Semantic Interpreter *Pythia* : GIS-based Expert Knowledge Algorithm for Automated Geotechnical Soil Profiling of Applicable Data

Maria Apostolos Papadopoulou[1,2]

[1] Department of Civil Engineering, Alexander University of Applied Sciences of Thessaloniki, Macedonia, Thessaloniki, Greece
Email: dm.papadopoulou [AT] gmail.com
[2] Thessaly Region Local Government, Larissa, Greece
Email: m.papadopoulou [AT] thessaly.gov.gr

ABSTRACT— ***Although the fields of geospatial data are growing rapidly, the result is still not satisfactory for the needs of engineers. No systematic information is available about geotechnical subsurface soil conditions and underground artificial infrastructures. This old-age problem is two-fold: (a) inadequate available digital geotechnical data, and (b) no concepts to improving the applicability and to updating data for engineering applications. On the second, the paper proposes the innovative GIS-based model-driven data processing methodology implemented into an expert knowledge algorithm named Semantic Interpreter Pythia (thereafter SI). From the point of view of geotechnical engineering, the subject of SI is the automated multi-thematic geotechnical soil profiling (GSP) by which it determines the geometry, the properties and the stratigraphy of the site-specific subsoil. From the point of view of geographic information science, the subject of this expert is to relate multi-thematic sets of data from databases, to interpret these data with a specialized data fusion model and, ultimately, to lead to unified information in a core relational database. The paper presents the innovative idea of this algorithm to propose the development of automated SI tools by the modern GIS and internet technology. These tools could help disseminate useful and up-to-date data for a wide range of uses. Based on the experiences distilled from an extensive geotechnical case study, the paper specifies what content is appropriate for engineering studies. The notions of data applicability and geotechnical semantic interpretation arise.***

Keywords— Geotechnical Engineering, Geographical Information Systems (GIS), Semantic Interpreter (SI), Online Engineering (OE), Spatial Reasoning/Intelligence, Spatial Data Infrastructure, Data Integration

1. INTRODUCTION

Although the fields of geospatial data are growing rapidly, the result is still not satisfactory for the needs of engineers. No systematic information is available about geotechnical subsurface soil conditions and underground artificial infrastructures. Engineers need digital, dense referring to spatial distribution, meaningful, and applicable data for the specific uses of civil engineering studies. It is indicative that this inadequacy exists, although there are national and regional projects over the last fifteen years to create a unified spatial data infrastructure (SDI) (e.g., [1] in the European Union) and individual development of geotechnical (geographic, spatial) databases and interoperable Geographic Information Systems (GIS) (e.g., [2] in Greece). Of course, massive efforts at national level should be more intensified. In addition, spatial data infrastructures should, in any case, include geotechnical information. However, the problem is not just how to investigate, gather, and digitize quantities of raw data (e.g. field measurements, laboratory tests). Case studies on automated microzonation studies by interoperable GIS [3] find that, even in the cases of ideally dense availability of spatial raw data at a local area, the applicability of these data is most often restricted regarding of fitness for use for the various geotechnical applications. The term "applicability" is herein used to denote the internal data consistency which is critically dependent upon meaningful and composable data. The present paper points out that the old-age problem of geotechnical data inadequacy is two-fold:

(a) inadequate available digital geotechnical data, and

(b) no concepts to improving the applicability and to updating data for engineering applications.

On the second, the present research proposes the innovative GIS-based model-driven data processing methodology implemented into an expert knowledge algorithm named Semantic Interpreter *Pythia* (thereafter SI). From the viewpoint

of geotechnical engineering, the subject of this innovative type of semantic interpretation of databases is to create automated multi-thematic geotechnical soil profiles (GSP) (section 3.4). A GSP is a fundamental cross-section concept in geotechnical engineering. The term "multi-thematic" in the present research refers to information from a range of actual investigations provide data of geotechnical concern. From the point of view of geographic information science, the subject of the expert SI is to relate multi-thematic sets of data from databases, to interpret these data with the specialized data fusion model and, ultimately, to lead to unified information in a core relational database.

The study considers that the SI's multi-thematic GSP provides advantages (section 3.4) which in general are: (1) A multi-thematic GSP includes the geometry, the properties and the layering of the sites subsoil; (2) It simulates the geotechnical subsurface soil conditions in a meaningful manner thus representing a fundamental input to be used by various geotechnical methodologies; (3) If users can getting ready GSP from SI, they will relieve of demanding processes and uncertainty; (4) During the GSP process, SI at the same time reduces the data inconsistency of a database content improving its semantics-related applicability; (5) The SI concept could share GSP with the modern interoperable GIS so that GIS understand geotechnical semantics and disseminate useful and up-to-date data content for wide use.

The paper presents the innovative idea of this algorithm to propose the development of automated SI tools by the modern GIS and internet technology. The extensive case study used *GeoSeism* [3], which is an interoperable GIS designed to input GSP through interoperability with SI. SI models geotechnical semantics and feeds with applicable geospatial data the various geotechnical methodologies constitute *GeoSeism* (section 3.2). The mobile SI version [4] is another application of this concept. It has taken over to allow working in situ, both on- and off-line, utilizing information and communication technology (ICT), remote sensing and internet technologies. Based on the experiences distilled on what content is suitable for engineering studies and when it is inconsistent, SI developed a specific data fusion model, called DIKW, to process data from geotechnical databases (section 3.5). However, it is the data inconsistency that makes this process much more demanding. The problem is then transmitted from data fusion to resolving the heterogeneity of semantics by inferring of likely complementary data. The notions of data applicability and geotechnical semantic interpretation arise (sections 4.2, 4.5). The Author believes that Semantic Interpreters like SI *Pythia* could play an important role in the exchange of geographic information, allowing SDI and spatial databases to improve their content avoiding duplicated efforts, inconsistencies, delays, confusion, and waste of resources. The aim of SI is dedicated to representing information about the subsurface soil of the real world in a form that a geotechnical engineering application software could use. If thus engineers are able to get ready GSP as input then they could more easily proceed in their specific evaluations related to site effects (e.g., earthquake-generated ground response, soil liquefaction, floods and water flow, settlements, shear failures, loading situation on underground pipes, etc.).

2. BACKGROUND

2.1 The Notions of Geotechnical Geospatial Semantics, Knowledge Representation and Reasoning, Knowledge Expert Algorithms, and Semantic Interpretation

The term "geospatial data" (or GIS data or geodata or georeferenced data or geographic data) refers to data which are pertaining to space (spatial) and, at the same time, have explicit information about their geographic position within a GIS (on the spatially enabled database of the vector map or the geo-referenced satellite image). Spatial data refer to features or phenomena distributed in the three-dimensional space which have physical and measurable dimensions (e.g., the roof-depths and the space shape of a soil stratum, the space position and the spatial distribution of a variable, the earthquake-generated vibrations and the site effects). Accordingly, in the present research, the term "geotechnical geospatial semantics" [3] refers to the understanding of the meaning of geographic entities of the real world pertaining to the engineering semantics, both to the cognitive (human perception) and to the digital concepts of meanings (digital world). Note that, geographic data and information are defined in the ISO/TC 211 [5] series of standards as data and information having an implicit or explicit association with a location relative to the Earth.

Knowledge representation and reasoning (KR^2) [6] is the field of artificial intelligence (AI) dedicated to representing information about the world in a form that a computer system can utilize to solve complex tasks. KR^2 incorporates findings from psychology and logic to make knowledge representation and reasoning, such as to apply rules and relations of sets and subsets. Examples of knowledge representation formalisms include semantic nets, systems architecture, frames, rules, and ontologies. Examples of KR^2 applications include a diagnosing a medical condition or having a dialogue in a natural language.

An expert computer system [7] is a form of software close to the same AI field. It is designed to emulate the reasoning and decision-making ability of a human expert so that identifying facts and most often solving problems in an automated manner. Its code is represented mainly by if–then rules. It basically includes an inference engine and a knowledge base. It may collaborate with a detecting system. Most often it is used for giving explanations or drawing inferences or debugging problems. In general, it deduces a fact based on a known fact and established rules. Examples of automated reasoning engines include inference engines, theorem provers, and classifiers.

Semantic Interpretation is a term which is more acceptable to refer to the natural language understanding component of dialog systems that holds the conversation of a human with a coherent structure. According to [8], the goal of interpretation is to binding the user utterance to a concept, or something the system can understand, during dialogues in a text, speech, graphics, haptics, gestures, or other modes for communication on both the input and output channel. It is creating a database query based on user utterance. The same term is also used to describe conventional "interpreters" which are related either to desktop applications or to syntactic data converting or to search processes.

The same term "Semantic Interpreter" is selected to name the present algorithm SI *Pythia*, because it has just a similar role in the understanding of a larger coherent structure. Nevertheless, compared to the above component, the design, implementation and subject here are essentially different. SI *Pythia* is a reasoning and decision-making application software, closer to the KR² field of AI. Its subject is to estimate geotechnical parameters (semantics) by processing the content of spatial relational databases. However, "semantic interpreter" is the only term could refer to the notion of "semantic interoperability" as a concept in the simulation theory and the model-based information technology.

2.2 Current Trends on Interoperable GIS and Semantic Interoperability

Among the advantages which make the interoperable GIS a sought-after concept is that modelling can extend continuously to lead to current or future interchange of operations and that it can improve the quality of spatial data for the benefit of the open GIS data sources [3]. Aspects which preserve the timeless value to express the current trends for interoperable GIS concepts, such as *GeoSeism* or more complex ones are: "**Interoperability** allows for the analysis of data in addition to the straight exchange" [9]. "Interoperability is the ability of systems to provide services to and accept services from other systems and to use the services so exchanged to enable them to operate effectively together" [10]. "The development of interoperable GIS has long focused on the need for technically unrestricted interchange of both spatial data and traditional GIS operations and analysis" [11].

The above indicate an extent in which a system can manage another system. A necessary ability is to getting tools to work together and to manage each other, aiming to exchange data and interpret that shared data. This ability includes an extension to the semantics of spatial data and the management of their meanings. This is semantic interoperability. It is the ability of two or more systems or elements to exchange information and to use the information that has been exchanged taking advantage of both the structuring of the data exchange and the codification of the data (e.g., including vocabulary) so that the receiving ICT can interpret the data. It is therefore concerned not just with the packaging of data (syntax), but also with the simultaneous transmission of the meaning (semantics) with the data [12]. *GeoSeism* implemented a first concept of interoperable GIS which can model geotechnical semantics to ensure semantic interoperability, as well as allows for current or future interchange of geotechnical operations and applicable data for the benefit of engineers and GIS. Based on the experiences distilled from the development and application of *GeoSeism*, a descriptive definition is [3]: "Interoperable GIS is any GIS-based information system that comprises components which share data and impact over organized datasets, procedures or means in order to achieve commonly accepted goals. The advantages they offer are the diffusion of applicable for specific purposes geographic data, the more automated production of geographic information, the exploitation of the advantages of the internet, the saving of hardware, software or/and resource ware, such as the compliance (use and rights) to inter-operate application algorithms with GIS data sources. They tend to an interoperable interpretation over available geographic data structures. They can impact and cooperate with GIS-based software applications and databases within one or more GIS operations (collection, storage, retrieval, management, visualization, and visual exploration in the Earth's space) and spatial data processes (modelling, process, analysis)".

One conclusion from the above is that the trends tend to the interoperable sharing and impact and to the use of common standards on semantics. The case of *GeoSeism* [3] finds that the knowledge algorithm SI could evolve more so that interoperating with the future interoperable GIS, which (similarly to *GeoSeism*) will specialize in providing automated engineering subsurface information. SI could thus bridge all these GIS with the growing global development of GIS data sources.

2.3 The Problem of Geotechnical Data Inadequacy and Low Applicability

The problem of spatial data inadequacy is general. Efforts to eliminate this known problem have gradually led to today's active efforts for open data and open standards such as to the technology of web mapping, hybrid applications, GIS distributions, volunteered geographic information, and the like. These efforts, either directly or indirectly, indicate that there is a need to create qualitative geographic information that would be useful for a better understanding of natural phenomena and environmental processes.

Referring to geotechnical data, the engineering communities focus the problem of inadequacy on the low availability of raw geotechnical data from actual in situ measurements (e.g. SPT or CPT field-measurements, geophysical in situ measurements, laboratory tests, etc.) and the lack of digital repositories [2]. Thematic and geographic digital maps which are most often distributed in raster formats do not provide geotechnical digital data. The product of microzonation studies is the best source of geotechnical data. Yet, it is not often digitized and much less often properly organized into

accessible databases so as to allow data retrieval and reuse. A database in the sense of a banking application which holds concepts such as entities, attributes, tuples and relations. Although geotechnical databases form a powerful tool for any engineering study and well-planned geotechnical investigation, geotechnical databases have not been developed for a great percentage of European cities. The content of the existing ones remains rather limited either.

On the other hand, engineering case studies [3] find that even in cases where there is an ideal density of data in a local area, the quality of data remains largely sparse, rather ambiguous, and not applicable. The term applicability is used to express the wealth content of a database (section 4.2). According to the present study, the content of databases has to be under special process to become wealth for high performance in engineering applications. Furthermore, there is no implemented framework of geotechnical data and tools to interactively connect with each other in order to use and update spatial data in a standardized and efficient manner. The inadequacy of geotechnical data is still large and the efforts in the domain of geotechnical data seem to be in a rather primary stage. This inadequacy insists albeit the aforementioned global efforts for SDI. The creation of a global SDI of geotechnical data is necessary to reduce the persistent inadequacy. The ongoing massive effort to create unified spatial data infrastructures at national level should obviously be intensified. The insight of SI is to evolve in such a manner that the web mapping to provide spatially enabled databases of consistent and detailed geospatial (GIS) information about the geometry and the engineering properties of the subsurface soil formations. Specific-purpose algorithms are needed for this purpose.

3. APPROACH, RESULTS

3.1 Purposes and Architecture of SI Pythia

The methodology and initial implementation of the innovative algorithm Semantic Interpreter *Pythia* (SI) developed (and tested) in the context of a doctoral thesis [13]. The purpose of SI is to determine the geotechnical subsurface soil conditions. For this purpose, it currently performs a one-dimensional (1D) GSP (section 3.4).

Unlike conventional expert systems, which are divided into an inference engine and a knowledge base, the architecture of SI consists of two subsystems: the procedural code and a database management system (dbms). The code includes stored procedures with embedded SQL processes for evaluations and reasoning. It is programmable allowing for reviews and extensions. It is implemented into an autonomous app allowing flexibility so that it is able to interoperate with spatial dbms or interoperable GIS. Currently, SI is coupled to the dbms *Kallipateira*. *Kallipateira* is a multi-thematic geotechnical relational dbms. It relationally integrates both raw and processed data of soil information. Raw data are the inputs from external data sources; currently from *HelGeoRDaS_uTH* [2] (section 4.1). Processed data are the intermediate and the final outputs from the apps which interoperate each other in the interoperable GIS called *GeoSeism.* Processed data also include the outputs from SI; which at the same time are inputs for *GeoSeism* (section 3.2).

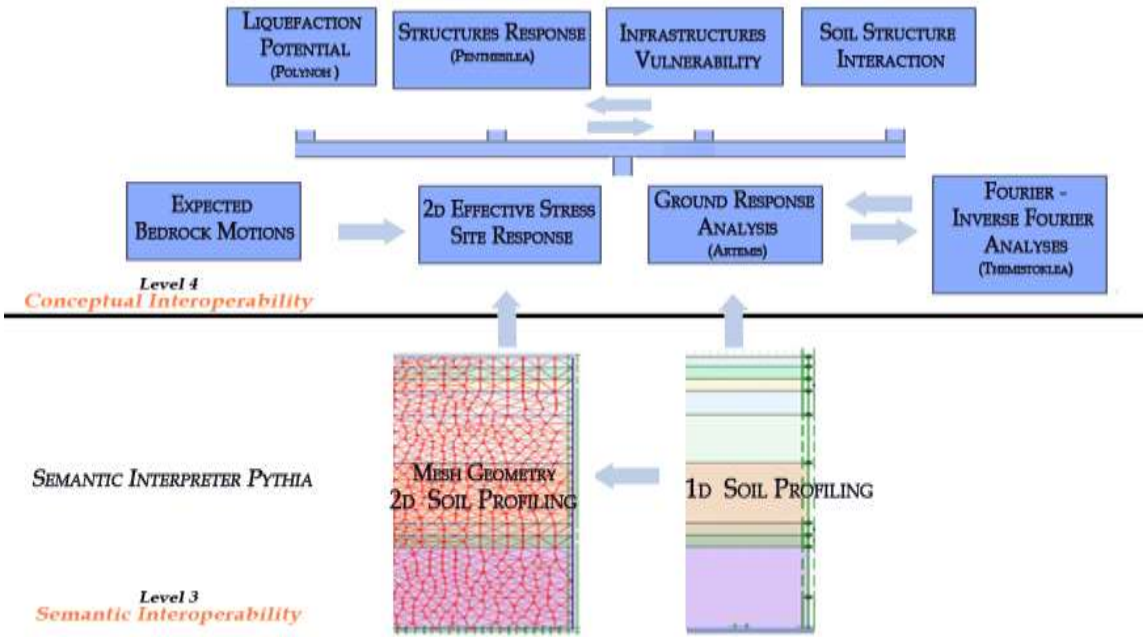

Figure 1: SI (Level 3) Establishes Geotechnical Semantics to Increase the Productivity of *GeoSeism* and the Related Interoperable GIS (Level 4) [Source: [3]]

3.2 SI *Pythia* and Interoperable GIS *GeoSeism*

GeoSeism is a GIS-based application software which is intended to provide more automated seismic microzonation studies by utilizing the interoperable GIS technology [3]. Its methodology estimates the earthquake-generated ground response and lateral phenomena taking into consideration the geotechnical subsurface soil conditions. The automated and much more effective manner to obtain this geotechnical information found to be the interoperation with SI (section 4.2).

Some notes about *GeoSeism*: It is designed to interoperate with a range of standalone (specific-purpose) apps and relational dbms. It portrays an implemented combination of technical along with semantic and conceptual interoperability maturity, in which the highest level of interoperability consists of more than one apps. Currently it uses the same core dbms with SI. Unlike a distributed dbms, which consists of loosely coupled parts that share no physical components, the interoperable GIS concept depicts a communicating framework allows for both data and model sharing to system members. Members (apps) can use data and impact each other. SI attempts to improve the geotechnical semantic interoperability of the system. **Figure 1** depicts some procedures (in the apps) of the interoperable GIS *GeoSeism*. These are mutually interoperating each other in order to elaborate automated microzonation studies (Level 4: conceptual interoperability level). It also depicts the assistance of SI (Level 3: semantic interoperability level) which interoperates with *GeoSeism* to perform the demanding preparatory work of GSP. This way, SI establishes common geotechnical semantics to the apps which *GeoSeism* interoperates.

3.3 Sequence of Events in the Operation of SI Pythia

The sequence of events in the operation of SI *Pythia* is as follows. **Figure 2** depicts the order of data integration, data homogenization, and data fusion. These terms are proposed in the present paper to name the included processes. Due to not having any absolute definition in the literature, the naming is put under discussion in the present community.

Stage 1- Search Retrieval.

The algorithm first attempts to acquire data from the core dbms grounding an SQL query. Next, in a future version, it will connect to a web search tool in order to search for available data through the GIS data sources. The search requests the geographic coordinates (site-specific retrieval) or the identifier (ID) of a certain soil formation or area (ontology retrieval). An ID is an ontology taxonomy which SI establishes to allow for better retrieval of multi-thematic spatial data, easy programming, and 3D depiction of the subsurface soil. In cases of lack of relevant data, the algorithm continues to search for data in close proximity to the required site. After collecting one or more positive replies from databases, the algorithm has either to select the more reliable one or, else, to record all of them in independent records.

Stage 2- Data Integration.

The algorithm attempts to integrate the found sets of multi-thematic raw data by recording them in separate records and these records in corresponding thematic tables of the relational core database. All records follow the organization of the relational data model. It consists of a general table contains the general information of a site (e.g. identity name, coordinates, address, type of investigation, elevations, etc.) and of related multi-thematic tables. Every record of the general table is recorded in the corresponding records of the related thematic table with one-to-many relation. Data integration also involves the modifying of relations between data. For example, when the algorithm has to relate thematic tables about laboratory measurements and hydrological data together, both of which refer to the same site, then SI has to incorporate all of this complementary information.

Stage 3- Data Homogenization.

Data homogenization includes editing tasks and processing tasks. The former aims to improve the layout in the files (e.g., updates anomaly, brings the wealth of records together, to alter the existed tuples, to place in hierarchical order the candidate keys of soil layering, and the similar). Double or triple records are not removed but kept in the dbms to increase confidence. Note that, the present research considers as “double” or “triple” records merely the ones come from the same or different sources and refer to overlapping sites and different type of investigations. For example, one record may refer to geotechnical investigations and the double one may refer to geophysical investigations, but both of them refer to the same site. On the contrary, if the records refer to the same type of investigation then their fields have to be put under consolidation. The latter tasks attempts either to create **complementary fields** in a record or to create **complementary records**. For example, in cases of redundancy of records, if the end-user reads only one of the original records, then he will perform a task-cycle with insufficient inputs. Thus, the algorithm has to consolidate the complementary fields together, so that the algorithm reads all of them at once. Note that, the double or triple records are not considered as “complementary” records. These are not included in this case and are not consolidated. On the other hand, creating complementary records is not as simple as it may seem. This entails the need to interpolate data from the nearby GSP to fill the fields of a current GSP. Similarly to a clustering algorithm, this is an iterative procedure divided into the following steps: (1) sub-divides into thin sub-layers the thickness of the deposit (this process creates empty complementary records); (2) assigns the depths of the in situ samplings (geotechnical or geophysical) data; (3) assigns the middle-depth between two successive of these samplings; (4) matches the fields of nearest sampling to an empty

record; (5) assigns the depths of the laboratory samplings; (6) matches the fields of nearest sampling to an empty record; (6) converge clusters to the empty fields performing some fuzzy logic techniques.

The field matching process to the new records employs criteria based on Euclidean distance. These are optional:

(a) **Nearest Neighbours and Means:** Nearest neighbour (NN) is a simple technique associates the nearest sampling to the geometric (or mass of uniform density) centre of the sub-layer. NN is a well-known clustering algorithm that selects or groups the most similar values.

(b) **Gravity Ranking:** The adjacent to the centroid sampling is graded according to how much of the usable data these hold. This method is a modification of the NN algorithm. It divides the dataset values into different clusters.

Stage 4- Data Fusion.

It is the target-stage of all procedures. It creates GSP (section 3.4) and, at the same time, follows the DIKW model to improve the applicability of data (section 3.5).

Figure 2: Data Integration, Data Homogenization, and Data Fusion processes

3.4 Multi-Thematic Geotechnical Soil Profiling (GSP) of SI Pythia

A **geotechnical soil profile** (GSP) is a fundamental cross-section concept in geotechnical engineering. It attempts to simulate a real-world cross-section of a soil deposit or rock mass under the ground surface, either in a schema or in a spatially enabled database for application use. The latter is what SI creates. There is no strict standard on how to create an appropriate soil profile and what to include in it. For most applications, it is a limited input of data which are manually typed. It is included in the input stage and there is no processing stage to improve these data. It is modified to suit a particular individual software, for its specific syntactic and semantic tasks, without being able to be used in another application. Depending on the application, it may be a one-dimensional (1D), or a two-dimensional (2D), or a three-dimensional (3D) soil profile. It may transmit much or less information. The anyway goal of GSP is to determine the soil geometry and properties as a function of depth that the section crosses. Geometry includes the thickness, the inclinations, the boundaries, and the discontinuities of the soil layers. Properties include the natural and engineering properties (and parameters) of the soil layers; determined on the basis of the available actual samplings investigated the stratigraphy at various depths. An integrated GSP also include the bedrock and the groundwater levels. A site-specific profile typically displays a vertical soil column beneath a required geographic point of the ground surface. In any case, the GSP is a fundamental input for most geotechnical applications (e.g., site characterization, seismic ground response analysis, liquefaction potential, spatial analysis). It is a fundamental because it provides detailed data, soil characterization and distribution of properties in relation to depth.

The term **"multi-thematic"** geotechnical data (or information) is related to geotechnical ground conditions (subsurface soil) and includes: geotechnical data from geotechnical field tests (SPT, CPT) and from laboratory tests, geophysical data (Crosshole, Downhole) from geophysical field tests, geology data (surficial lithology, rock units, bedrock) from geology descriptions, hydrologic data from groundwater measurements, topographic data (coordinates, elevations, contours, discontinuities), and relevant data about the properties and geometry of soil strata, the geology and the aquifers. **Figure 3** depicts the themes of multi-thematic geotechnical data. Details about the geophysical and geotechnical tests can be found in many geotechnical engineering books (e.g., [14]. These data come from a number of related thematic databases (lithology, laboratory, geophysical tests, etc.) of actual investigated locations (IL). An available IL is the site which the actual investigations refer to.

SI is introduced as a modern concept for creating a site-specific automated multi-thematic GSP. It is a special process which subdivides soil deposits into thin sub-layers and then relates the multi-thematic soil data to determine the soil properties and geometry as a function of depth. In the end, the process automatically puts the output into a permanent data structure so that this output is anytime accessible for a variety of purposes. **Figure 4** presents an example of a subdivision of soil layers into thinner sub-layers. The algorithm creates complementary records for every new sub-layer. Then it matches appropriate data to every new record. These data come from a number of related thematic databases (lithology, laboratory, geophysical tests, etc.). Advantages of the GSP of SI *Pythia*, are: (1) Multi-Thematic Data; (2)

Thin Sub-Layers; (3) Deep SP; (4) Maximum Information; (5) Geographical-Referenced Sites; (6) Detailed Automated Information; and (7) Improved Data Applicability (Data Consistency and Composability).

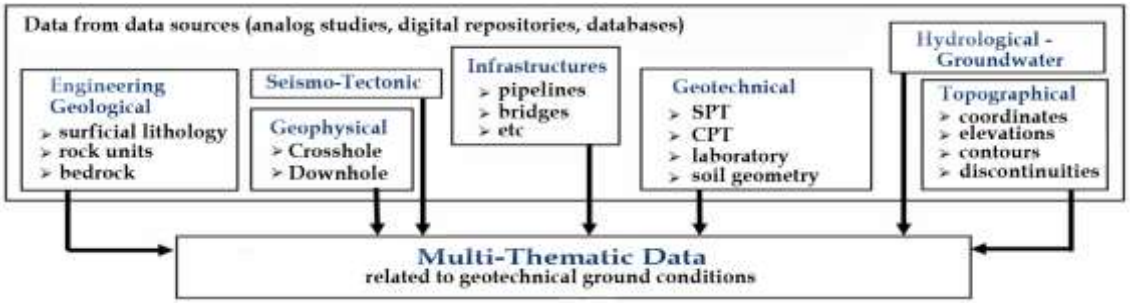

Figure 3: Multi-Thematic Geotechnical Data That SI Utilizes

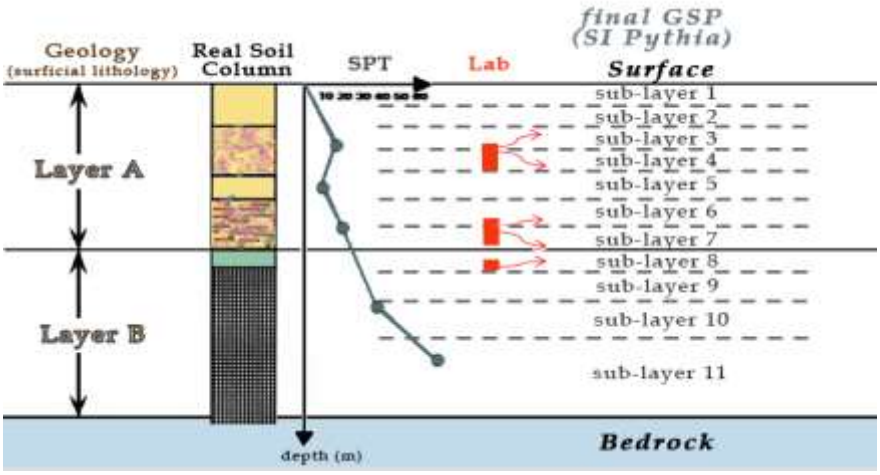

Figure 4: An Example on How a Geotechnical Soil Profile (GSP) is Created by SI

The presently materialized steps of GSP are the following from 1 to 7. The total steps include:

Step 1- Reference System (RS). SI defines a local RS for every site of the examined area. The RS represents the soil column which vertically crosses the ground from a geographic point on the Earth's surface to the quasi-bedrock depth.

Step 2- Multi-Thematic Data. SI gathers multi-thematic data from actual investigations about the required site and its vicinity and associates them to the defined RS.

Step 3- Quasi-Bedrock. SI determines the material and the depth of a quasi-bedrock.

Step 4- Sub-Division of Layers. SI subdivides the unified or the multi-layered deposit into thin sub-layers.

Step 5- Properties of Soil Sub-Layers. SI distributes in situ tests and laboratory properties to the sub-layers following to criteria (section 3.3: stage 4).

Step 6- Vertical Interpolations. SI extents the soil layering to the quasi-bedrock depth by interpolating. Then, SI determines the properties (materials) of the new layers.

Step 7- Complementary Properties (and Parameters). SI evaluates complementary data (soil properties and parameters) for every soil sub-layer. For example, it hyphenates the soil classification, the corrected number of SPT blows, the shear waves velocity depending on the in situ tests, the plasticity category, the effective soil stresses, etc. The result of this Step is a meaningful 1D GSP.

Step 8- Mesh Geometry. SI distributes the soil properties (and parameters) to a required mesh geometry using spatial interpolation techniques between adjacent soil profiles. The result of this Step is a meaningful 2D GSP.

Step 9- Spatial Distributions. SI distributes the soil properties (and parameters) to the space using analytic methods. The result of this Step is a meaningful 3D GSP and spatial ontologies.

3.5 DIKW Data Fusion Model and Applicable Geotechnical Data

DIKW is a specific-purpose methodology designed together with SI to improve the applicability of the data SI employs. DIKW code includes reasoning intertwined with the GSP processes so as to assist and correct step-by-step the workflow. Because SI processes the total content of geotechnical data, thus DIKW has to face all of the various semantic-related inconsistency problems met during GSP. **Table 1** presents a catalogue of such cases, such as actions and methods to deal with them. Reducing inconsistency increases data applicability. Note that, the cases detected during the currently materialized SI levels range from 1 to 5 (section 3.4).

In essence, DIKW methodology is a type of data fusion process. The term "data fusion" is defined as "the act or process of combining or associating data or information regarding one or more entities considered in an explicit or implicit knowledge framework to improve one's capability (or provide a new capability) for detection, identification, or characterization of that entity" [15]. Among the most known models which develop detailed data fusion are the JDL model [16], which was the first attempt to provide a detailed model and a common terminology over data, and the Dasarathy model [17].

Table 1: Systematic data inconsistency problems SI faces related to semantics of data

Problem	***SI's Action***	***Method***
Low Data Quality	Data Cleansing	(Various Fixed Methods)
No Available Data at a Soil Column	Searches Data in the Proximity	Spatial Analysis
Shallow Soil Column	Interpolates	Approaches
No Data about One or More Soil Layers	Calculates Mean Values of Both Sides (Op -Down)	Estimation
No Data on Some Fields	Approaches on the Basis of Adjacent Records	Approaches
No Soil Characterization	Estimates Using a Unified Classification	Estimation
Insufficient Data for Soil Characterization	Approaches Using Alternative Classifications (Class)	Approaches
Insufficient Data for Alternative Class	Approaches on the Basic of Fuzzy Logic (FuzL)	Approaches
No Data about the Groundwater Llevel	Searches Data in the Proximity	Spatial Analysis
No Data in the Proximity	Fuzzy Logic (FuzL)	Approaches
New Data are Available	Updates all Existing Data Sets	(Repeats Fixed Methods)
Other Particular Systematic Heterogeneity	Successively: Searches, Approaches, Alternatives, FuzL	(Various Case-Specific)

DIKW Data Fusion includes a predefined order of data processing procedures in a hierarchical concept of levels. Each superior level is comprised of information, which is increasingly refined (and generalised) as one progresses upwards. This is because the information that comes as output from previous (i.e. inferior) level, acts as data-input to the adjacent level for the subsequent computations. However, the semantic content of requests between each hierarchical level needs to be unambiguously defined: what is sent is the same as what is understood by the receiving layer. The definition of the content of each layer needs also to be unambiguous. For example, spatial analysis is classified at a higher level than GSP, while soil layer characterization is classified at a higher level than the standalone values of the properties, and so on. A fundamental data fusion process is the construction of associations between one or more data elements and another is the evaluation of these elements. The current DIKW designed to serve SI which creates a meaningful data clustering for semantic interoperability. It can therefore find wide applications independently of *GeoSeism*. **Figure 5** depicts the DIKW data fusion framework. The processing stages are layered as follows. The currently materialized levels are the ones from 1 to 3.

level 1— Multi-thematic data level.

The base-level of this pyramid consists of multi-thematic data. This level stores records of collected raw data with no processing besides data integration and homogenization (section 3.3). The data fusion of this stage combines diverse input data sets into a unified (fused) core database by relating records and tables together.

level 2— Soil formations refinement.

It is the first level of Information, in the sense that fused data are here more informative and synthetic than the original multi-thematic inputs. One or more data from the lower level are used to estimate (or approach) new data (fused data). This level may be followed by increment, reduction or replacement. Fused (and consolidated) data sets contain attributes and metadata which might not have been included in the original data sets. The output of this stage includes more accurate and comprehensive information about the soil formations parameters and properties.

level 3— Site Soil profile.

This second level of Information fusion focuses on a higher level of inference. It evaluates the integrated GSP. The overall database now constitutes a meaningful representation of all layers. Data fusion might be viewed as set combination wherein the sets of the lower level are retained. Consequently, information of this level is composable and can be used to evaluate a higher level. In addition, the information from the base to this level is considered as "adjective"

information, because it is based solely on real multi-thematic data with improved confidence. This stage completes data consolidation and data fusion processes.

level 4— Spatial Distribution.

This level evaluates the spatial impact of the site-specific GSP of the previous level. It establishes relationships between the objects based on approaches of the probability theory. A future code extension will perform analytic methods to determine the hypothetical soil profiles by spatial distributions. This information is neither adjective nor composable. Yet, it drives to the creation of a complete 3D GSP.

level 5— Data on Map.

This level connects objects on the map.

level 6— Knowledge.

This level provides data management and open source code. The aim is to educate users on the know-how about the employed methods and the standards.

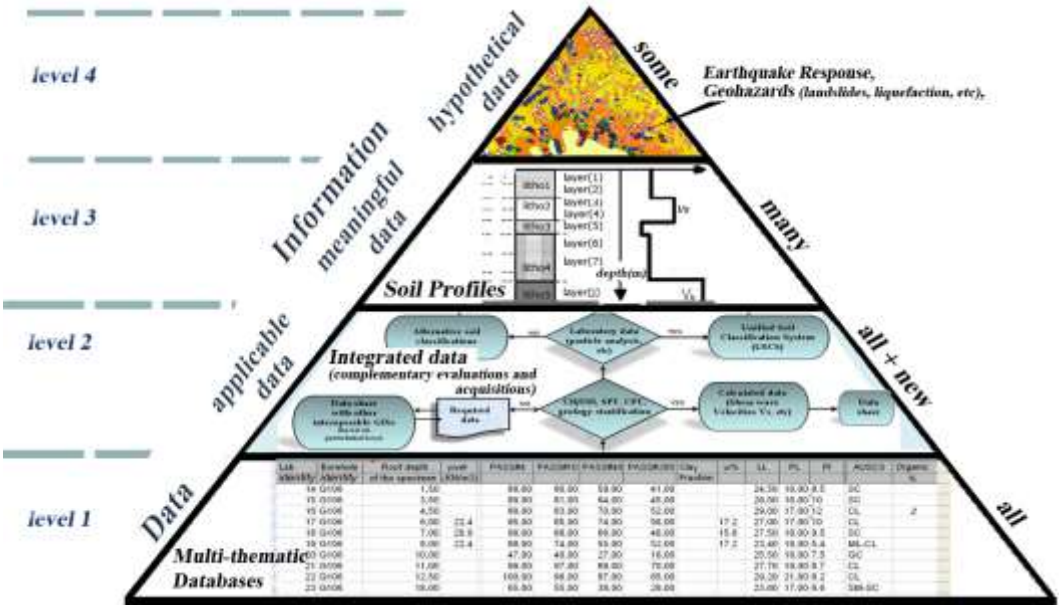

Figure 5: The DIKW Data Fusion Model

4. CASE STUDY, FINDINGS, DISCUSSION

4.1 Case Study

The case study elaborated in the context of the same doctoral thesis [13]. SI interoperated with *GeoSeism*. The role of SI focused on processing the raw content of the external relational dbms *HelGeoRDaS_uTH* so as to provide ready GSP to *GeoSeism*. The standalone *HelGeoRDaS_uTH* and the core *Kallipateira* are compatible. The former represents an external data source feeds SI with a remarkable content of multi-thematic data of more than 342 IL. This content comes from all the reliable available sources investigated the (multi-thematic geotechnical) ground conditions of an entire typical Greek city.

Apparently, it was an extensive case study. Consider that, the SI's GSP subdivides thin soil (sub)layers (of 0.60 – 1.00 m thickness) and the usual number of soil layers ranges from 30 to 150 in every IL. Consequently, the study examined much more than a million soil data values. As expected, various research experiences distilled. One of them it was the experience to handle (by the code of SI) problematic semantics and related transmitted problems at the modelling level of interoperability. A cumulative survey on the interoperability-related experiences is quoted below. Note that, the findings about the geotechnical methodology (seismic microzonation analysis and benchmark problems) of *GeoSeism* are subjects of subsequent related papers.

4.2 Applicable Data, Data Composability, Geotechnical Data Inconsistency, Measuring Applicability

The term "**data inconsistency related to geotechnical semantics**" is proposed in the present research to indicate the semantic-related problems which restrict data applicability. These problems may lead either to operational failures or to rarely detected errors in the output data. Apart from the syntactic heterogeneity, which is always apparent, the most of

these problems are related to the semantics and are rarely detectable. Data quantities do not always guarantee applicability. The code has to acquire or calculate or approach complementary data and to properly establish a common terminology, language, methods, principles, standards, etc. This work is quite demanding and requires much know-how. A variety of semantic-related inconsistency in the input data becomes much more important in the case of the interoperable GIS. Low composability reflects the internal inconsistency of databases. Inconsistency problems are widely transmitted especially when more than one GSP are input for analysis together. This phenomenon indicates that data are not applicable. As a consequence, there is no "automated" seismic microzonation study and the so-called in the literature is rather an exaggeration. The role of the semantic interoperability level is to appropriately process the meanings of the exchanged data so that these are understood by the end-users. **Table 2** presents a catalogue of inconsistency cases SI faces related to the semantics of data. Of course, extensions to the code are necessary to improve these procedures and to solve more semantic-related inconsistency problems.

The term "**applicable (geotechnical) data"** emerged in the present research as:

(a) Data which are meaningful. In the sense that they include as many as possible themes of multi-thematic data.

(b) Data which prove composability regarding of fitness for use. In the sense that composability reflects the capability of data to evaluate a new data element based on two or more other ones.

(c) Data which prove the organizational feasibility of the data structure. In the sense that the data model holds concepts (e.g., relational structure, candidate and alternate keys, ontology) which organize data elements so that these better represent the properties of the real world.

Data composability meant the ability to evaluate (or approach) a target-element of data based on available data sets or elements. Composability is obviously increased in a meaningful database because the latter allows for new data evaluations. Problems associated with semantics (meanings) restrict composability. DIKW provides data values of all intermediate and final levels (section 3.5). This advantage allows the independent use of each level's data to serve various purposes. It is also positive for the productivity of data.

Table 2 presents indicative data productivity measurements (applicability test) before and after improving by SI the input data. Note that, these results come merely from the steps 1 to 5, as these are outlined in section 3.4.

Table 2: Indicative data applicability test before and after improving data by SI

Problems against Applicability	***Before Improvement (%)***	***After Improvement (%)***
Empty Cells of Investigations (IL – Investigated Locations)	87	87
Cells of Sparse Investigations (< 6 IL/cell)	55	55
Depths < 30 % of Total Depth	83	83
Depths = 30 – 80 % of Total Depth	15	23
Depths ≥ 80 % of Total Depth	2	87
No Data about One or More Soil Layers	8	8
Certain Fields of the Records with No Data	88	78
Corrections over the In Situ Tests	0	100
No Soil Characterization	32	20
Insufficient Data for Soil Characterization	16	12
Approach a Soil Characterization	0	12
Alternative Classifications	0	12
Lack of Laboratory Tests	64	62
Lack of Groundwater Data	26	6
Up-to-Date (Active Communication)	0	3
Other Heterogeneity (e.g., Confuse Empty Fields with Zero)	3	0

An applicable data model represents information about the world in a form that an interoperable GIS could productively use. A measure of applicability could be the degree of data elements which an interoperable GIS can productively use from the core database [3]. Based on the present case study, another measure of applicability could be the level to which the DIKW model is effectively materialized (see **Figure 5**).

4.3 The Importance of the Automated Multi-Thematic GSP

Most of the proposed SI's GSP steps are not effective in typical geotechnical software because the latter do not perform any elaborated processing stage for this purpose. Users have then to process data in a manual manner. They also need to know the know-how on a variety of cases. The processing of inputs and results of a

large number of investigations in a city-scale area is a manual work. Similarly, data are manually and repeatedly transferred (after every task-cycle) from one database to another. Therefore, a soil profile usually cannot include more than a few manual entries or requires non-automated modifications. As a consequence, processes cannot easily integrate automated update and reuse of more than one soil profiles at a time (task-cycle). This type of data processing is too obsolete to take advantage of the challenges of modern technology as well as the available internet-based data sources. On the other hand, neither the development of multi-thematic geotechnical database contents is an easy work nor to add new records over existing relationships. These manual works should be avoided due to the very high probability of error, labor intensity and time-consuming use.

However, technological advances aim at drastically reducing dependence on manual methods. To this end, SI is designed to provide a proper database with organized detailed GSP information. This concept is much more accessible and applicable than the prior technology. By the help of this automated algorithm, users could avoid the preparation of a whole GSP which includes demanding (long, complex and uncertain) processes, and would easily obtain meaningful, up-to-date, ready, organized and standardized information.

4.4 1D Multi-Thematic GSP as a First Meaningful Data Clustering

The quality of SI tested through the performance of *GeoSeism*. Because seismic microzonation studies include a wide range of engineering knowledge, SI had to be able to serve many such geotechnical methodologies. It found to be a flexible solution, SI to create just a detailed and multi-thematic GSP. Because a GSP is the first meaningful data clustering, in the sequence, it can be used as an input to various other geotechnical methodologies. However, the latter methodologies should be necessarily adapted to read their inputs by the SI's database. Specifically:

The GSP is a fully defined but implementation-independent model capable of representing the building block from which more complex operations to be constructed. For example, GSP is the most common model before a numerical ground response analysis.

The GSP could establish semantic interoperability to the interoperable GIS so as to improve the applicability of data. If GIS obtain ready GSP [3], then both engineers and GIS will avoid much demanding work from the level 1 to the level 3 (section 3.5).

The GSP allows for further modelling, because many of the 2D and 3D GSP models could be relatively straightforward extensions of the 1D GSP model. A future version could scatter each layer's properties over a mesh network of squares or rectangles. This GSP could then easily help to identify the site-specific properties of the 2D or 3D space. **Figure 1** depicts a general idea on evolving the 1D to a 2D GSP.

4.5 Classifications of SI's Geotechnical Semantic Interpretation and Data Fusion

Through the development and implementation of SI, three new concepts had arisen: First, the particular DIKW model of data fusion (section 3.5). Second, the notion of "Applicable (Geotechnical) Data" (section 4.2). Third, the notion of "Geotechnical Semantic Interpretation". The latter meant the spatial data process methodology of SI which is designed to overcome a variety of data inconsistencies related to the geotechnical semantics of data and to model the geotechnical semantics of multi-thematic data in databases so that ensuring ready applicable geotechnical information within the interoperable GIS.

Similarly to expert algorithms, the general problem SI addresses is to interpret (identify) the facts. It examines and modifies relationships over given table-entities. Depending on the framework [18] which classifies expert systems applications, the present application shows traits of more than one category from "Prediction" until "Design". SI differs from expert systems in that its processes are more complex than a mere rules engine. Particularly, it consists of two subsystems (section 3.1). The dbms represents data. The code represents the processing of data. Data are the known facts. Data processing flows through stored-type procedures and embedded SQL processes.

Similarly to the usual data fusion processes, the novel concept DIKW aims to produce more consistent, accurate, and useful information than that provided by individual data. Compared to Dasarathy and JDL (section 3.5): JDL is oriented toward the differences among the input and output results, regardless of the employed fusion method. Dasarathy model differs from the JDL model with regard to the adopted terminology and employed approach. It provides a method for understanding the relationships between the fusion tasks and employed data, whereas the JDL model presents an appropriate fusion perspective to design data fusion systems. DIKW transmits the problem from data fusion to resolving the inconsistency of semantics by inferring of likely data elements (e.g. soil layers' characterizations). It employs different approaches at every level. It is most oriented towards the creation of complementary data. Compared to the above models, the fundamental DIKW process is the construction of associations between one or more data elements and to evaluate these elements.

5. CONCLUSIONS

The paper proposes the innovative GIS-based model-driven data processing methodology implemented into an expert knowledge algorithm named Semantic Interpreter Pythia (thereafter SI). From the point of view of geotechnical engineering, the subject of SI is the automated multi-thematic geotechnical soil profiling (GSP) by which it determines the geometry, the properties and the stratigraphy of the site-specific subsoil. This concept aims to fully process data and share GSP with the modern interoperable GIS so that GIS understand geotechnical semantics and disseminate useful and up-to-date data content for a wide range of uses. Given the lack of related standalone data processing experts in this field (because this processing is much application-dependent), the paper specifies what content is appropriate for engineering studies. It finds that a GSP is the fundamental input for various engineering methodologies. A multi-thematic GSP is obviously meaningful. SI's multi-thematic GSP simulates in a meaningful manner the geotechnical ground conditions and can be used for various purposes while at the same time relieving users of demanding process and uncertainty. If thus engineers were able to get ready multi-thematic GSP as input then they could more easily proceed in their specific evaluations related to site effects (earthquake-generated ground response, soil liquefaction, etc.). From the point of view of geographic information science, the subject of the expert SI is to relate multi-thematic sets of data from databases, to interpret these data with the specialized data fusion model and, ultimately, to lead to unified information in a core relational database. However, the inadequacy of data makes this multi-thematic GSP process much more demanding. The problem is then transmitted from data fusion to resolving the inconsistency of semantics by inferring of complementary data. Based on the findings of an extensive case study, SI developed the specific data fusion model called DIKW. The final meaningful output is automatically organized in the core database and can be directly applicable by users (human or machines). The notions of data applicability and geotechnical semantic interpretation arise.

6. REFERENCES

[1] **INSPIRE**, Directive 2007/2/EC of the European Parliament and of the Council, establishing an …Infrastructure for Spatial Information in the European Community, at http://inspire.ec.europa.eu/

[2] **Papadopoulou** Maria. A. (under review). "*HelGeoRDaS_uTH*: A Multi-Thematic Geotechnical Database System About The Subsurface Soil And Underground Infrastructures In Thessaly (Greece)", International Journal of Spatial Data Infrastructures Research (http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/487).

[3] **Papadopoulou** Maria A. (2019- accepted) "Geotechnical Geospatial Semantics and Interoperable GIS: the Case of *GeoSeism* for Automated Seismic Microzonation Studies", Asian Journal of Engineering and Technology Vol. 7, No. 1 (ISSN: 2321 – 2462).

[4] **Papadopoulou**, M. A., G. S. **Ioannidis** (2018 a). "Mobilizing the Semantic Interpreter *Pythia* – Teaching Engineering Students to Integrate GIS and Soil Data During In Situ Measurements", In: Auer M., Tsiatsos T. (eds) Interactive Mobile Communication Technologies and Learning. IMCL 2017 (Advances in Intelligent Systems and Computing), vol. 725, Springer, Cham, 2018 a. DOI https://doi.org/10.1007/978-3-319-75175-7_19

[5] **ISO/TC** 211, "Geographic_information/Geomatics", "Geographic_data_and_information".

[6] **Wikipedia**, "Knowledge representation and reasoning (KR, KR², KR&R)".

[7] **Wikipedia**, "Expert Computer System".

[8] **Wikipedia**, "Semantic Interpretation".

[9] **Marr** J. Andrew, **Pascoe** T. Richard, **Benwell** L. George, "Interoperable GIS and Spatial Process Modelling", Proceedings of GeoComputation '97 & SIRC '97, 1997.

[10] **Wikipedia**, "Interoperability".

[11] **Goodchild** Michael , **Egenhofer** Max , **Fegeas** Robin , **Kottman** Cliff , "Interoperating Geographic Information Systems", (Book), 1991. DOI https://doi.org/10.1007/978-1-4615-5189-8

[12] **Wikipedia**, "Semantic Interoperability".

[13] **Papadopoulou** Maria A. "Automated methodology for seismic hazard microzonation studies of interoperable geographic information systems – The case study of a Hellenic city", Doctoral Thesis, Department of Civil Engineering, University of Thessaly, Greece, 2017. (in Greek, at . https://www.didaktorika.gr/eadd/handle/10442/40605)

[14] **Kulhawy** F. H. and P. W. Mayne. "Manual on estimating soil properties for foundation design", Report EL6800. Electric Power Research Institute, Palo Alto, 306 p, 1990.

[15] **OGC** - Open Geospatial Consortium, "Fusion Standards Study, Phase 2", Engineering Report, Date: 2010-12-13, Reference number of this document: OGC 10-184, Category: (Public) Engineering Report, 2010.

[16] **Steinberg** A. N., **Bowman** C. L., & **White** F. E., "Revisions to the JDL Data Fusion Model", ERIM International, Inc, 1999.

[17] **Dasarathy** B. V., "Decision Fusion", IEEE Computer Society Press, 1994.

[18] **Hayes-Roth** Frederick, **Waterman** Donald, **Lenat** Douglas, "Building Expert Systems", Addison-Wesley, (Book), 1983. ISBN 0-201-10686-8