

Ranka Stanković
(Faculty of Mining
and Geology, Belgrade)

IMPROVEMENT OF GEODATABASE QUERIES WITHIN GEOLISS

Abstract: We present how resources and tools developed within the Human Language Technology Group at the University of Belgrade can be used for improvement of queries for the geodatabase within the Geological information system of Serbia (GeolISS). The first section of this paper introduces the approach to GeolISS development, whose main goal is the integration of existing geologic archives, data from published maps on different scales, newly acquired field data, Intranet and Internet publishing of geologic information. The Faculty of Mining and Geology of Belgrade University realized the GeolISS project financed by the Ministry of Environmental protection during the period 2004–2006, and from that time many institutions working with geologic data have been using GeolISS. In the second section implementation details are presented, while section 3 presents GeolISS search functionalities. Efficient exploration of such a valuable source of geologic data in view of the diversity and amount of archived data, which is of paramount importance for the quality of results obtained by the query, can be substantially improved by using various lexical resources, such as morphological dictionaries and a geological dictionary. These lexical resources used within WS4QE (Workstation for query expansion) enable semantic and morphological expansion of the query, the latter being very important in highly inflective languages, such as Serbian, which is outlined in section 4. Evaluation of query improvement is described in section 5.

Key words: GIS, digital archiving of geologic data, query expansion

1. GIS approach to digital archiving of geologic data

The main goal in many geological research projects is no longer to create a single geologic map but rather a database, from which many types of geologic and engineering geology maps can be derived. This requires a database design or data model that is sufficiently robust to manage complex geologic concepts such as three dimensional (spatial) and temporal relations among map units, faults, and other features. A geographic or geospatial information system (GIS) is a system for capturing, storing, analyzing, managing and presenting spatially (georeferenced) data and related attributes. Modern GIS technologies are based on digital information, with various methods used for creating digital data. The most common method of data creation is digitization, where a hard copy map or research plan is transferred to digital form by means of a computer-aided design program and geo-referencing capabilities.

Geological data correspond to real geological objects, such as geological units, boreholes, wells, tectonic structures, veins, which are represented in a GIS by digital data. There are two general approaches used for storing data in a GIS: the raster and the vector method. The first method represents geological features as images, whereas the latter uses geometrical shapes, namely different types of geometrical objects, for their representation. There are advantages and disadvantages to using both the raster and the vector data model to represent reality. Raster data sets record a value for each point in the area of interest, which may require more storage space than the representation in vector format that stores only the

necessary data. Vector data can be displayed as vector graphics used on traditional maps, whereas raster data will appear as an image. Vector data are easier to register, scale, and re-project, which can simplify the combining of vector layers from different sources. Vector data are more compatible with the relational database environment. They can be part of a relational table as a normal column and processes using a multitude of operators. The space required for storage and sharing of vector data is usually much smaller than the space required for raster data. Another advantage of vector data is that they can be easily updated and maintained.

In GeolISS vectorization of geologic maps is chosen as the approach to digitization of geological structures, namely geospatial data in general, as well as for their digital archiving. Different geological features are expressed by different types of geometry:

- Zero-dimensional points are used for geological features that can best be expressed by a single reference point; in other words, a simple location. For example, the locations of wells, boreholes, structural measurement or stratification.
- One-dimensional lines or polylines are used for linear features such as boundaries, tectonic structures, faults, fold access, dyke or vein.
- Two-dimensional polygons are used for geographical features that cover a particular area of Earth's surface. Such features may include lakes, park boundaries, buildings, city boundaries, or land uses.

Each of these geometries is linked to a row in a database that describes their attributes. Vector features can be made to respect spatial integrity through the application of topology rules such as 'polygons must not overlap'.

2. GeolISS

The Geological information system of Serbia (GeolISS) represents a repository for digital archiving, query, retrieving, analysis and visualization of geological data. The development and implementation of the GeolISS is managed by a team from the Faculty of Mining and Geology of Belgrade University and financed by Ministry of the Environmental protection. The main goal of the implementation of GeolISS is the integration of existing geologic archives, data from published maps on different scales, newly acquired field data, Intranet and Internet publishing.

GeolISS is a tool that allows users to create interactive queries (user created searches), analyze spatial information, edit data, maps, and present the results of all these operations. Data modeling is inspired by different geological models [1], [2], interchange schemes [3], and standards proposed by ISO TC211 [4]. The design is also significantly influenced by the Ontology Web Language (OWL). GeolISS is implemented using ESRI ArcGIS technology [5], and designed to function as a personal geodatabase and SDE enterprise geodatabase on MS SQL server 2000.

The logical framework of GeolISS implementation is based on five packages of classes: concept, observation, spatial entity, description and metadata [6]. **Concept** 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/interpretations, or to specify attribute values. **Observation** implements field data records and measurements, the basis for classification, interpretations and modeling of geological features. Any observed property can be expressed as a text, number, picture and geometry (location). **Spatial entity**, which is treated as observation locations and mapped/interpreted geological occurrences, are implemented in the geodatabase geometrically as 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 interpretation, e.g. a collection of properties with assigned values (e.g. attributes) that characterize some geological occurrence. The description tables of basic geological entities (mapped unit, lithology composition and geological structure) contain common attributes that specify the purpose and context of the description. The description purpose attribute makes the intended function of a description explicit, e.g., default description, necessary property description, identifying property description, or instance description. **Metadata** keep track of data source, links to the bibliography, the person, organization, and project responsible for original data acquisition, and the processing steps involved in automating the information.

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 also an extension of ArcGIS especially designed for data entry in GeolISS database [7]. They are implemented in MS Visual Studio 2005 and support data entry in both personal and enterprise geodatabase. They are also implemented to work as a standalone application which handles thematic non-spatial tables in SQL Server 2000.

Figure 1 depict the GeolISS working environment for visualization within ArcMap (part of ESRI ArcGIS), with a list of selected layers of spatial data on the left-hand side and the corresponding map on the right-hand side. In the top part of the snapshot is the GeolISS toolbar that enables the selection of appropriate data management tools. The user can select one or more objects from the map and use the GeolISS toolbar for opening the user interface for description of spatial data. The previously selected geological unit is shown in figure 2. 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.

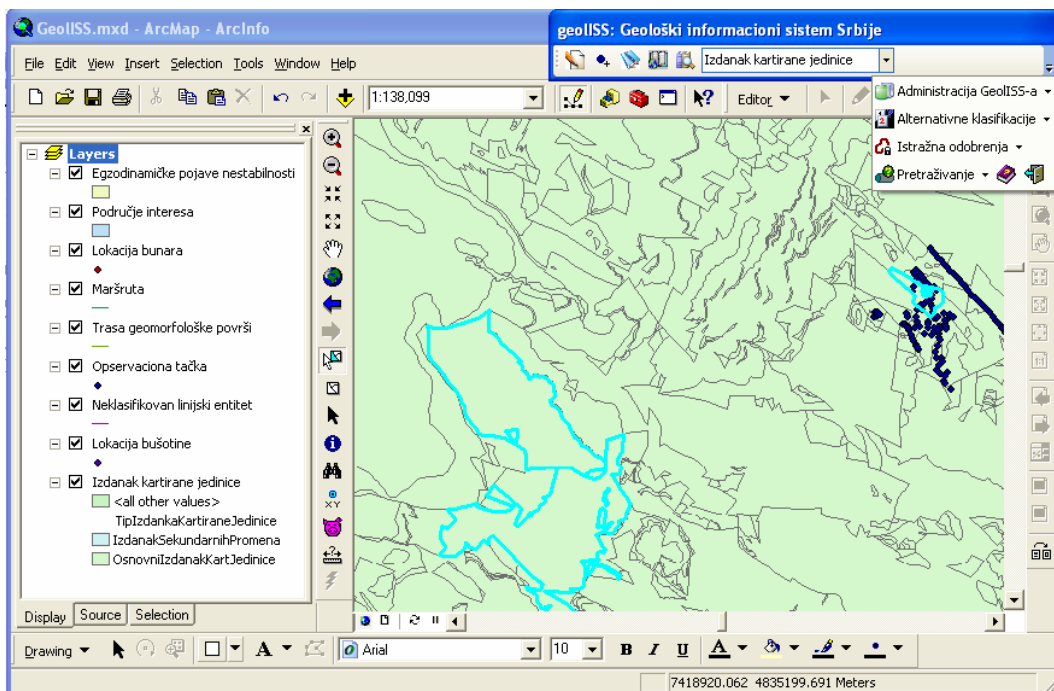


Figure 1. Geological unit selected in GeolISS environment with related data

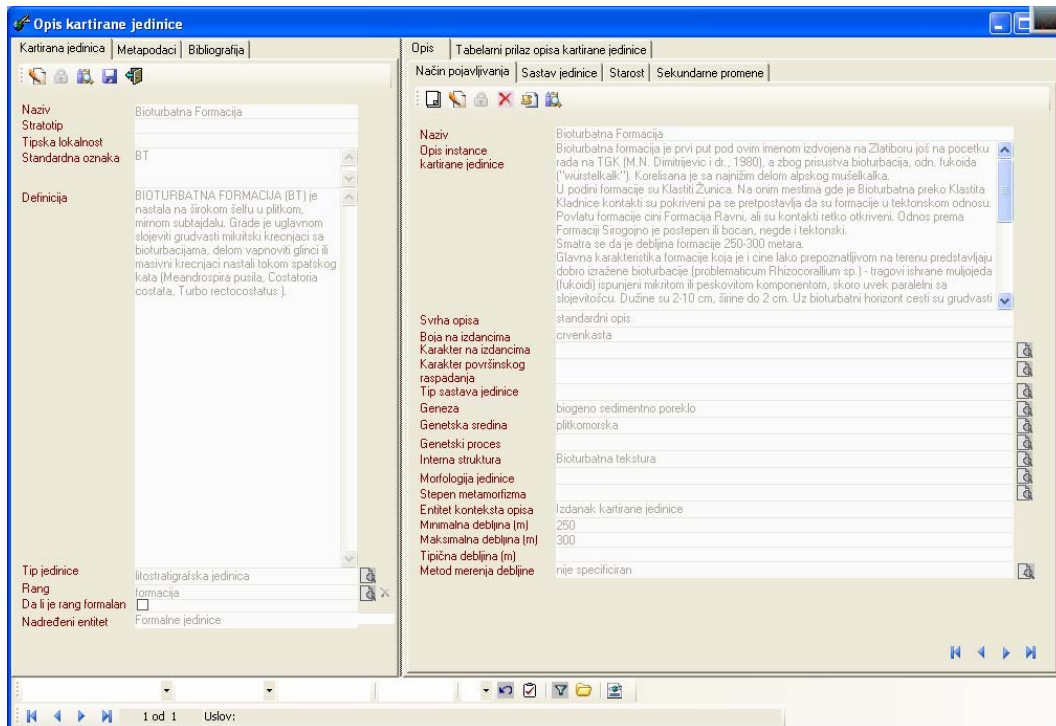


Figure 2. Related data of selected geological unit

3. Search in GeolISS

Having in mind the diversity and amount of data archived in GeolISS it is obvious that efficient exploration of such a valuable source of geologic data is of paramount importance, and hence this problem has to be handled very carefully. Spatial queries differ from SQL queries: they allow the usage of geometry data types (points, lines and polygons), and they take into account spatial relationships between these geometrical objects such as distances, disjoints, intersects, touches, crosses, overlaps, contains, as well as the computation of lengths and areas. 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 joining of spatial and attribute tables
- SQL queries creation using the GeolISS query building tool
- ArcMap spatial query aggregation within GeolISS forms.

The spatial object search using GeolISS predefined meta queries is the one most frequently used. Parameters for queries are defined in the geodatabase, but the user can update and insert new search definitions. There are 46 query criteria defined for ten spatial object types, each query having several attribute options. Figure 3 presents the form with query parameters: class of spatial objects, criteria name and description, typical questions related to criteria, and a list of query attribute parameters. For each query attribute a mapping name (field name in the geodatabase) is defined, as well as header text (to be displayed in query construction), field type with coded values for string 'C', number 'N,I', date 'D', and domain 'X', with the domain name given separately.

Figure 3. The form for maintenance of query parameters

Figure 4 depicts the form for spatial object search for *geologic unit outcrop* 'IzdanakKartiraneJedinice', but in general it can be used for any spatial entity: borehole, well, geologic surface path etc. The user selects among the search criteria offered, in our example, for the *geologic unit* 'Kartirane jedinice' (table with description of spatial object), the field search *Name* 'Naziv', operator *contains* 'sadrži' and *value* *Zlatar*. Many query conditions can be combined with Boolean operators. As a support for novel users typical questions for selected search criteria are presented on the right side of the form. In the bottom part of the form is the WHERE condition for the SQL query, which the user can edit provided that he/she is familiar with the SQL query language.

After query submission, the appropriate layer is displayed with highlighted spatial objects that satisfy query conditions. It is further possible to automatically open the data form with selected (highlighted) objects and to perform further filtering, exploration and data management.

Figure 4. Form for spatial objects searching

4. Web service for query expansion

Selection of words chosen for a query, which are of paramount importance for the quality of results obtained by the query, can be substantially improved by using various lexical resources. Morphological dictionaries enable morphological expansion of the query, very important in highly inflective languages, such as Serbian. The geological dictionary, developed within GeolISS, supports semantic and multilingual expansions of the query. The Human Language Technology group at the University of Belgrade (HLT) has been developing various lexical resources over a long period, the resources reaching a considerable volume to date [8]. They include morphological e-dictionaries and finite state transducers, which offer the possibilities for solving the problem of flexions 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 has produced an integrated and easily adjustable tool, a workstation for language resources, labeled WS4LR, which greatly enhances the potential of manipulating each particular resource as well as several resources simultaneously [9]. This tool has already been successfully used for various language processing related tasks including query expansion [10]. A part of the WS4LR system is the web application WS4QE (Workstation for Query Expansion) with accompanying web services, which provide for management of tasks on the web.

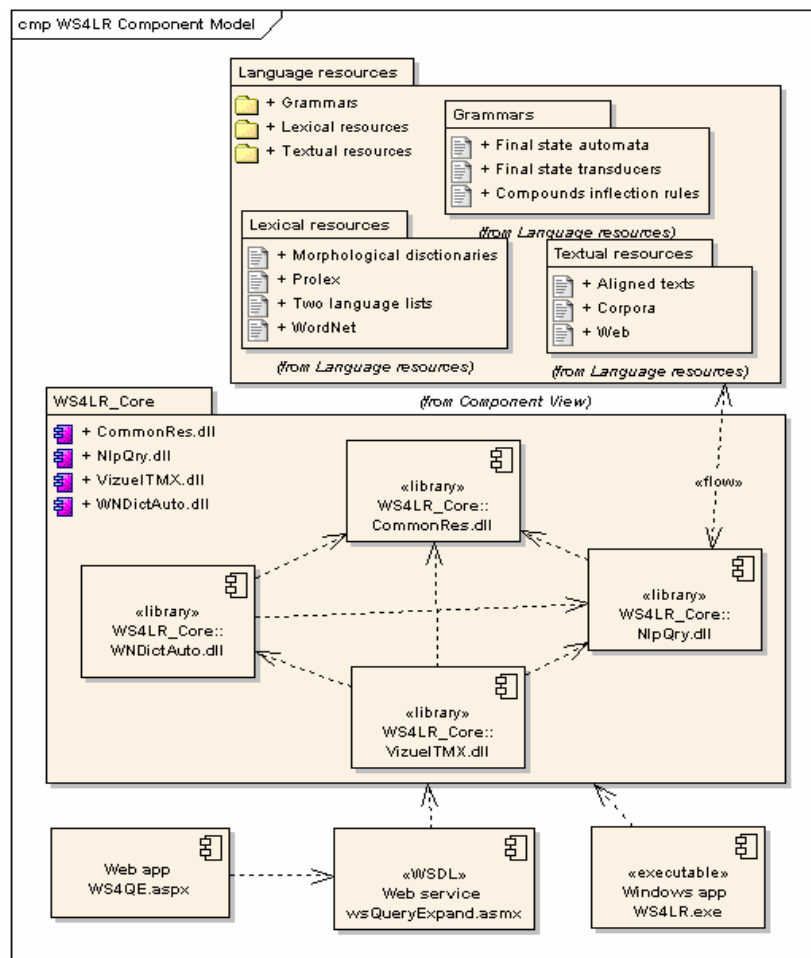


Figure 5. WS4LR component diagram

WS4LR, written in C#, is organized in modules which perform different functions. A Component diagram (Fig. 4) illustrates the pieces of software that make up the WS4LR system. The diagram on figure 4 demonstrates some components and their inter-relationships. The core of the system WS4LR_Core is used by two components: the stand-alone Windows application WS4LR.exe and the web service *wsQueryExpand.asmx*. Web application WS4QE.aspx manages user query request, then uses web service in order to expand the user query, submits the expanded query to the search engine and finally presents retrieved result.

WS4LR handles simultaneously several types of resources, one of them being the system of morphological dictionaries of Serbian simple words and compounds in LADL format. Morphological dictionaries in the same format exist for many other languages, including French, English, Greek, Portuguese, Russian, Thai, Korean, Italian, Spanish, Norwegian, Arabic, German, Polish and Bulgarian.

Another important resource handled by WS4LR is the Serbian wordnet [11]. A wordnet is composed of synsets, or sets of synonymous words representing a concept, with basic semantic relations between them forming a semantic network. Each synset word or “literal” is denoted by a “literal string” followed by a “sense tag” which represents the specific sense of the literal string in that synset, while the interlingual index (ILI) enables the connection of the same concepts in different languages, a feature that can be used, among others, for cross-language information retrieval.

For expansion of queries with proper names WS4LR is using Prolex, a multilingual database of proper names which represents the implementation of an elaborate four-layered ontology of proper names [12] organized around a conceptual proper name that represents the same concept in different languages.

WS4LR also handles aligned texts. A pair of semantically equivalent texts in different languages, such as an original text and its translation, that are aligned on a structural level (paragraph, sentence, phrase, etc.) is known as an aligned text or bitext. The standard format for representing aligned texts is the Translation Memory eXchange format (TMX) that is XML-compliant [13]. Expanded query can be applied on TXM documents in order to retrieve aligned segments that correspond to search criteria in the source and target language. A filtered TMX document is transformed into XML, TXT and HTML output files by XSLT transformations.

The developed web application receives the user query, and subsequently uses the local web service WS4QE to expand the query and forward it to the Google search engine using the Google AJAX Search API. Google AJAX Search API is a Java script library which enables the embedding of Google searches into personal web pages or web applications. This library is composed of simple web objects which perform “inline” search using numerous Google services (Web Search, Local Search, Video Search, Blog Search, News Search and Book Search). The web service returns the required information in XML format, which is being received and converted to appropriate application structures (string, array, table, etc.).

5. GeolISS query expansion

Having in mind the vast amount of textual data in GeolISS (observations, interpretations, field work day books, etc.), morphologic and semantic query expansion is implemented in GeolISS search functions. Apart from HLT lexical resources mentioned, for semantic query expansion the geological dictionary developed within GeolISS can also be used, as a taxonomy with definitions for each entry, synonyms and bibliographical references, as well as equivalent terms and definition in another language (presently only English equivalents are in the database).

For illustration purposes, the query for geological unit retrieval with the term *clay* in the description field was submitted twice: once without and once with morphological expansion. The word *clay*, in Serbian ‘*glina*’, has inflected forms ‘*glina, glinama, glinom, glinu, glini, gline*’, and thus the expanded query introduces all these inflected forms in the WHERE part of the SELECT query. The characteristic parts of unexpanded and expanded queries are:

- KlasifikacioniTermin IN (SELECT InstancaID FROM KartiranaJedinica WHERE Opis LIKE '%glina%')
- KlasifikacioniTermin IN (SELECT InstancaID FROM KartiranaJedinica WHERE Opis LIKE '%glina%' OR Opis LIKE '%glinama%' OR Opis LIKE '%glinom%' OR Opis LIKE '%glinu%' OR Opis LIKE '%glini%' OR Opis LIKE '%gline%')

Figure 6 shows on the left-hand side the results for the unexpanded query, with 51 geologic units highlighted, and on the right-hand side results for the expanded query, with 118 geologic units. In the latter case, the recall was doubled, and precision was not reduced. Generally speaking, queries often need to be ‘fine tuned’ in order to obtain an optimal balance between recall and precision.

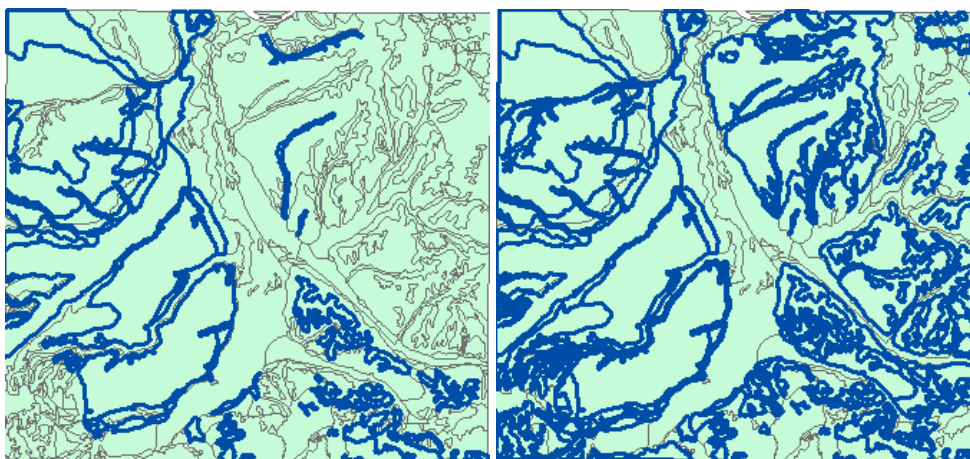


Figure 6. Selected geologic units with the original and morphologically expanded query

Semantic expansion can include synonyms, obtained by geological dictionary, for example:

- Fold ‘*nabor* → *bora*’
- Block faulting ‘*blok stene* → *stenoviti blok, monolit*’
- Abyss, Swallow hole ‘*bezdan* → *ponor*’
- Lower Pliocene ‘*donji pliocen* → *pont*’
- Artesian spring ‘*arteški izvor* → *izvor uzlaznog tipa*’

For illustration purposes of semantic expansion, the query for geological unit retrieval with the term *Lower Pliocene* in the description field was submitted twice: once without and once with semantic expansion. The word *Lower Pliocene*, in Serbian ‘*donji pliocen*’, has synonym ‘*pont*’, and thus the expanded query introduces both forms in the WHERE part of the SELECT query. Figure 7 shows in the upper part the results for the unexpanded query, with 13 geologic units highlighted, and on lower side results for the expanded query, with 33 geologic units.

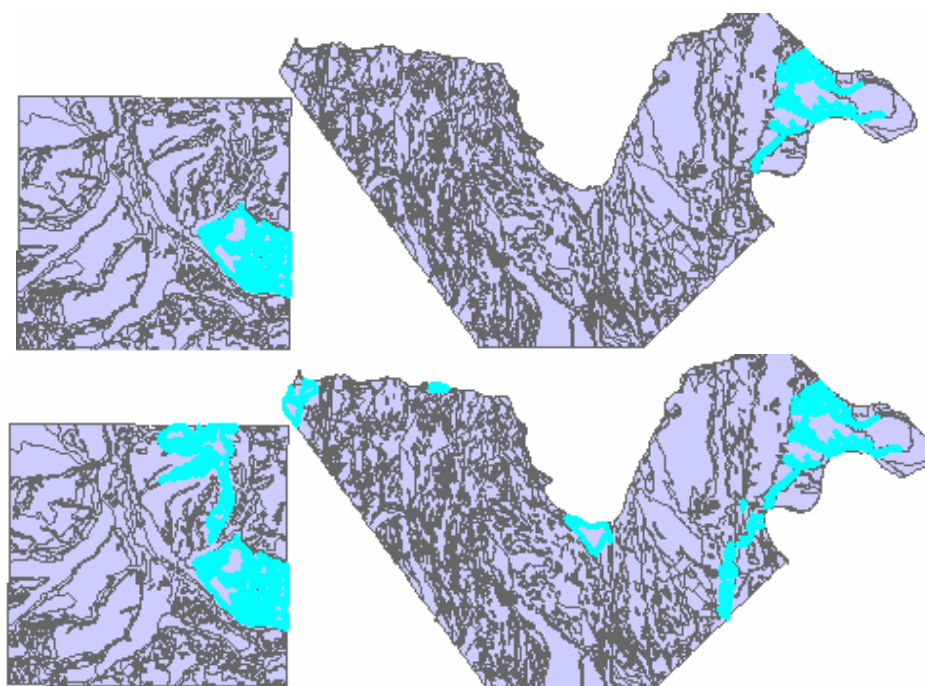


Figure 7. Selected geologic units with the original and semantic expanded query

6. Conclusion

GeolISS has proved as a substantial and reliable foundation for digital archiving, query, retrieving, analysis and visualization of geological data. GeolISS has been implemented in ESRI ArcGIS technology, designed to function as a personal geodatabase and SDE enterprise geodatabase on MS SQL server 2000. The initial version of the Geological information system of Serbia, although having some limitations, presents a solid base for digital recording and management of geological data. Future upgrading must be guided by users and any future GeolISS development must follow the evolution of geospatial standards as well as technology trends. First results of the integration of the WS4QE web service with GeolISS demonstrates that resources and tools developed within the HLT Group at the University of Belgrade can substantially improve query results for the GeolISS geodatabase. A further integration of developed lexical resources, namely the spatial multilingual database of named entities, can be used for a multilingual symbolization of maps.

References

- [1] Richard, S.M., Matti, Jonathan, 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. <http://pubs.usgs.gov/of/2003/of03-471/richard1/index>.
- [2] Richard, S.M., 2003. *Geologic map database implementation in the ESRI Geodatabase environment*, in Soller, D.R., ed., *Digital Mapping Techniques 03—Workshop Proceedings*, U.S. Geological Survey Open-File Report 03-471, p.169–183, accessed at <http://pubs.usgs.gov/of/2003/of03-471/richard2/>
- [3] GeoSciML interchange scheme from CGI, IUGS (<https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/GeoSciML>).
- [4] International Organization for Standardization Geographic Information/Geomatics project (ISO TC211, <http://www.isotc211.org>)
- [5] ESRI: GIS and mapping software (<http://www.esri.com>)

- [6] Blagojević B., Trivić B., Stanković R., Banjac N., (2008) “*Short note about implementation of Geologic information system of Serbia*” u časopisu Zapisnici Srpskog geološkog društva, Srpsko geološko društvo, Beograd.
- [7] ESRI Developer network (<http://edn.esri.com>)
- [8] 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.
- [9] Krstev, C., Stanković, R., Vitas, D., Obradović, I. (2006). “WS4LR: *A Workstation for Lexical Resources*”. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 2006, pp. 1692–1697
- [10] Krstev, C., Vitas D., Stanković R., Obradović I., Pavlović-Lažetić G. (2004) “*Combining Heterogeneous Lexical Resources*”, in Proceedings of the Fourth International Conference LREC, Lisabon, Portugal, May 2004, vol. 4, pp. 1103-1106.
- [11] Krstev C., Pavlović-Lažetić G., Vitas D., Obradović I.: “*Using Textual and Lexical Resources in Developing Serbian Wordnet*”, Romanian J. Information Science and Technology, Romanian Academy, vol. 7, No. 1–2, pp. 147–161, (2004)
- [12] Krstev, C., Vitas, D., Maurel, D., Tran, M. (2005). “*Multilingual Ontology of Proper Names*”. In Proc. of Second Language & Technology Conference, Poznań, Poland, April 21–23, Wydawnictwo Poznańskie Sp. z o.o, Poznań...
- [13] TMX 1.4b specification, <http://www.lisa.org/standards/tmx/tmx.html>