ACADEMIA DE STUDII ECONOMICE - BUCURESTI
Fondata prin Decret Regal la 6 aprilie 1913
Institutul de Studii Doctorale

# Designing and implementing a semantic e-learning platform

**Coordonator ştiinţific:**

**Prof. Univ. Dr. Ion Smeureanu**

**Doctorand:**

**Adam-Nelu Altăr-Samuel**

Bucureşti, noiembrie 2014

# Contents

**Keywords:** e-learning platform, semantic web, text classification algorithms, domain ontologies, web services, semantic mapping

**Summary**

The current thesis' final objective was the design and implementation of a software eLearning platform, based on a detailed study of the most important semantic web technologies, including emerging semantic eLearning standards.

The users of the platform can input texts from different areas of knowledge and the infrastructure will suggest the domain of the text. This includes the design and implementation of an algorithm that combines the results of several text classification algorithms, in order to improve the accuracy of the overall classifier. The platform will then suggest appropriate studying materials for the inferred domain. For this purpose, the platform features a component that continuously searches the internet for new learning materials to add to the knowledge base.

The thesis is structured in five chapters.

**The first chapter** introduces the basic concepts and technologies that describe the semantic web, as well as the improvements it can bring to eLearning.

The semantic web is an extension of the current web that allows searching and combining data fast and effortlessly. This new approach is based on collecting, processing and publishing information interpretable by machines and metadata expressed in RDF (Resource Description Framework). At present, the content of the World Wide Web is designed to be read by humans, not reused by applications. The semantic web will complement the current web, creating and environment in which software agents will be capable of processing various sophisticated tasks, an environment in which data will have a well-defined meaning. Hence, there is hope that, in the near future, computers will not only be capable of displaying data, but of „understanding" it.

The characteristics of the semantic web, i.e. well defined meaning of concepts and automatically processable metadata, used by appropriate software agents, establish an efficient approach for satisfying the requirements of eLearning. Learning materials can be interpreted semantically and, at the request of users, reorganized in order to create

new didactic modules. Based on the user's requests and preferences, learning materials and other information considered relevant can be combined in a simple and intuitive manner. This process is based on semantic queries and navigation through the learning materials and is possible through the use of ontologies, which provide exact definitions of concepts and notions.

The following table presents the benefits eLearning can bring to the classical learning systems and the way in which the semantic web can help in providing these benefits.

| Characteristic | Classical learning | eLearning | The semantic web |
|---|---|---|---|
| Delivery | **Push**<br>Instructor determines agenda | **Pull**<br>Student determines agenda | Knowledge items (learning materials) are distributed on the web, but they are linked to commonly agreed ontologies. This enables construction of a user-specific course, by semantic querying for topics of interest. |

| Characteristic | Classical learning | eLearning | The semantic web |
|---|---|---|---|
| Responsiveness | **Anticipatory**<br>Assumes to know the problem | **Reactionary**<br>Responds to problem at hand | Software agents on the semantic web may use a commonly agreed service language, which enables co-ordination between agents and proactive delivery of learning materials in the context of actual problems. The vision is that each user has his own personalized agent that communicates with other agents. |
| Access | **Linear**<br>Has defined progression of knowledge | **Non-linear**<br>Allows direct access to knowledge in whatever sequence makes sense to the situation at hand | Users can describe the situation at hand (goal of learning, previous knowledge,…) and perform semantic querying for the suitable learning material. The user profile is also accounted for. Access to knowledge can be expanded by semantically defined navigation. |
| Symmetry | **Asymmetric**<br>Training occurs as a separate activity | **Symmetric**<br>Learning occurs as an integrated activity | The semantic web offers the potential to become an integration platform for all business processes in an organization, including learning |

| Characteristic | Classical learning | eLearning | The semantic web |
|---|---|---|---|
| | | | activities. |
| Modality | **Discrete**<br>Training takes place in dedicated chunks with defined starts and stops | **Continuous**<br>Learning runs in the parallel to business tasks and never stops | Active delivery of information (based on personalized agents) creates a dynamic |
| Authority | **Centralized**<br>Content is selected from a library of materials developed by the educator | **Distributed**<br>Content comes from the interaction between the participants and the educators | The semantic web will be as decentralized as possible. This allows an efficient combination of information. |

| Characteristic | Classical learning | eLearning | The semantic web |
|---|---|---|---|
| Personalization | **Mass produced** Content must satisfy the needs of many | **Personalized** Content is determined by the individual user's needs and aims to satisfy the needs of every user | A user, using his/her personalized agent, searches for learning materials customized for his/her needs. The ontology is the link between user needs and characteristics of the learning material |
| Adaptivity | **Static** Content and organization/taxo -nomy remains in their originally authored form, without regard to environmental changes | **Dynamic** Content changes constantly through user input, experiences, new practices, business rules and heuristics | The semantic web enables the use of distributed knowledge provided in various forms, enabled by semantic annotation of content. The distributed nature of the semantic web enables continuous improvement of learning materials. |

**The second chapter** is a review of the components of the infrastructure that represents the final objective of this thesis. It presents the main components of the platform, as well as the most important algorithms proposed and implemented within the infrastructure.

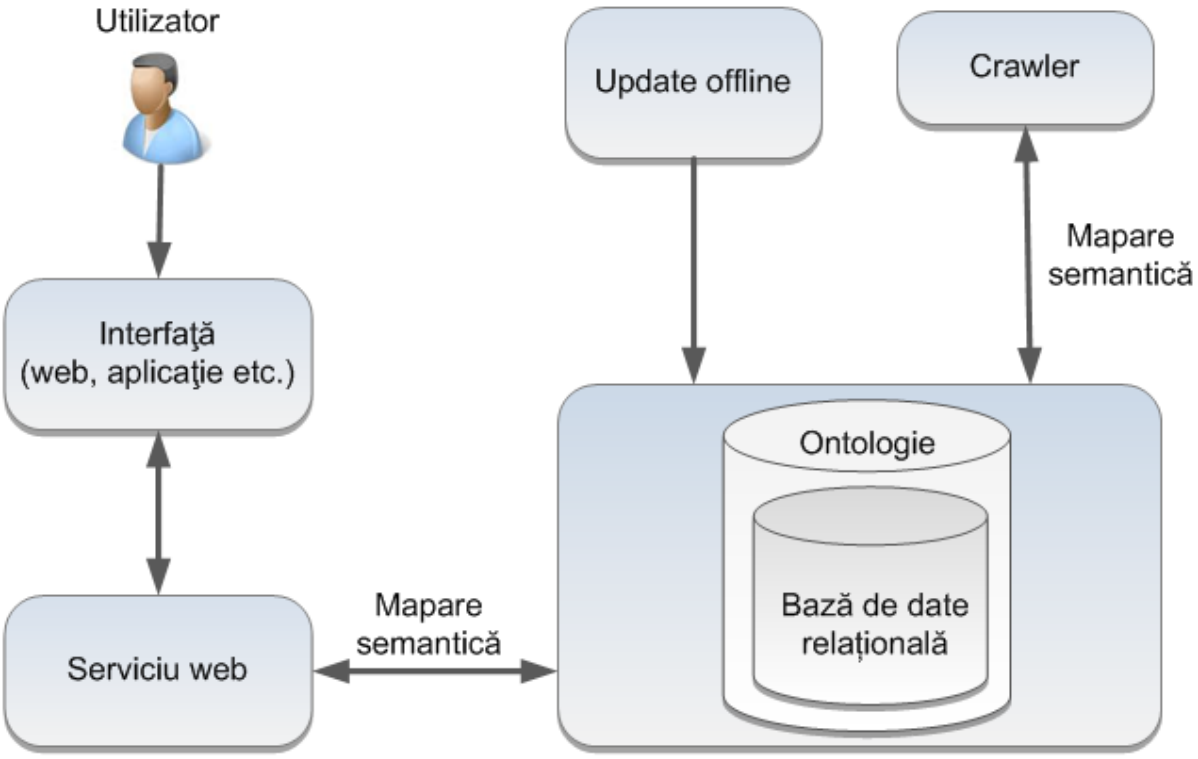
**The third chapter** consists of an analysis of the current state of research in the field of the semantic web, with emphasis on the formal logics used for deducing statements regarding concepts and the relationships between the concepts (a crucial aspect for an efficient implementation of the semantic web). Next, the paper presents a comparative study of RDF Query Languages, as well as a comparative study of the emerging semantic eLearning technologies.

**The fourth chapter** presents the concepts and the technologies used for implementing the eLearning infrastructure developed in the thesis. The chapter also includes a detailed description of each component implemented within the infrastructure. This includes the design and implementation of a component destined for assisted teaching, which includes the identification of the domain of interest, based on training and classification algorithms, as well as a component that provides learning materials for the identified domain of interest.

The domain is established by using the ontology of domains implemented within the infrastructure, which contains a hierarchy of 330 domains, out of which 289 are “leaf” domains (leaf nodes of the tree). These nodes play an important role within the infrastructure, since all the learning documents are linked to exactly one these domains.

The fourth chapter also includes a review of the algorithms used for preprocessing the training documents and the texts inputted by users, before applying the classification algorithms.

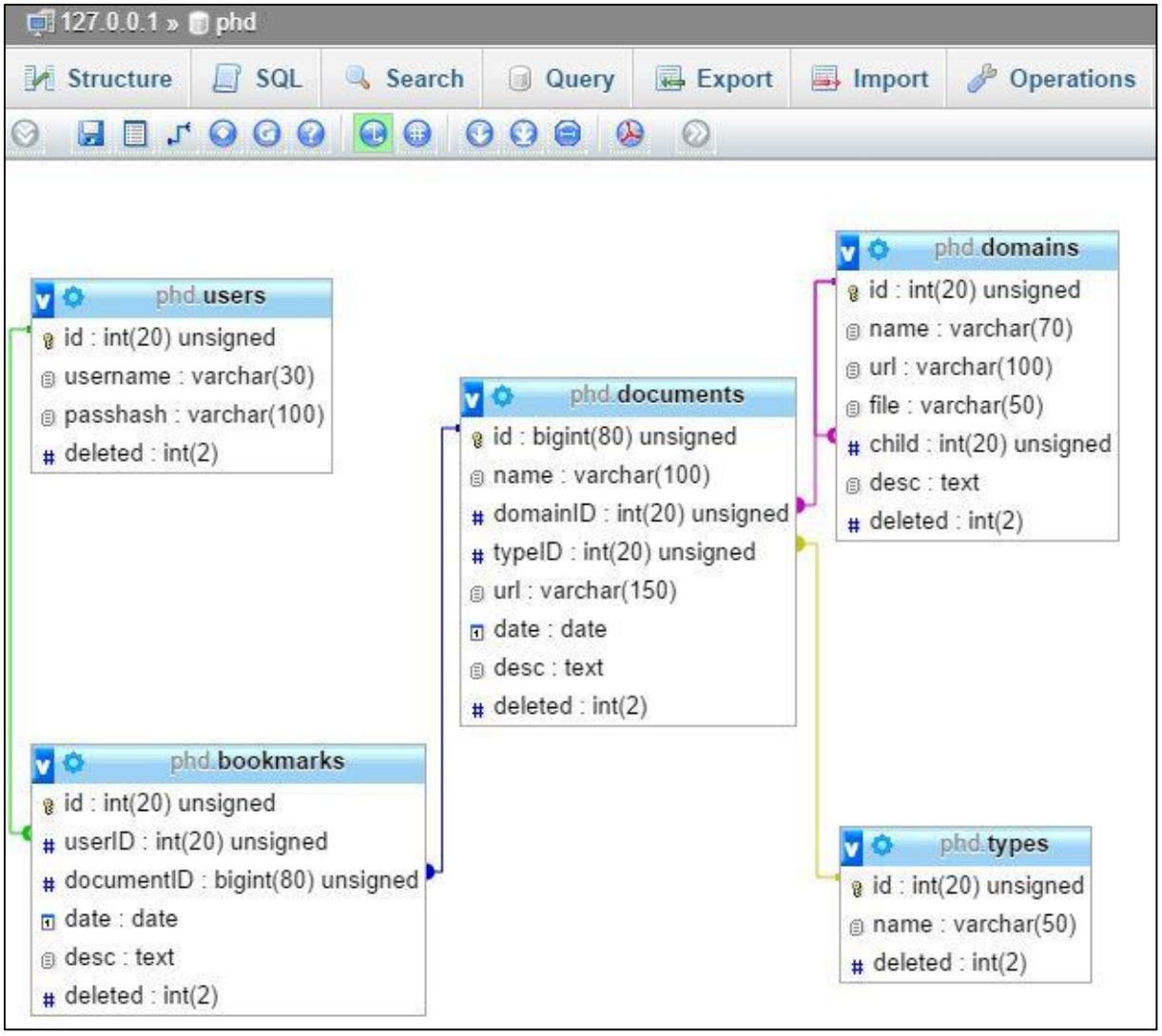
The following figure presents the components of the implemented eLearning platform.




The knowledge base of the infrastructure consists of a relational database, which contains information about the ontology of domains used within the platform, the learning materials provided and the users that have access to the platform.

A very important component of the infrastructure is the ontology of domains, on the basis of which the learning materials are classified. It is represented as a tree of domains and sub-domains. Each “leaf” node has its own training document, used by the classification component, in order to identify the user’s text’s domain. The training document is a text file that contains relevant information for that particular domain.

The learning materials can be of various formats, namely video, audio and text and they are associated to exactly one “leaf” domain. All users have access to any learning material provided by the infrastructure, which can also be bookmarked by users, for a later review.




The figure above presents the structure of the database. The domain hierarchy is represented in the table *phd.domains,* which contains information about the name of the domain, its description, as well as the path to the training document associated with the domain (if the domain is a "leaf" domain). In order to ensure the hierarchical representations of the ontology of domains, the table contains a foreign key that points to another record of the same table (*child* references *id*).

Besides the initial learning materials provided by the platform, a component destined for automatic search of learning materials has also been designed and implemented. This component is called a **Crawler,** and its purpose is to browse the internet in order to find new learning materials.

The web can be considered a graph in which the pages represent the nodes of the graph, and the links contained within the pages represent the edges that connect the nodes. To browse the web means to visit every node of this graph.

The crawler does breadth-first searches (BFS) of connected, non-cyclical sub- graphs found within the graph. This means parsing a web page and analyzing its content, in order to identify the links contained within. If the visited page is in HTML format, the crawler extracts all the links contained within the page and adds them to the breadth-first search queue. Pages that contain links to other sites queued for visiting will be interior nodes of the graph, while the others will be leaf nodes. Besides analyzing the html content of the page, in order to identify the links, the crawler also classifies the content found within the html (inside the <html> tag), based on the domains defined in the ontology. The classification is done by the web service implemented within the infrastructure.

The web service's task is to identify the domain of the text inputted by the user, for which the platform will suggest learning materials.

For establishing the domain of interest the text goes through two stages. The first stage consists of pre-processing tasks, such as stemming the words and eliminating the words considered irrelevant for the classification. For this, the platform implements Porter's Stemming Algorithm proposed in the 1980 article "An Algorithm for Suffix Stripping" and an algorithm based on the statistical Chi-Square test. The second stage consists of using actual classification algorithms, including a new algorithm that combines results of various classifiers, in order to identify the domain of the text.

**The fifth chapter** is reserved for the detailed presentation of the text classification algorithms implemented within the infrastructure.

Supervised classification of a document implies identifying the class it belongs to, based on a set of classes and a set of training documents associated with each class.

The eLearning platform proposed in this paper implements three well known text classification algorithms, k-Nearest Neighbour, k-Nearest Neighbour + TF-IDF(Term Frequency – Inverse Document Frequency) and the Naïve Bayes classifier.

The fifth chapter contains a detailed presentation of all these algorithms, including their formal definitions and their representation as logical schemas and as pseudocode.

The results of these classifiers are combined, using the new algorithm proposed in this paper, thereby creating an overall classifier that performs better than each of the classifiers, used individually.

The output of each classifier is represented as a vector, with the $i^{th}$ component of the vector representing the rank of the $i^{th}$ class in the ranking of classes produced by that particular classifier. Then, the score of each classifier, will be calculated as the sum of the distances between the vector associated with that particular classifier and all the vectors associated with all the other classifiers. The winning ranking will be chosen as the one that has the lowest score, thereby improving the classification process, in comparison to using each classification method separately. This is demonstrated in the study of the performance of the algorithm proposed in this paper, which can be found at the end of the fifth chapter.

The thesis ends with a short section of **Conclusions,** which also contains directions for future research.

The **Bibliography** contains relevant titles for the proposed topic.

The **Annexes** section contains the code files used for programming the components of the platform, as well as a list of papers published during the PhD period.