collections of terms and text definitions of domain objects or collections of possible values for properties. Terms in the vocabularies are used to classify observations/interpretations, or to specify attribute values. *Observation* implements field data records and measurements, the basis for classification, interpretation and modeling of geological features. *Spatial entity* is treated as observation location and mapped/interpreted geological occurrence, and implemented in the geodatabase geometrically by points, lines and polygons. This approach provides for visualization of any geological feature and its cartographic presentation. *Description* is implemented as an instance of observation or

[1] ESRI: GIS and mapping software, http://www.esri.com

interpretation, e.g. a collection of properties with assigned values (e.g. attributes) that characterize some geological occurrence. *Metadata* keep track of data source, links to the bibliography, the person, organization, and project responsible for original data acquisition.

Some relationships between records from different tables in GeolISS database are implemented as semantic nodes. Those relationship classes comprise direct relationships, interrelationships, observation relationships and metadata relationships.

GeolISS data management tools are an extension of ArcGIS especially designed for data entry in the GeolISS database. They are implemented in MS Visual Studio 2005 and support data entry in both personal and enterprise geodatabases. They are also implemented to work as a standalone application which handles thematic non-spatial tables in SQL Server 2000. Each type of geological spatial data has an appropriate form for data management, with the possibility of adding various related data, such as geochemical and geophysical analysis, different types of measurements, stratigraphic age, lithology, etc.

3. Lexical and Terminological Resources

The HLT Group has been developing various lexical resources over a long period, and they have reached a considerable volume to date (Vitas et al., 2003). They include morphological e-dictionaries and finite state transducers, which offer the possibilities for solving the problem of flections in queries, and electronic thesauri, ontologies and wordnets which offer various possibilities for automatic or semi-automatic refinement of queries by adding new words to the set of words initially specified by the user.

The HLT Group also produced an integrated and easily adjustable tool, a workstation for language resources, named WS4LR, which greatly enhances the potential of manipulating each particular resource as well as several resources simultaneously (Krstev et al., 2008). This tool has already been successfully used for various language processing related tasks including query expansion (Stanković, 2008a). A part of the WS4LR system is the web application WS4QE (Workstation for Query Expansion) with accompanying web services, which provide for management of these tasks on the web. This tool was integrated with GeolISS to enable the use of lexical resources within this GIS application.

Morphological dictionaries of simple words and compounds are in the so called LADL format (Courtois et al., 1990), and basically consist of lemmas accompanied by inflectional class codes, which enable automatic production of all inflectional forms. The Serbian morphological dictionary of simple words contains 122,000 lemmas, which can generate approximately 1,400,000 different lexical words. The Serbian morphological dictionary of compounds contains about 4,300 lemmas (generating more than 70,000 different forms) and it is being constantly upgraded.

Inflectional finite state transducers (FST) for the inflection of both simple and compound words have been developed within the Unitex system[2]. These transducers play an important role in the query expansion application WS4QE, by enabling a more elaborate query expansion that can significantly improve retrieval performances. The use of transducers is especially important in the case of compounds.

For instance, if a query is performed with the compound keyword *kvarcna stena* ‘quartzose, quartz rock’, three inflectional transducers are used: one for the inflection of the adjective *kvarcni* ‘quartz’, one for the inflection of the noun *stena* ‘rock’ and one for the compound as a whole, which takes care of agreement conditions between the adjective and the noun.

The compound lemma for the given keyword has the following form: *kvarcna (kvarcni.A2:aefs1g) stena (stena.N600:fs1q), NC_AXN*. The A2 transducer generates 13 forms of the adjective *kvarcni*: *kvarcni, kvarcna, kvarcne, kvarcno, kvarcnoj, kvarcnih, kvarcnima, kvarcnim, kvarcnomu, kvarcnom, kvarcnu, kvarcnoga, kvarcnog*, while the N600 transducer generates seven forms for the noun *stena*: *stena, stene, steni, steno, stenu, stenama, stenom*. Thus, there are 91 possible combinations: 13 (from A2) x 7 (from N600). However, due to the third inflectional transducer (*NC_AXN*) which controls agreement, this query expands into only 10 combinations of an adjective form and a noun form:

> *kvarcna stena* AND *kvarcne stene* AND *kvarcnoj steni* AND *kvarcnu stenu* AND *kvarcna steno* AND *kvarcnom stenom* AND *kvarcnoj steni* AND *kvarcnih stena* AND *kvarcnim stenama* AND *kvarcnima stenama*

thus disabling false retrieval.

Due to the abundance of compounds in Serbian, the development of a comprehensive dictionary of Serbian compounds is a tedious task. In the attempt to alleviate this problem, we have developed a procedure for automatic creation of lemmas for a given list of compounds (Stanković, 2008b). This procedure is based on rules and relies on data from morphological dictionaries of simple words.

Automatic creation of lemmas for compounds is of special importance for technical applications, as is the case here. Namely, it often happens that a technical term, which is frequently a compound, is not in the morphological e-dictionary of compounds.

For example, in order to determine the third inflectional transducer for *kvarcna stena* ‘quartz rock’, the following rule of the abovementioned procedure, imbedded in the query expansion application WS4QE, was used: if the first word is an adjective in nominative singular and the second word is a noun in nominative singular and their gender and animate/inanimate category agree, then the transducer NC_AXN should be used:

```
<Rule ID="4" CFLX="NC_AXN">
  <Word ID="1" POS="A"
    Case="1" Anim="$a" Gen="$g" Num="s" />
  <Word ID="2" POS="N"
```

[2] http://www-igm.univ-mlv.fr/~unitex/

Case="1" Anim="=$a" Gen="=$g" Num="s" />
</Rule>

In addition to morphological dictionaries, wordnets in XML format are also used for expanding queries with related words as well as for bilingual searches. The Serbian wordnet (SWN), conceived within the Balkanet project (Tufiş, 2004), presently encompasses about 14,800 synsets. Synsets in SWN are connected via the Interlingual index (ILI) (Vossen, 1998) with synsets lexicalizing same (or similar) concepts in English within the Princeton wordnet, which is publicly available.

Figure 1 depicts an example of a WordNet hyperonym/hyponym tree for Serbian and English, whose synsets belong to domain 'geology'.

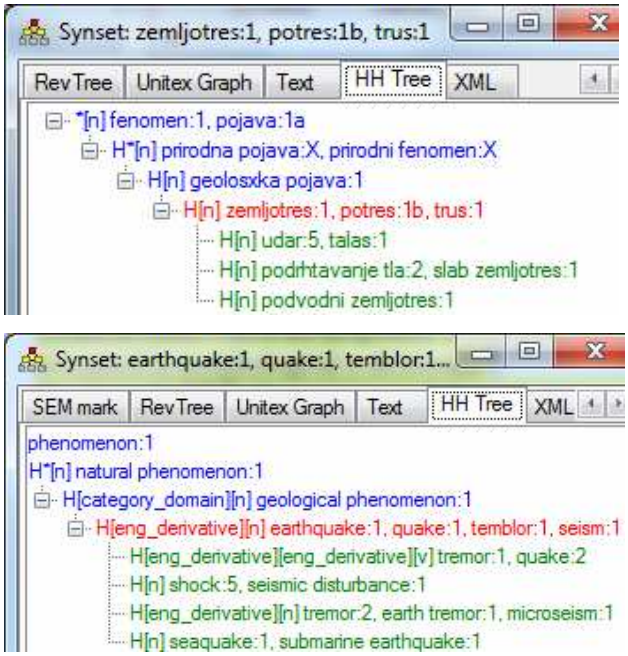

Figure 1: An HH tree in Serbian and English WN

Finally, for semantic query expansion we also used the geological terminological database GeolISSTerm developed within GeolISS. GeolISSTerm represents the core of GeolISS, and it is implemented as an aggregation of geological vocabularies, collections of terms and text definitions of things thought to exist in a domain or collections of possible values for properties. The terms in the vocabularies are used to classify observations and interpretations, or to specify attribute values. The GeolISSTerm database is publicly available for browsing on http://www.rgf.bg.ac.rs/GeolISSTerm/. GeolISSTerm can be viewed as taxonomy with definitions for each entry, synonyms and bibliographical references, as well as equivalent terms and definition in other languages. GeolISSTerm supports semantic and multilingual expansions of the query, as well as multilingual map annotation. In the initial phase around 3,500 Serbian concepts were entered with their English equivalents. Besides English equivalents, several French, German and Russian terms were also entered for multilingual testing purposes.

Figure 2 presents a part of the interface for GeolISSTerm management with an example of a tree view for the concept *Piroklastična stena'* Pyroclastic rock', which has entries in English, Serbian and French:

{ID=2356, Name=*Piroklastična stena,* Def= *Stena nastala depozicijom i litifikacijom piroklasticnih naslaga; fragmenti ove stene obrazovani su direktnom eksplozivnom fragmentacijom,* Synonym= *Piroklastit,* Hyperonym= *2123*}

{ID=*12345*, Name=*Pyroclastic rock,* Def=*A volcanoclastic rock formed by direct explosive volcanic activity*, OrigID=2356, Lng=EN)

{ID=*12367*, Name= *Roches pyroclastiques,* Def= *Une pierre volcanoclastiques formées par l'activité volcanique explosive direct*, OrigID=2356, Lng=FR)

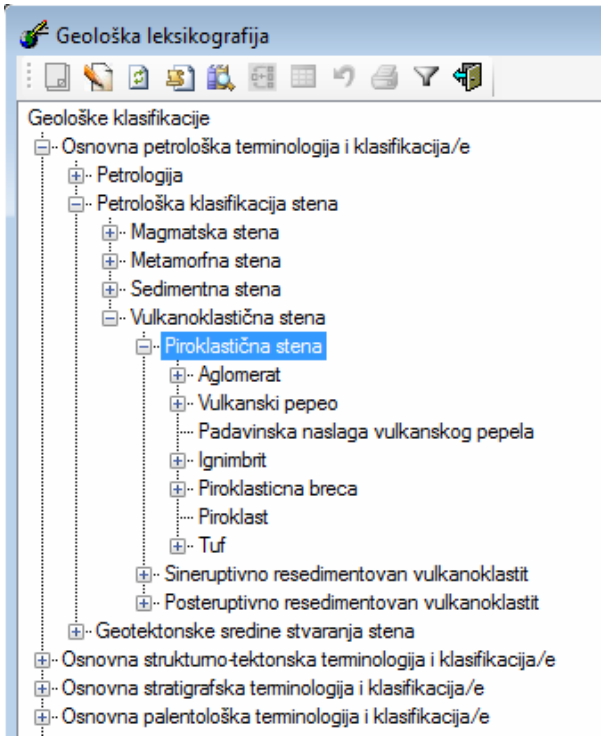

Figure 2: GeolISSTerm HH tree in GeolISSTerm

Further development is underway, as well as concept restructuring and synchronization in compliance with the ISO 1087-1 TMF standard (ISO, 2000). As to the physical (internal) implementation, recommendations of the ISO/TC 37 LMF (ISO, 2008) and ISO 16642 TMF (ISO, 2003) standards were followed.

4. Multilingual Map Annotation

In cartography, annotation is the text or graphics on a map that provide substantial information for the map reader. Annotation may identify or describe a specific map entity, provide general information about an area on the map, or supply information about the map itself.

In general, the placement of descriptive text, or label, onto or next to features on a map is known as labelling. In ArcGIS, it refers specifically to the process of automatically generating and placing descriptive text for map features. A label in ArcGIS is dynamically placed and its text string is derived from one or more feature attributes. This is very useful if data is expected to change or if maps

are created at different scales, and/or for different users.
The user can specify dynamic labelling for all features in a layer, or use label classes to specify different labelling properties for features within the same layer. For example, in a layer of geologic units, the user might label those with a general type of geologic composition, genesis, morphology term, geologic process, metamorphic grade etc.
Development of GeolISSTerm and its integration within GeolISS provides for automatic map annotation in different languages (fig. 3). This is of great importance as it makes visualization of geological data practically language independent, thus substantially broadening the group of potential users.

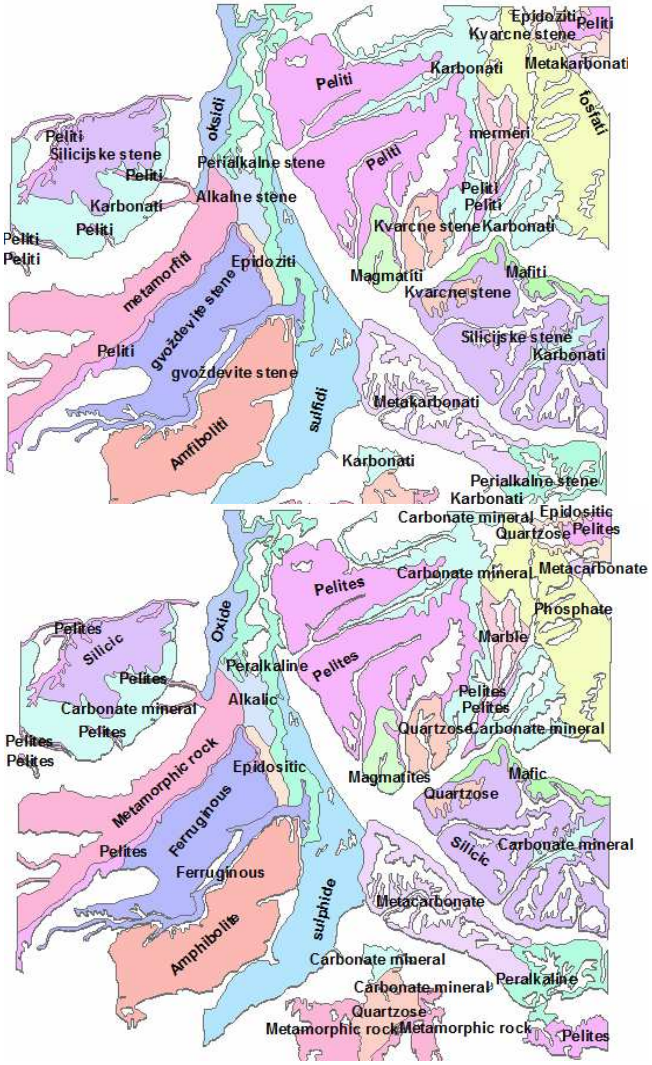

Figure 3: An example of automatic map annotation in English and Serbian

5. GeolISS Query Expansion

Having in mind the diversity and amount of data archived in GeolISS, efficient exploration of such a valuable source of geologic data needs special attention. Data retrieval from the GeolISS geodatabase is provided on several levels: searching on the level of interface forms, spatial object search using GeolISS predefined meta queries, symbolization support with integration of spatial and attribute tables, SQL queries creation using the GeolISS query building tool and ArcMap spatial query aggregation within GeolISS forms.
Spatial object search using GeolISS predefined meta queries is the most frequently used one. Templates for queries are defined in the geodatabase, but the user can update and insert new search definitions. Selection of words chosen for the query can be substantially improved by using lexical resources, morphological dictionaries and transducers in the first place.
For illustration purposes, the query for retrieval of geological units containing 'limestone' ('krečnjak' in Serbian) in their description field was submitted twice: once without and once with morphological expansion. The first query contained only the word 'krečnjak'. In the second query all inflected forms of this word, generated using morphological dictionaries and transducers, namely 'krečnjak, krečnjaka, krečnjaku, krečnjakom, krečnjače, krečnjaci, krečnjacima, krečnjake', where used.
The query was expanded by introducing all inflected forms in the WHERE part of the SELECT query. The unexpanded query retrieved 95 geologic units, while the expanded query selected 249 geologic units. Figure 4 shows part of the results as shaded areas on the left hand side for the unexpanded query, and on right hand side for the expanded query. With the expanded query the recall was more than doubled, while at the same time precision was not reduced. In general, queries often need to be 'fine tuned' in order to obtain an optimal balance between recall and precision.
To illustrate semantic expansion, the query for geological unit retrieval with the term 'facie' ('facija' in Serbian) in their description field was submitted twice: once without and once with semantic expansion. The first query contained only the word 'facija', whereas the second, expanded query included related terms 'litofacija, petrofacija, biofacija' obtained from GeolISSTerm. While the unexpanded query retrieved 24 geologic units, the expanded query retrieved 26.
Besides related terms obtained from the terminological dictionary, semantic expansion can include synonyms from Serbian (and English) wordnets, as for example 'izvor, vrelo' (eng. 'spring, outflow, outpouring, natural spring')..

6. Conclusion

The initial results obtained by integration of the WS4QE web service with GeolISS demonstrate that resources and tools developed within the HLT Group at the University of Belgrade can substantially improve query results for the GeolISS geodatabase. Results also indicate that improvements are available using another resource, the terminological database GeolISSTerm. The multilingual character of GeolISSTerm and the fact that the concepts of Serbian and English wordnets are related by the Inter-Lingual-Index provide for the sharing of GeolISS with the international GIS community. This feature will be

further enhanced by upgrading GeolISS on the basis of users' requests as well as the evolution of geospatial standards and technology trends.

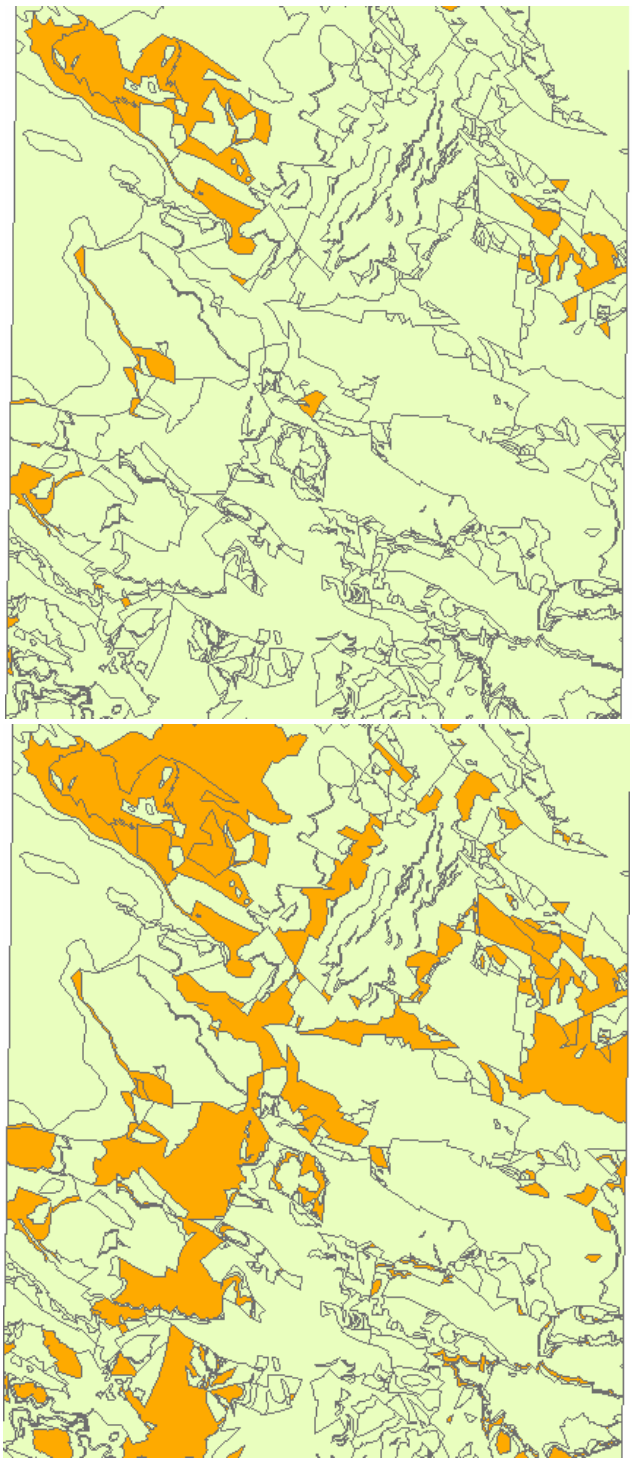

Figure 4: Selected geologic units with the original and morphologically expanded query

7. References

Richard, S.M., Matti, J., Soller, D.R., (2003). Geoscience terminology development for the National Geologic Map Database, in Soller, David R., ed., *Digital Mapping Techniques '03—Workshop Proceedings*, U. S. Geological Survey Open-File Report 03-471, p. 157-167.

SEE (Solid Earth and Environment) Grid, (2009). GeoSciML - the CGI Datamodel and Encoding http://www.seegrid.csiro.au/twiki/bin/view/CGIModel/GeoSciML.

ISO (2009). TC211, Geographic Information/Geomatics. http://www.isotc211.org

Blagojević B.,Trivić B., Stanković R. Banjac N., (2008). "Short note about implementation of Geologic information system of Serbia", *Zapisnici Srpskog geološkog društva za 2007. godinu*, Srpsko geološko društvo, Beograd, pp. 37-44.

Vitas D., G. Pavlović-Lažetić, C. Krstev, Lj. Popović, I. Obradović (2003): „Processing Serbian Written Texts: An Overview of Resources and Basic Tools", *Proceedings of the International Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece, November 2003, S. Piperidis, V. Karakaletsis (eds.), pp. 97-104.

Krstev C., Stanković R., Vitas D., Obradović I., (2008) "The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines", in *Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 28-30 May 2008, European Language Resources Association (ELRA), 2008.

Stanković R. (2008a). „Improvement of geodatabase queries within GeolISS", *Review of the national center for digitization 12/2008*, University of Belgrade, Faculty od Mathematics, pp.65-74.

Courtois, Blandine; Max Silberztein (eds.) (1990). *Dictionnaires électroniques du français*. Langue française 87. Paris: Larousse

Stanković R., (2008 b). „Improvement of Queries using a Rule Based Procedure for Inflection of Compounds and Phrases", Polibits (37) 2008, Special section: Natural Langugage Processing, Journal of Research and Developement in Computer Science and Engeneering, ed. Grigori Sidorov, Centro Innovacion y Desarrollo Tecnologico en Computo, Instututo Politecnico Nacional, Mexico, pp. 14-20.

Tufiş, D. (ed.), (2004). Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology. Bucureşti: Publishing house of the Romanian academy, Vol. 7, No.1-2.

Vossen, P. (ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers

ISO (2000). 1087-1, Terminology work -- Vocabulary -- Part 1: Theory and application.

ISO (2008). ISO/TC 37/SC 4 N453, N330 Rev.16, ISO FDIS 24613, Language resource management — Lexical markup framework (LMF).

ISO (2003), 16642, Computer applications in terminology -- Terminological markup framework (TMF).