Semantic Web Tools for Categorization Greek Texts on the Internet: the MeDa13 standard and TeGO ontology

Chrystalla Neofytou
Open University of Cyprus
Steliou Siepi 10
8020, Pafos, Cyprus
00357-99563568
chrystalla.neofytou@st.ouc.ac.cy

ABSTRACT

The wider question of this study is the suitability of existing Web search engines for the needs of school education. It examines the relevance to the teaching objectives of the results returned by the search process given a query and its (stated or unstated) purpose in the context of an educational activity. The particular field of teaching and research interest is Modern Greek in Cypriot secondary education (third high school grade, G9). For the purpose of the research, the exploratory-descriptive and experimental approaches were adopted.

The goal of this work is to automate the categorization of Greek texts available in the Internet into textual genres, according to their external (structural/format) and internal (linguistic, stylistic) features, using Semantic Web technology, i.e. metadata and ontologies. The long term vision is to build a semantic search engine that categorizes its results into textual genres as understood and used in high school teaching. The results so far highlighted the need for multi-categorization of texts, mainly due to their hybrid and multimodal nature. The contribution of this paper lies in the construction of a semantic search engine suitable for the needs of school education which returns results classified into textual genres, in accordance to their external and internal characteristics. An algorithm to categorize texts into genres was designed and tested on the basis of the standard metadata Dublin Core (DC). By adapting DC model to the needs of the present work, a model called 'MeDa13' (MetaData13) that includes thirteen elements, was designed. In addition, TeGO ontology (Textual Genres Ontology) was constructed in order to clarify concepts and terms that are stated in the definitions of textual genres. TeGO was developed specifically for teaching Modern Greek focusing on textual genres. In this paper, we describe the proposed models and present the target objectives and outcomes.

Categories and Subject Descriptors

H.3.4. [**Semantic Web**]: Semantic Web tools – Metadata and Ontologies – *structure for automatic achieving Greek text categorization.*

General Terms

Algorithms, Documentation, Theory, Language.

Keywords

Semantic Web tools, Metadata, Ontologies, Automatic Text Categorization (ATC), Greek Language Teaching.

1. INTRODUCTION

Searching and collecting information from the Internet is usually performed by entering queries in search engines. These queries use keywords related to the search objective. Since this objective is typically not explicit, the standard return of such keyword-based queries includes a large number of irrelevant and inappropriate results, failing to semantically relate the query with the subjective goal of the search. By entering a query in a keyword-based search engine, the search is conducted in huge databases that contain copies of web pages in order to match the keywords. Subsequently, an algorithm calculates the relevance of the content of web pages with these keywords. The frequent and large deviation of the results returned by search engines compared to the expected and desired results is due to search engines' manufacturing weakness to perform all the necessary conceptual correlations and also to identify and understand the relationships between concepts and definitions of objects or entities, as defined and understood by humans.

Semantic Web, built on the foundations of the existing web, seems to give the solution to overcome these disadvantages of World Wide Web (WWW). It builds data that relate to real objects and entities of the real world and support associated functions with these (metadata). Semantic Web aims to the structuring of information and the improvement of the existing web, placing (in the web) a semantic layer that allows machines (computers) to understand and process (human) information effectively [1]. Its vision is to equip machines with the necessary knowledge (in the form of dictionaries) so as to be able to 'read', process, interpret and understand the semantic content that concepts and terms carry, resulting to the efficiency and responsiveness of the search engine to the primal objective.

The essential difference between 'traditional' keyword-based and semantic search engines is that the latter do not operate on the basis of keywords but on the basis of semantics of concepts or terms that refer to real objects and entities of the world, describing a dataset (metadata) and representing knowledge (ontology). This means that the semantic search engine returns results related to the purpose (objective of the search) after associating (correlating) concepts and objects with the most appropriate semantic content. Therefore, Semantic Web bridges the communication gap between machines and humans based on the idea of a shared lexical and semantic basis. The basis is built on the construction of a commonly, between computers and humans, perceived semantic framework of concepts and terms that refer to objects and entities of the real world, reflecting, on the one hand, the conceptual relations between them and, on the other hand, representing the relevant knowledge (knowledge representation). Metadata and ontologies, as tools of semantic web technology, enable the readability of data from the machine (machine readable data) and, by creating vocabularies that describe concepts or terms of objects and entities of the world, ensure a common understanding of the domain that the ontology refers to. In this way, knowledge sharing and reuse is achieved.

1.2. RESEARCH OBJECTIVES

This study examines the suitability of existing web search engines in relation to the content and the intended learning outcomes of school education, focusing on language teaching. The key research question is whether Semantic Web, as an evolution of the existing web, with the available tools, i.e. metadata and ontologies, can be used to achieve a better organization of the online data (information). The goal is to place the Internet in the service of school education, so as to give the teacher-user the possibility of option of the most appropriate material (data) for his/her teaching amongst the available information. In the present work, this can be achieved by automated text clustering.

Taking into account the content and teaching objectives of language course for the third high school grade (G9), this paper proposes the automatic categorization (classification) of Greek texts on the internet into textual genres. The corpus of texts examined consists mainly of Greek texts, while all texts written in a foreign language on the Internet are grouped in a separate category especially designed for the needs of this research. Text categorization is performed in accordance to the main external (structural) and internal (linguistic, stylistic) characteristics of the texts and the automation of the process is achieved by using tools of Semantic Web, i.e. metadata and ontologies. The ultimate goal is to build a system (model) for structuring information that digital texts contain, seeking to identify and retrieve specific data (from the texts) that distinguish one (textual) genre from the other. The long-term vision of this work is to construct an optimal, and also ideal for the needs of school education, semantic search engine that returns the results of the search process categorized into genres.

1.3. SPECIFIC RESEARCH QUESTIONS

Firstly, the paper examines the effectiveness and relevance of web search engines, in relation to the content and the intended learning outcomes of school education, focusing on language teaching in secondary education in Cyprus (study of the third grade of high school). The research is oriented to the utilization of the Internet as part of an educational activity, with the aim of gathering specific information, in accordance to the educational activity. In particular, the objective of the web search is to find and retrieve texts of different textual genres.

Second, the survey seeks to give answers whether Semantic Web, as the next step of the current Web, and new generation semantic search engines, are the solution to ensure best (most relevant and appropriate) results considering the context and learning objectives of language teaching. The study explores whether metadata and ontologies (Semantic Web tools) can be used for the better organization of the results returned from the search.

1.4. RESEARCH ASSUMPTIONS

The present work proposes the automatic categorization of Greek texts into textual genres, starting from two research acknowledgements. One is the formulation of general definitions for the textual genres examined and, the second is the necessity of multi-categorization of texts. Specifically, the aim is to articulate 'good' but fuzzy definitions of genres. The accuracy and detail of these definitions on the one hand, and the search for answers to more literary questions on the other, (such as *what is literature and what kind of texts it includes, how can we define the value of a literary text etc.),* do not help the objectives of the research. The goal is to formulate general definitions that describe textual genres and list their main text differentiating characteristics. The aim is the applicability of these definitions in a computing environment so as the texts on the Internet will be categorized (classified) into genres automatically. The definitions, at a later stage, will be rewritten in a programming language, understandable by machines, and the algorithm for the automatic categorization of texts will be developed based on these (definitions).

Furthermore, the survey highlights the necessity of multi-rather than single categorization of texts, implemented by the return of classified results in gradient. The goal is not the absolute categorization of a text into a single genre but its classification in the most appropriate categories, according to the degree of conformance to the definitions, i.e. the number and frequency of the external and internal relevant characteristics appearing in the text. Moreover, as noted in the international bibliography, it is not possible to have a single categorization of texts because of their heterogeneity which is due mainly to their hybrid and multimodal character [2]. The coexistence of stylistic, structural and linguistic elements from different (textual) genres in one text aiming to produce meaning, and/or the use of other semiotic resources (modes) outside of language, for example, video and audio, are factors that confirm the necessity for multi-categorization. In the relevant literature, is underlined that the phenomenon of hybrid (nothogenon) texts is not related solely to technological developments, but it appears as a writing technique, mainly for literal texts, some years before, and is referred to as 'intertextuality' [3]. In contrast, multimodal texts appeared as new (textual) genre within the context of the evolving Information and Communication Technology (ICT), creating a new communication and linguistic reality, and, consequently, underscoring the need to redefine and broaden the meaning of the text to include the new form of digital (usually multimodal) texts.

1.5 RESEARCH METHODOLOGY

The methodology of the research approach was chosen and developed according to the general field of interest and specific scientific and research requirements and objectives of the survey. The research is in the field of computing and ICT, focusing on Semantic Web. It examines the suitability of the existing search engines of WWW having as an objective the use of Internet in school education, particularly, in language teaching in Cypriot secondary education (study of third high school grade, G9). For the purposes of research we adopted the exploratory-descriptive and experimental approach [4]. The information about structural and functional characteristics of both technologies (Semantic and World Wide Web) and the existing search engines were gathered by conducting a thorough literature review (exploratory-descriptive approach). In order to examine the suitability of the results that the existing keyword-based search engines return, using search engine Google, one of the most popular search engines, a simple experiment of online searching to gather specific information about language was conducted (experimental approach).

The textual genres that were included in the list of categories, under which texts from the Internet are categorized into genres, were gathered from the textbook of Modern Greek language of the third grade. The literature review that followed helped to wordage the general definitions and to list the main external and internal characteristics of each textual genre examined. The next step was the text gathering from the internet. Implementing both a random and semi-directed data sampling methodology, we collected one hundred texts from the internet, which were categorized into genres according to the following five criteria: source, content (subject), language, data (elements of) writing and authorial intention (purpose). In this phase of research, the results of categorization were recorded and checked by two groups of teachers, which consisted of ten primary teachers (first group) and ten secondary teachers (second group). The audit was implemented with the distribution of a questionnaire which sought to demonstrate the correctness and validity of the formulated definitions of the genres and categorization criteria. The requested result was the convergence of the results of the initial and control phase classification. In case of discrepancy or ambiguity of the classified results, we examined the likelihood that the disagreement is due to the hybrid nature of the texts, taking into account the possibility of mistaken judgment either on the part of the researcher or the participating teachers.

2. THEORETICAL FRAMEWORK

2.1. World Wide Web and Search Engines

Internet is considered as one of the most direct and fast means of communication and information. The usefulness of the Internet in all areas of human activity is located, mainly, in seeking information, e-commerce and in synchronous (e.g. chat) or asynchronous (e.g. email, blogs) communication. The results of previous study [5] have shown that, for education in particular, the main online activity of users (teachers and students) working in a learning framework, is searching and gathering information

Search engines are special programs used to perform online search for retrieval of digital information from data on the Internet, making Web accessible and searchable to any user. Existing search engines operate on the basis of keywords while giving users the choice of composite or advanced search in order to put restrictions on their search, e.g. regarding the language of the returned results, the file type, etc. After entering the keyword in the search engine, a specific program ‘runs’, which has as a task to look into the database that contains copies of web pages that are automatically selected by the computer network server and are placed into the database of the search engine. After selecting a link on a page, the search engine retrieves the current version of the chosen web page from the computer network server. Special computer programs called “spiders” (online robots) locate pages that may be included in the database by following the links on pages that already exist in the database of the search engines. The next step is page indexing that is performed by a computer program that identifies the text, the links and the page content. Then, the text is stored in the database of the search engine and gives the command for matching keywords. If the content of a register entry in the database of the search engine coincides with the keyword, it is returned as a result. This kind of searches (keyword-based) is related more to the spelling of the word than with semantics. One of the most popular keyword-based search engines is Google. Google is a link-based search engine which apart from the model that it uses of information retrieval is taking into account the importance of hyperlinks. In order to determine the relevance or importance of a website, regardless of the query worded by the user, Google performs a sorting algorithm called PageRank, the result of which changes with the monthly re-indexing [6]. Moreover, Google filters the results as to their similarity, leading to the omission of similar entries. All related to the keyword results are displayed as a list of links that includes relevant pages, which are either classified with the help of an algorithm and depending on the browser in accordance to the degree of users’ views, or listed thematically or randomly. The number of results is usually enormous and so the user needs to choose the most appropriate results for his/her purpose of search. After some brief readings of the results returned and, sometimes, after the necessary query reformulation in order to reduce the numerical results, presumably the user gets the answer (information) he/she was looking for.

Matching keywords with web pages as a gathering-information-method for the existing search engines, highlights the ineffectiveness on searching and retrieving information (IR) from the Internet, especially with regard to digital files that are no texts but combine other semiotic means such as a video, an image or a photo. In these cases, the method cannot be implemented as the search engines are looking for specific data formats into these digital files. The search is based on verbal descriptions that accompanies files and have the form of metadata (data about data). The fact that these written descriptions rely on human factor involves the risk of incorrect description of the file and, consequently, the possibility of failure of the search engine to locate and retrieve the wanted digital archive. Furthermore, the dynamically changing and heterogeneous nature of the Internet, associated with the possibility for users to retrieve and modify the available information, if permitted by the creator of the website, or to enrich the web content by uploading new information, lead to the creation of a chaotic and anarchic digital environment. The fluidity that characterizes the distribution of web information implies an inability to control web content regarding the suitability and reliability of the available information. This underlines the necessity of human intervention in the search process to gather the most appropriate information. It is concluded that the structural and operational weakness of current applications (search engines) for automating the process of searching and retrieving information from available digital resources, each file type (e.g. text, video, photography, picture, etc.), presents a great risk of a failed search.

In addition, the weakness of keyword-based search engines to understand the relation of concepts with objects or entities of the real world and the semantic correlation between the terms in a sentence is due to lack of semantics in the registered information. While the whole process fascinates users, at the same time, it is indifferent to computers, since as machines are incapable of understanding the registered information. As a result a large communication gap between humans and machines is created, since, on the one hand, human has the ability to read and interpret a word, phrase or sentence by giving them the right connotations resulting to general or specific findings, or logical extrapolations after citing two or more truthful sentences and, on the other hand, computer is unable to make such automated correlation and (reasonable) inference. For example, reading the sentence *"Ο Μιχάλης είναι μεγαλύτερος από τον Αντρέα"* (Michael is older than Andreas) concludes that *"Ο Αντρέας είναι μικρότερος από τον Μιχάλη"* (Andreas is younger than Michael). Computer is unable to reach at this logical conclusion, since as a machine it cannot understand the relative age of the two subjects, which is expressed in the sentence using the comparative degree of the adjective "μεγάλος" (old), and the syntax used for comparing two objects in Greek language. Furthermore, the phenomenon of multiplicity, i.e. the use of one word to describe different objects, for example, the word "language" gets the interpretations "anatomical organ", "communication tool", "fish species" etc., and the phenomenon of synonym, i.e. the existence of two or more words that describe the same object or situation but differ in style, performance or expressive significance, for example, the word "clever" and its synonyms "smart", " intelligent", "very clever" etc., make it even more difficult for the search engine to understand the question (query) set by the user. Semantic Web, seeks to bridge this communication gap between humans and machines, based on the creation of a common semantic basis that will allow automation of these functions with minimum human intervention with a view to produce meaning and recover (retrieve) the appropriate information.

2.2. Semantic Web, Metadata and Ontologies

Semantic Web (Web 3.0), as the evolution of the existing web, provides a common syntax and vocabularies for creating statements understandable by machines. It is based on the agreement f a common language for describing logic and the use of this language for exchanging proofs. The aim is to enrich semantically the online information so that it is machine readable and understandable. Computers are equipped with the appropriate knowledge and ability in order not only to be able to 'read' the information contained in the files (web pages), which is written in natural language so as to be understandable and usable by humans, but also to understand the semantic content and relationships between terms, concepts and content. However, it should be noted that the term "semantic" is not inconsistent with the word "global" that describes the broadest network of users, but instead is related to the handling of data (information) available online. The term, taken from linguistics, refers both to the study of semantics of the information that the user sees and read, and, to the information that the machine 'sees' and 'reads'. Specifically, in linguistic terms, semantics seek to examine the relationship of signifiers, i.e. words, phrases, signs and symbols of a language, and their significance (denotata).

Semantic Web, inspired by the creator of the current Web, Tim Berners Lee, is an initiative of the World Wide Web Consortium (W3C), which aims at structuring and organizing online information in order to be process able and interpretable, not only by humans but also by machines [7]. According to the scientific team of W3C, Semantic Web is a web of data that refers to the existence of a common schema and configuration data (standardizing) [8]. In contrast to WWW technology where data are collected mainly by file sharing, with Semantic Web data, derived from different sources, is unified and correlated. Furthermore, Semantic Web refers to the existence of a common language to record the manner in which data is related to real objects or entities of the world. According to Taibi, Gentile & Seta (2005), the goal is to equip computers with tools that enable them to process (human) information effectively putting a semantic layer on the existing web. Its basic principles are to maintain the content distributed on the Internet, to represent and retrieve information, and also, to represent the concepts of specific areas (e.g. education), with the use of ontologies. Creating a commonly perceived, between humans and computers, working semantic framework, Semantic Web seeks to automate search functions and information retrieval from the web, achieving data sharing and reuse.

"Agents", "metadata" and "ontology" are terms directly linked to Semantic Web technology. Software agents are special programs undertaken to 'look' on the Internet and collect on behalf of the user information from various sources with semantic content [9]. Metadata are "data about other data" or "information for further information". According to Tim Berners-Lee (1997), metadata is information about web resources or other objects, which is machine understandable. It is structured information that describes, explains, locates or facilitates the retrieval, uses or manages an information resource. Metadata describe entities aiming to their identification, recognition, discovery and management. Metadata are usually embedded in HTML documents or stored in a database and then linked to the objects they describe. Archiving of digital files and their maintenance, by enriching content with structured (condensed) information contained in metadata, allows future reuse of these data. Administrative information included in metadata safeguard specific elements, necessary to preserve electronic documents due to any changes in formats. In addition, structuring of web content ensures interoperability, i.e. the ability of different systems with different software to exchange data with minimal cost content and functionality. The concept of semantic interoperability is associated with the use of metadata, the related content described and the subject area to which the content refers to.

The term "Ontology", derived from the philosophy, is used in the computing industry, primarily in the field of Artificial Intelligence and is closely related to Semantic Web. Aristotle explains that ontology, i.e. the science of ‘the being’ or reason for ‘the being’, is "the metaphysical study of the nature of life and existence," while to the discipline of computing, ontology is understood as a form of representation knowledge about the world and, in particular, as a representation of the importance of terms in a dictionary and the relationships between those terms. According to Gruber [10], ontology is "a shared and common understanding of a domain that can be exchanged between people and systems applications. It is a formal, explicit specification of the distributed (shared) conceptual representation (conceptualization)". The term 'explicit' means that the type of concepts used and the constraints on the use of these concepts is identified with clarity; while the term 'shared' reports that the ontology should reflect knowledge of common acceptance within a community [11]. 'Semantic representation' (conceptualization) refers to an abstract model of phenomena in the world where the concepts related to them are identified. In related interpretations, the term 'rigid' is added, which indicates that the ontology must be machine readable.

Ontologies are used for describing a specific domain of knowledge in which terms and relations between them are clearly defined, so there is a commonly accepted terminology (glossary), which makes it possible to connect the content of the information society, mechanically processed by a computer, with the meaning given by humans. The ontology defines the formal semantics of information, facilitating its processing by the computer, and also sharing and reuse of the knowledge. Knowledge to ontologies is formulated using five categories of components. These are the classes, relations, functions, axioms and instances. A hierarchical structure that supports mainly inheritance relationships IS-A (subclass of) relations and HAS-A state the relations between the terms (concepts).

3. DATA GATHERING

3.1. Categories of Textual Genres

In the present work, the categories of textual genres examined are formed according to the texts included in the language schoolbook of the third high school grade (G9). Different text genres are selected according to the teaching subject and the stated learning objectives [12]. The schoolbook includes essays, poems, historical, literary, scientific, journalistic, advertisting, interpretative texts, theatrical plays, brocuures, comics etc. Also, it includes texts from the Ancient Greek literature and texts written in a foreign language.

This paper proposes the classification of (digital) texts in six major categories of textual genres having as a criterion the authorial intention (purpose). The first category called *Informative* includes texts that inform the reader on various matters such as advertising or journalistic texts, posters, brochures, leaflets etc. The second category called *Explanatory* includes texts that explain/interpret a word or phrase, a concept or term, or texts that give instructions, e.g. a cooking recipe. Dictionaries, encyclopedias, captions (short explanatory text that accompanies a painting, a picture, a sketch etc.), recipes are some of the subcategories of the category 'Explanatory'. The third category called *Literary* includes prose and poetry texts. These texts aim mainly at readers'amusement and the expression of writer's feelings and thoughts. Some subcategories of texts that belong to this category is the narrative text, the historical text, the folklore, the poem, the epigram. The fourth category called *Artistic* aims at reader's entertaining. This is a special category of textual genre that includes texts using other alternative ways of expression, not the language, such as a photo (use of light) or a sketch (use of lines). The fifth category called *Multimodal* includes multimodal texts, that is texts using other semiotic resources (modes) outside of language. For example, a video in which three modes (language, image and sound) are used combinationally. These type of texts inform or/and amuse the reader. The sixth category called *Foreign_Language* texts is specifically formed for the purposes of this work and includes all texts written in a language other than Greek.

3.2. Texts' sampling from the Internet

The collection of text samples from the Internet was completed in two phases. The first phase of text sampling took place in May-June 2012 and was a random selection of texts from various websites. The second phase took place in September-October of the same year and concerned a semi-guided selection of texts from specific websites (e.g. Portal for Greek Language) with literary or similar content. Google search engine was used for searching and collecting the samples. The keyword used for the search was the word "language" for the reason that is directly related to the subject of this research (language teaching). Google returned 5,120,000,000 results of which, finally, 851 results appeared in 85 pages (links of pages). Firstly, the (links of) web pages were chosen in a systematic way, one page per ten (i.e. first, eleventh to eighty-first page). In total 91 results were collected, from which the results appearing in an odd number were selected (i.e. first, third, fifth result etc). Subsequently, these results were tested and evaluated for their appropriateness of content (one text was rejected due to its unsuitability) and the wanted diversity in textual genre. With the completion of the first phase a hundred texts were gathered and, subsequently, examined to ascertain their suitability. The audit concerned both the content of the texts and the diversity of textual genres (at least three sample texts for each textual genre). Regarding the content of the collected texts, all texts with insulting, offensive, racist, extreme nationalist, sexually explicit and related content, were rejected and replaced with more appropriate texts, collected in the second phase of sampling. From this procedure forty six texts were gathered, while the remaining fifty four were gathered from the second phase. The fact that the list of the hundred samples included mostly informational, journalistic and advertising texts, in a way, imposed the implementation of the second phase of sampling, in order to gather texts from various genres e.g. literary genre (historical texts, poems etc.). At the final stage, all selected texts were grouped according to five cretiria (source, content, language, data of writing, author's intention).

3.3. Categorization of Texts

From the corpus of one hundred texts collected, seventy were selected and categorized. The next step was the evaluation and verification of the proposed categorization that was performed by two groups of teachers. The first group consisted of ten teachers of primary education and the second of ten teachers of secondary education. The distributed control questionnaire consisted of four parts. Part A (Introduction) included general information for the collection of the texts, the purposes of the evaluation process and general instructions to the participating teachers; Part B (Definitions of Textual Genres) included the definitions of textual genres examined, grouped into six major categories; Part C (Example) included a text example with the answer card, that follows each text, completed. The answer card included the proposed categorization and the fields for the criteria 'source' (given by the researcher), 'data of writing' and 'authorial intention', which teachers had to fill in so as to justify their agreement or disagreement to the proposed categorization. Part D (Text Processing) included seven texts that differ as to the source, the style and the content. Each questionnaire had at least one wrong categorization to prevent interlocking results due to given answers, and to gather more reliable test results. For the same reason, participants in each group had different texts to examine but the same texts were examined by participants from the other group. Then, the results were gathered and recorded. The results verified the initial results of the proposed text categorization and confirmed the correctness of the criteria applied. The next step was the automation of the process using these criteria for the development of the algorithm.

4. ALGORITHM FOR ATC

The development of the algorithm for automatic classification of Greek texts into genres using metadata was completed in two stages. The first stage focused on the design and development of the algorithm for the automatic text classification (ATC), and the second in the implementation and evaluation of the tool used (metadata).The algorithm was developed relying only on the text and on the basis of the command *if-then*. The five criteria used for the text categorization (source, content, language, data of writing and authorial intention), were used as the conditions (statements) in the algorithm. In the main body of the command *if-then*, the external (structural/format) and internal (linguistic, stylistic) characteristics of the textual genres examined were recorded. These criteria (conditions) are some of the elements that compose the metadata model that is designed in the next stage in order to automate the process of text categorization.

At this phase the algorithm was developed in pseudo code form and so understood only by humans. ATC algorithm 'runs' in two levels. At the first level of performance, having as a condition (criterion) the authorial intention (purpose) and, only for the category *Foreign_ Language* having as criterion the language of the text, the algorithm gives as a result the categorization of texts in six major categories of textual genres. At the second level of performance, having as a condition (criterion) the source of the text and having as hidden conditions (nested statements) the criteria subject (content) and data of writing, the algorithm returns the results of more specific categorization, i.e. the subcategories of textual genres. For example, an informative text (at the first performance) is subcategorized as journalistic (at the second performance).

1st application:

```
 If a text aims to inform the reader (purpose)
 then it belongs to the category Informative.
```

2nd application:

```
If the text aims to inform the reader (purpose)
    and if  is taken from a newspaper or magazine (source)
    and if it discusses current issues (economy, politics, social)
    (subject/content)
     and if it derogates from the linguistic norm e.g. omission of
    the article
    and it uses (data writing)
        1. simple everyday language
        2. formulaic expressions
        3. literary words
        4. scientific terminology
        5. passive syntax
        6. new and unusual composite word pairs
        7. evaluative adjectives for formulating judgments and
           comments
then it belongs to the category Journalistic.
```

5. METADATA MODEL MEDA13

Metadata model MeDa13 is specifically designed for the needs of this research. The proposed model is constructed on the basis of Dublin Core (DC), one of the most widespread metadata standards [13]. The goal is to automate the process of text categorization into textual genres. DC is implemented on the basis of metalanguage XML and RDF and uses fifteen elements for describing digital objects such as video, audio, images, text and websites for easy identification and retrieval. The components of DC are Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights. MeDa13 is designed in accordance to the requirements of the present work and aims to identify and recover (from digital texts) the data (elements) that clearly indicate the (textual) genre of each text. It describes all kind of texts, including both hybrid and multimodal texts.

MeDa13 includes thirteen elements, ten of which are taken from DC. These are the elements Title, Creator, Subject, Publisher, Date, Source, Type, Description, Language and Rights. The three elements that complement MeDa13, taken from the algorithm that was developed previously, are Modification Date, Data Writing (DAW) and Authorial Intention (AUI). MeDa13 standard has two applications. Simple application is used to categorize texts in six major categories of textual genres and includes only five elements (type, description, language, data writing and authorial intention). Composite application is used for the more specific categorization and includes all thirteen elements.

6. THE ONTOLOGY TEGO

The development of the ontology Textual Genre Ontology (TeGO) was completed in six stages. The first stage is defined as the field of knowledge and use of ontology TeGO (finding the purpose of the ontology). The second step is the apprehension of ontology notions and relationships between concepts and terms that refer to the concepts and relationships in the ontology. In the third stage axioms, rules and constraints (restrictions) are defined. Axioms are always true sentences that express common perceptions generally recognized by concepts and, also, help to limit the interpretation of the concepts contained in the ontology. Rules are statements in the form of *if-then* sentences (condition-finding), describing the logical conclusions which can be drawn based on certain assumptions. Finally, restrictions are typical descriptions of what must be true in order for an instance to belong to a class. In the fourth stage, the ontology TeGO will be encoded in a specific language (e.g. OWL), followed by evaluation. In the sixth phase, documentation and implementation of the ontology developed will be performed.

TeGO is developed in the broader field of language teaching and it describes concepts that refer to the textual genres examined in the present work. The proposed categories of textual genres are defined as classes of the ontology developed. These are the categories of textual genres *Informative, Explanatory, Literary, Artistic, Multimodal and Foreign_Language*. The first relationship defined for the classes is the relationship *has_a (property)*. Each class has two attributes (properties), *Language Identity* (LangID), which takes the values 0 (=non Greek texts) and 1 (=Greek texts), and *Authorial Intention* (AUI), which takes the values 0 (=not clear purpose: to inform or entertain the reader), 1 (=purpose: to inform the reader), 2 (=purpose: to entertain the reader). Class *Foreign_ Language* has one attribute

(LangID) that distinguishes this class from the other five classes. Table 1 shows the values of attributes LangID and AUI for the basic classes of TeGO. Each subcategory of textual genre is a subclass a class. For example, subclass *Advertising* is a subclass of the class *Informative*. *Is_a* relationship indicates that each subclass has same attributes and values with those of the class that is a subclass of. However, each subclass, in addition to characteristics that inherits from its class has the attribute Data Writing (DAW) that describes specific characteristics of the textual genre, which are not inherited from its class but are unique. Data Writing refers to the external and internal characteristics that distinguish one textual genre from another, such as metrics and rhythm in a poem. Table 2 shows an example of coding class *Informative*, given the values of attributes LangID and AUI for its subclasses. Subclasses of all classes of ontology TeGO are coded as the example.

Table 1. Attributes for basic classes of TeGO

Classes	Attributes	
	LangID	**AUI**
Informative	1	1
Explanatory	1	1
Literary	1	2
Artistic	1	0
Multimodal	1	0
Foreign_Language	0	-/-

Table 2. Attributes for subclasses of class Informative

*Not marked at this stage

Classes	Attributes		
	LangID	**AUI**	***DAW**
Advertising	1	1	-
Journalistic	1	1	-
Chronograph	1	1	-
Scientific	1	1	-
Poster	1	1	-
Brochure	1	1	-
Interview	1	1	-

7. CONCLUSIONS AND FUTURE WORK

The core observation that underlies this paper is that, in the case of web technology utilization in school education, internet is useful mainly for the results it gives. The target is to ensure the relevance of these results in relation to the stated learning objectives and teaching content. As noted in the introduction of this paper, the long vision of this work is to build an optimal semantic search engine in order to overcome any weaknesses of existing search engines and to place Internet in the service of school education. As recorded, the proposed metadata model MeDa13 includes descriptive, structural and administrative information of any type of digital file -not only texts- related to the subject area of education, in particular, of language teaching (genres). TeGO ontology that is built for the needs of the present work interprets and represents these concepts, which are related to the genres discussed in the language course of third high school grade (G9). MeDa13 and TeGO will be the tools for the construction of the Optimal Semantic Search Engine (OPSSE). Future work will focus on the completion of the steps in developing TeGO, on the implementation and thorough evaluation of the proposed models (MeDa13 and TeGO), and, finally, on the construction of OPSSE.

8. ACKNOWLEDGMENTS

Many and sincere thanks and gratitude to my supervisor of my doctoral thesis, Mr. Thanasis Hatzilacos, Professor of Information Systems at Open University of Cyprus.

9. REFERENCES

[1] Taibi, D., Gentile, M., and Seta, L., 2005. A Semantic Search Engine for Learning Resources. Recent Research Developments in Learning Technologies.

[2] Kessler, B., Nunberg, G., and Schütze, H., 1997. Automatic detection of text genre. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the conference, 7–12 July, Madrid, (pp. 32–38). [San Francisco, CA]: Morgan Kaufmann Publishers, 1997.

[3] Chandler, D., 2009. Intertextuality. Semiotics for beginners. http://users.aber.ac.uk/dgc/Documents/S4B/sem09.html

[4] Korres, C., 2011. Methodology of educational research: Quantitave approaches to research. Athens, 2011.

[5] Neofytou, C., 2008. Language Teaching & ICT: training procedure, trainers' and trainees' aspects on ICT in Cypriot Secondary Education. Master thesis. Department of Applied Linguistics. School of Philosophy. Aristotle University of Thessaloniki. Thessaloniki, Greece 2008.

[6] Mbalkizas, N., 2006. Google Search Engine. Presentation 3. In teacher training for the usage and implementation of ICT in teaching practice (Training Level II). 2006. http://users.sch.gr/nikbalki/epim_kse/files/Parousiaseis/Google_SearchMachine.pdf

[7] Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. Scientific American. May, 2001.

[8] W3C Semantic Web Activity http://lpis.csd.auth.gr/mtpx/sw/

[9] Kanellopoulos, D., 2012. The benefits of Semantic Web in e-business. Modern Technical Inspection. Issue Nov-Dec 2012. http://www.technicalreview.gr/index.php?option=com_content &task=view&id=684

[10] Gruber, T. R. 1993. "A translation approach to portable ontology specifications". In: *Knowledge Acquisition*. 5: 199–199.

[11] Gaitanou, P. & Gergatsoulis, M., 2006. Ontology management: extended study on the main problems and presentation of existent ontology library systems. In *Proceedings of the 15th Pan-Hellenic Academic Libraries Conference, pp. 136-150, Patra, Greece, 1-3 November,* 2006.http://conference.lis.upatras.gr/files/2.04.FullText.pdf

[12] Pedagogical Institute of Cyprus. Teacher's book: Modern Greek Language, 3^{rd} Grade. Cyprus, 2008.

[13] Dublin Core. http://el.wikipedia.org/wiki/Dublin_ Core

[14] Berners-Lee, T., 1998. Semantic Web Road map. . September 1998. http://www.w3.org/DesignIssues/Semantic.html

[15] Hatzilacos, Th., 2011. WWW search environment for K-12 Education, Seminar at the Open University of Cyprus. Nicosia, Cyprus 2011.