Ontology-based Data Extraction in the Scholarship-Related Content

Anna Yushtina

Master's Thesis
Master of Applied Computer Science
30 ECTS
Department of Computer Science and Media Technology
Gjøvik University College

Avdeling for
informatikk og medieteknikk
Høgskolen i Gjøvik
Postboks 191
2802 Gjøvik

Department of Computer Science
and Media Technology
Gjøvik University College
Box 191
N-2802 Gjøvik
Norway

Ontology-based Data Extraction in the Scholarship-Related Content

Anna Yushtina

2013/11/27

Abstract

Master Thesis on the topic "Ontology-based Data Extraction in the Scholarship-Related Content" is concentrated on the area of ontologies and on the research of the methods by which ontological concepts can be recognized in the text, enhancing its semantic meaning.

The use of ontologies can significantly improve semantic richness of the texts presented on the Web, but to be able to exploit all their capabilities, specific XML-based notations must be written to describe each and every resource. This is usually quite a big amount of human work, and the Thesis is seeking for the ways to decrease the amount of human resources, either by suggesting automatic or semi-automatic approaches for ontology-based information retrieval.

In the experiments conducted in the domain of scholarships, ontology for scholarships has been thoroughly evaluated, and the names of the disciplines were chosen as a target area for the further information retrieval research.

Discovery of the ontological concepts in the text was performed by, first, scraping the webpage for the target section, and then by implementing Boolean search method with and without prior preprocessing. Such approach demonstrated very good results, and with preprocessing roughly 70% of the disciplines were retrieved. Furthermore, extension of the ontology has been proposed as the way to increase extraction rate by 10%. Overall, 80% of the disciplines can be retrieved by our method.

Preface

I would like to say thank you - Thank You! - to all of the people who supported me in the process of writing my Thesis, people who believed in me at the times I myself did not. First and foremost I want to say thank you to my dear parents, whose support, love and caring I felt even from a land "Far Far Away", and special Thank You goes to my Dad, who was regularly sharing his valuable insights with me, inspiring, motivating and encouraging me to move forward. I love both you and Mom so much, and I wouldn't have been able to do it without you!

I want to say thank you to my supervisor whose skilful and supportive guidance and valuable expertise helped me to get where I am now, and for teaching me how important it is not to give up when it gets hard and you feel lost. Because it does get better - and now I know it. I believe that Thesis made me stronger, and I have inner faith that there is nothing I can't do, and that "Impossible is nothing" - if you work as hard as you can and never, ever give up, because most often you feel like giving up when you are the closest to your goal.

I also want to say thank you to my friends, Mariia, Vasilisa, Iryna, Parinaz, Eli, Xinwei, Gerardo, Andrew, Ruslan - and everyone else. You were my cheerleaders, and you believed in me more than I did. I can't express how grateful I am!

Contents

List of Figures

1 Introduction

1.1 Topic

The topic of the Thesis is concentrated on the area of Semantic Web in general and ontology-based information retrieval in particular.

The concept of the Semantic Web first appeared at the beginning of 21st century and, according to the T. Berners-Lee et al. [4], Semantic Web "is not a separate Web but an extension to the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

Semantic Web has emerged in a form of the evolutionary step of Web "eras": Web 1.0, Web 2.0 and, finally, Web 3.0, which Semantic Web is being referred as the central element of, [5]. While Web 1.0 represents static-fashioned approach to the websites, suggesting all users the role of inactive consumers of the content, Web 2.0 presents them the opportunities to generate the content themselves [6], introducing the concept of "Read-write Web" [7], a dynamic environment where everyone can be both a creator and consumer - and "sharer". Web 2.0 is a Web of Social Networks, the Age defined by the "wisdom of crowds", it "harnesses collective intelligence" and turns going-online programs and devices into web services, [8].

Web 3.0, on the other hand, as Jim Hendler suggests [5], can be viewed as "Semantic Web technologies integrated into, or powering, large-scale Web applications". So, what exactly is Semantic Web then, and what are its most distinguishing features?

Semantic Web is promoting the idea of "Web of Data", Linked Data, in contrast to traditional concept of "Web of Documents", [9], allowing documents to include not only plain data, but also "metadata" - data about the data, that adds semantic meaning to the hypertext. In general, World Wide Web Consortium (W3C) explains Semantic Web as following [10]: "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". The main tools of Semantic Web are RDF (Resource Description Framework) and OWL (Web Ontology Language). RDF is used to describe resources by the rules defined in RDFs (RDF Schema) or OWL ontologies. Both RDF(s) and OWL are written in XML (XML Extensible Markup Language).

According to Studer et al. [11], "An ontology is a formal, explicit specification of a shared conceptualization", its purpose is to define concepts and relationships between them in a certain domain of knowledge. One can view ontologies as a set of rules by which certain resource should be described. Single ontology could be used for describing infinite number of resources, unifying and structuring information about them, making it easier for machines to "understand" their meaning, group and retrieve pieces of information requested by users in efficient fashion.

Ontology-based Information Retrieval extends trivial capabilities of Information Retrieval methods by adding semantic meaning to the retrieved data, e.g., when we confirm that certain term refers to the certain instance in the ontology, it is automatically inferred that the retrieved

term is an instance of the higher class that is described in the ontology.

1.2 Problem Description

Even though the concept of Semantic Web appeared more than ten years ago, the topic is still quite under-researched and semantic tools are not widely used. As N. Shadbolt et al. wrote in 2006 [12], "This simple idea ... remains largely unrealized", and seven years later, the situation has changed only at the slightest. Ontologies are not commonly used on the vast majority of web resources, and there are not many ontologies that can be accessed publicly by direct search in Google, moreover, semantic engines designed for their search seem to be mostly outdated.

The main benefit of using ontologies on the Web is to enhance semantic meaning of the data presented in the documents - but ontologies alone are not able to provide extensive description of resources, that is what RDF-descriptions are for. RDF-descriptions are specific documents that describe the data presented on the webpage "by the laws" defined in the ontology in RDF (XML) format. RDF-descriptions should be written for each document (resource) individually, and although the creation of ontology doesn't require too extensive exploitation of human resources (at least it is one-time affair), writing descriptions to the documents is rather long and resource-demanding process. The Thesis aims to research the ways of automatizing or semi-automatizing the process of creation RDF-descriptions on the example of ontology and web-portal for scholarships, concentrating on a narrow issue of extracting information about the names of the disciplines (according to the ontology disciplines' names) from the website.

The area of scholarships has been chosen particularly because organizing and structuring information about different funding opportunities can help young people around the world to study abroad. Student mobility is becoming very common, especially between European countries, which are involved in Bologna Process that insures fulfilment of unified standards in education. Due to the open borders in Shengen area, the possibilities to travel to foreign countries to study are enormous for the most of Europeans, and the only thing that can stop people from exploring life "beyond" their homes is money. Scholarships often is the one and only solution for many, and letting more people know about such offers can make a significant impact on the exchange rates, encouraging fair competition among prospective students and giving them unique opportunity to get a taste of a different kind of life.

There are many scholarships portals that provide information about funding opportunities for young people all over the world, helping students and universities in fostering multicultural exchange - but, unfortunately, the data sometimes can be cluttered and repeated across different sources. That is where the need for scholarships' ontology came from, the substance which aimed to combine, unify and structure the information from different platforms for its further representation in a desirable and efficient for the end-user way. Ontology for scholarships, that was created in the course of Advanced Project Work [13], "Scholarship Ontology", has been designed to be able to describe major concepts of the scholarship-related domain of knowledge. Ontology has been created using Protégé editor, it has 78 classes and 33 properties (23 Object properties and 9 Data properties), and more than 700 individuals of classes.

1.3 Research Questions

The questions which are aimed to be answered in the Thesis, include:

1. What features should ontology for scholarships have, and how to evaluate and modify existing ones?
2. What kinds of information retrieval methods can be used for extracting knowledge in the scholarship content by the use of domain ontology?
3. How preprocessing can improve the results of the information extraction in the scholarship-related domain?

1.4 Motivation and Beliefs

Multicultural student exchange should be open to everyone who wants to experience living and studying abroad even for a fixed period of time. First of all, it is beneficial for young people whose character is being formed when interacting with others, and the more diverse views and culture those have, the more broad-minded, liberalized and understanding the person becomes. People can widen their life perspective by being concerned not just about the country they live in, but about the whole world altogether, learning to analyze, understand and care, and realizing that people from different countries are more similar to each other than TV news are used to demonstrate. Student exchange can really change lives, help young people to grow up, get independent and learn to take responsibility for their actions - the qualities which will create better society.

With so many amazing funding offers available for students, it is a pity that information about them is not thoroughly spread, well-known and introduced. One step in a right direction would be to organize and structure such information by the use of ontologies, so that it could be easily searched and browsed. That was a core motivation for creating Scholarship Ontology, which was designed for structuring available information about scholarships' announcements from different sources, with its further reorganization in a beneficial for perspective students way. Nevertheless, ontology alone is not a panacea, and it is not going to solve all existent problems on its own - it's not a solution by itself, rather than an algorithm, methodology for the solution, which is useless unless further works are performed.

The motivation for writing this Thesis has come from realization that although Semantic Web has a high potential, its principles are not yet commonly used, and one of the reasons is that it requires a lot of manual work, which can be quite costly. Defining the set of rules that describe the information structure in the certain domain (either in the form of ontology or RDF Schema) might be one-time task for the specialist, but describing continuously emerging resources by the help of these rules is time-demanding (although doesn't require high qualifications). That is why the idea of the Thesis is to try to automate or at least semi-automate the process of creating RDF-descriptions on the example of one specific website and domain ontology.

There can be found quite a lot of research on how ontology can be beneficial for data extraction, but the majority of them focus on the creation of complex systems for data mining, and there are no researches found for testing of the performance of straight-forward "baseline" information retrieval methods, such as analysis of Boolean search for strings matching, etc. Moreover, the

impact of preprocessing on the overall information retrieval performance has to be thoroughly researched as well.

The core belief is that combining and structuring information from distinct web resources by describing them with the help of ontologies will contribute to solving the problem of "data overload", which we are experiencing today due to the ever-growing amounts of web data. Moreover, the more people will integrate ontologies in their web solutions, the more chances, due to the Network Effects, for ontologies to become sort of a standard for handling newly published content.

1.5 Planned Contributions

Planned contributions of the Thesis will be:

1. Evaluation and modification of the Scholarship Ontology;
2. Analysis of the performance of the baseline information retrieval methods that are suitable for the data mining in scholarship-related content;
3. Analysis of the impact of Natural Language-based preprocessing techniques on the overall information retrieval results;
4. Results obtained in the course of the Thesis could serve as an example for future analysis of retrieval of other kinds of information based on ontology, and could help in implementation of similar research for other website - or the website which structure simplifies the ontology-based information retrieval could be developed.
5. Improved Scholarship Ontology itself.

2 State of the Art

2.1 What features should ontology for scholarships have, and how to evaluate and modify existing ones?

There can be defined different approaches to the ontology evaluation, and, basically, all of them describe the set of features that ontology should have. The theoretical base for ontology evaluation is presented below. The chapter is a partially rewritten and extended version of the text written in the Research Project Planning Course, [14].

There are different approaches to the evaluation of ontologies. Janez Brank et.al. defines the following: (2005, [1]):

1. *"Golden standard" evaluation:* ontology is being compared to the "golden standard" (often it is an ontology itself) [15];
2. *Application-reliable evaluation:* ontology is being used in specific application and its output results are afterwards evaluated [2];
3. *Data-based evaluation:* "expressiveness" of the ontology is being measured towards the certain corpus of data (documents, etc.) [16];
4. *Human-based evaluation:* experts decide what kind of evaluation they want to perform. Often several different types are combined together to achieve specific evaluation goals, [3].

Also in their work, Janez Brank et.al. stresses upon different layers of the ontology that might be evaluated, and gives short summary on what evaluation approaches are better to use for each of them (2005, [1]). So, the layers of the ontology that can be evaluated, are:

- *Lexical layer:* evaluation of terms, concepts and instances of the ontology;
- *Hierarchical layer:* evaluation of the consistency of the taxonomy, classes and subclasses (concepts with "is-a" relationships);
- *The level of semantic relationships of other kind* (that is usually what we call Object and Data Properties);
- *Context level:* in case if external ontologies are being reused, their content must be evaluated as well;
- *Syntactic level:* consistency of manually designed ontologies should be tested and evaluated;
- *Architectural level:* architecture of the ontologies must be evaluated.

Janez Brank et.al. (2005, [1]) suggest the following table that summarizes what kinds of evaluation are better to use when evaluating ontology on different levels, Fig. 1. Footnote 1 refers to using "Golden standard" evaluation method, comparing target ontology's syntax with the one of golden standard's.

Level	Approach to evaluation			
	Golden standard	Application-based	Data-driven	Assessment by humans
Lexical, vocabulary, concept, data	x	x	x	x
Hierarchy, taxonomy	x	x	x	x
Other semantic relations	x	x	x	x
Context, application		x		x
Syntactic	x[1]			x
Structure, architecture, design				x

Figure 1: Suitable evaluation approach with regard to ontology layer, [1]

Ontology evaluation methods are described more in-depth further.

Evaluation towards ”Golden Standard”

”A Task-based Approach for Ontology Evaluation” [2], a paper by Robert Porzel and Rainer Malaka, suggests to evaluate ontologies based on their performance rates compared with the ones of "Golden standard"'s. On practice, "Golden standard" is usually just a set of previously annotated answers.

Figure 2 shows practical implementation of the approach. To test this method, one needs to have specific application where they can test response of the ontology, which also refers to application-reliable evaluation to some extent. This method can be used for testing single ontology as well as multiple ontologies, in the last case, their "success rates" towards golden standard can be compared to each other.

Given experiment proved to have several shortcomings [2]:

- *insertion errors* that indicate superfluous concepts;
- *deletion errors* that indicate missing concepts;
- *substitution errors* that indicate ambiguous and off-target concepts.

”Golden standard” evaluation has also been performed by Maadche and Saab [15] who suggested another kind of its implementation: authors were trying to find similarities between several ontologies, therefore, one of the two ontologies was used as a "golden standard".

Application-reliable evaluation

The goal of the application-reliable evaluation method is to test performance of the ontology when working on certain tasks. Hence, data that being tested are the output of a certain application after using target ontology with it. Already mentioned Robert Porzel and Rainer Malaka [2], even though they were implementing ”Golden Standard” technique, were performing the actual comparison by the means of external application, so the method they used can be considered combined.

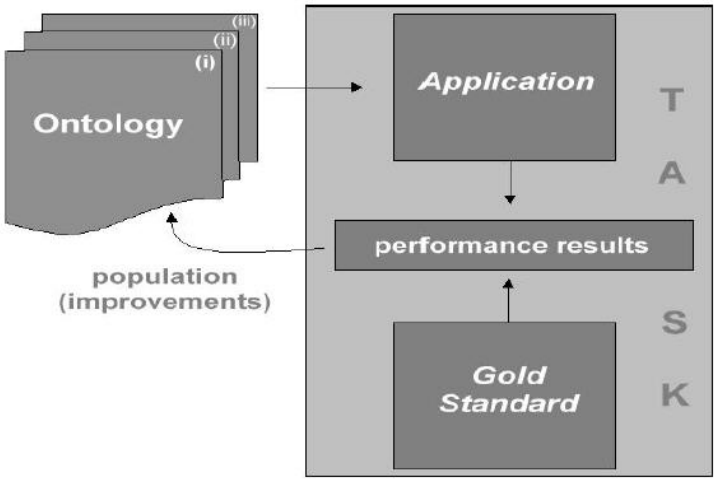


Figure 2: Test setup by Robert Porzel and Rainer Malaka: ontology evaluation towards golden standard, [2]

Nevertheless, application-reliable evaluation was proven to have a number of drawbacks [1]:

- When two ontologies are under the test, they both need to be designed in the same way in order to be evaluated fairly;
- Since application approach is task-oriented, it's often hard to draw conclusion on whether ontology is generally expressive, all we can state is the level of expressiveness when performing certain task;
- Internal processes of the performance of the application have to be known to the specialists in order to conclude whether results of the evaluation are valid and reliable.

Data-based evaluation

Data-based evaluation refers to comparison of the target ontology to the certain data that contains information about the ontology's domain of knowledge. Data is usually represented in a form of textual documents (corpus), which is one of the most accessible forms of knowledge [16]. This method measures the extent of correspondence between the terms in ontology and actual textual representation of the particular knowledge domain. Brewster at al. [16] in their experiment, for example, extracted a set of the relevant concepts from the given corpus with their following comparison to the terms of ontology. The amount of overlap was then measured.

Human-based evaluation

Human-based evaluation, as can be seen from the title, in formed with the help of human judgement and, therefore, usually combines different approaches "under one roof". Often results obtained from different methods are then "added" in a form of the weighted sum of per-criterion scores that are then calculated (2005, [1]). Many different criteria can be considered, for example, Lozano-Tello and Gómez-Pérez, [3], suggest 117, organized in five separate groups, "di-

DIMENSION: CONTENT *CHARACTERISTIC*	*TYPE*
CONCEPTS (FACTOR)	**(very_low, low, medium, high, very_high)**
Essential_Concepts	(very_low, low, medium, high, very_high)
Essential_Concepts_In_Superior_Levels	(very_low, low, medium, high, very_high)
Concepts_Properly_described_In_NL	(very_low, low, medium, high, very_high)
Formal_Specification_Of_Concepts_Coincides_With_NL	(very_low, low, medium, high, very_high)
Attributes_Describe_Concepts	(very_low, low, medium, high, very_high)
Number_Of_Concepts	(very_low, low, medium, high, very_high)
RELATIONS (FACTOR)	**(very_low, low, medium, high, very_high)**
Essential_Relations	(very_low, low, medium, high, very_high)
Relations_Relate_Appropriate_Concepts	(very_low, low, medium, high, very_high)
Formal_Specification_Of_Relations_Coincides_With_NL	(very_low, low, medium, high, very_high)
Arity_Specified	(very_low, low, medium, high, very_high)
Formal_Properties_Of_Relations	(very_low, low, medium, high, very_high)
Number_Of_Relations	(very_low, low, medium, high, very_high)
TAXONOMY (FACTOR)	**(very_low, low, medium, high, very_high)**
Several_Perspectives	(very_low, low, medium, high, very_high)
Appropriate_Not-Subclass-Of	(very_low, low, medium, high, very_high)
Appropriate_Exhaustive-partitions	(very_low, low, medium, high, very_high)
Appropriate_Disjoint-partitions	(very_low, low, medium, high, very_high)
Maximum_Depth	(very_low, low, medium, high, very_high)
Average_Of_Subclasses	(very_low, low, medium, high, very_high)
AXIOMS (FACTOR)	**(very_low, low, medium, high, very_high)**
Axioms_Solve_Queries	(very_low, low, medium, high, very_high)
Axioms_Infer_Knowledge	(very_low, low, medium, high, very_high)
Axioms_Verify_Consistency	(very_low, low, medium, high, very_high)
Axioms_Not_Linked_To_Concepts	(very_low, low, medium, high, very_high)
Number_Of_Axioms	(very_low, low, medium, high, very_high)

Figure 3: Characteristics of the ”Content” dimension of ontology evaluation criteria developed by developed by Lozano-Tello and Gómez-Pérez, [3]

mensions". Those dimensions are: content, language, methodology, tool and costs. Fig. 3 shows example of the Content Dimension characteristics.

Some authors describe other approaches to the ontologies' evaluation, [17]:

1. Logical (Rule-based);
2. Metric-based (Feature-based);
3. Evolution-based.

Logical Evaluation

Logical Evaluation, or Rule-based evaluation, uses rules that are defined inside the ontology and tests ontology for consistency. For example, if we say that certain property is functional, it means that for one domain value there will be only one range value. In the Scholarship Ontology, we have functional property "isLocatedIn" which has domain "Educational Institution" and Range "Location", which means that Educational Institution cannot have multiple locations.

Logical Evaluation checks ontologies for consistency, and is already embedded into the functionality of such ontology design software as Protégé [1] in a form of so-called Reasoners.

Metric-Based Evaluation

Metric-based Evaluation represents quantitative approach to the ontology analysis. Lozano-Tello's and Gómez-Pérez's [3] method with consideration of 117 distinct criteria described above is an example of such evaluation.

Another examples include works of H.Alani et al. [21], "Ranking ontologies with AKTiveRank" where authors retrieve ontologies based on their "relatedness" to the topic (term) suggested by the user.

[1] http://www.protege.stanford.edu/

Evolution-based Evaluation

We know that ontologies' evaluation has an iterative nature, that is, it has to be performed regularly, since with time certain concepts could lose their importance, or some new ones could evolve. Evolution-based evaluation describes this "timing" feature of ontologies, which, according to N.F.Noy et al., [18], can be defined by:

- Change in the domain: occurs when certain changes in a real world take place, and the knowledge model represented by ontology, correspondingly, has to be modified as well;
- Change in the conceptualization: occurs when the change of the viewpoint for the domain description is required, since any topic can be described in different ways and from different angles;
- Change in explicit specification: occurs when the language ontology is written in, has to be "translated" into another one, and preserving the semantics during such "conversion" can lead to certain problems.

Evolution-based evaluation has been performed by P. Plessers et al. [19] who used a version log for ontology change detection and P.Haase et al. [20], who were researching inconsistencies in the changing ontologies.

2.2 What kinds of information retrieval methods can be used for extracting knowledge in the scholarship content by the use of domain ontology?

There are two main approaches to the query-based document search, [22]:

- Statistical Search. In Statistical search results are retrieved and ranked statistically, as to to what extent document matches the query.
  In statistical methods query is considered to be a simple set of terms (words) in the document. Very often the terms undergo preprocessing: the words can be stemmed (lemmatized) in a way that the form of a user query is being matched with all forms of the queried word, [23], e.g., the word in the query could be "agricultural", but after preprocessing also other word forms will be considered ("agriculture"). Another form of preprocessing is creation of stop-word lists (the words that are frequently used in queries but don't really affect search results (such as "which", "that", etc)). Sometimes the terms in the queries are considered as phrases, that could be the case when certain words (terms) appear together in a specific order many times in the collections of documents. Some statistical search mechanisms make use of "n-grams"-search, where the text in a searchable document is divided into the set of n-grams (which are arbitrary strings of n consecutive items(words, characters, etc), [24], Figure 4).
  Following methods fall into this category:
  - Classical and Extended Boolean Search;
  - Vector Space;
  - Probabilistic.

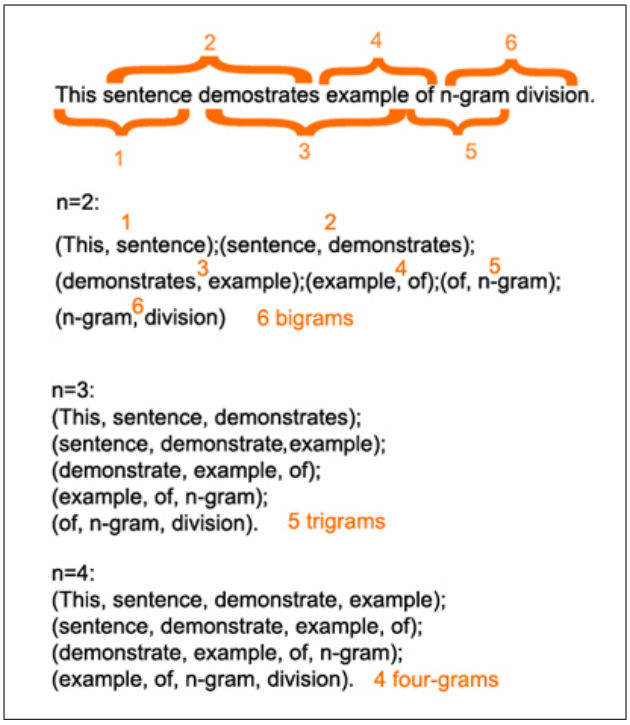


Figure 4: Division of text into n-grams

- Semantic Search. In Semantic search results are retrieved based on certain extent of syntactic and semantic analysis, and Natural Language Processing (NLP) techniques are exploited. Nevertheless, Semantic Search methods are usually used in conjunction with Statistical Search ones.

2.2.1 Classical and Extended Boolean Search

What are capabilities of Boolean Search?

Boolean Search (or Boolean keyword search) is a search method based on the Boolean Retrieval Model, information retrieval model where the query is given in the form of Boolean expressions of terms. Boolean expressions consist of search query terms combined with Boolean operators AND, OR and NOT, [25]. By this method, each document is considered to be a combination of words.

What types of queries can we generate using Boolean Retrieval Model? There can be many, Figure 5, [26]:

- AND - default operator, search query A AND B means that we are searching for both words to be existent in the documents, that is why, the more terms we are looking for, logically, the less results we will be getting. On practice, AND is a default operator, which means that if a query is written in a form A B, there is considered to be "AND" between them;
- OR - using operator OR between two terms, A OR B, means that the search will be conducted for finding the documents that consist either term A, or term B (or terms A and B together).

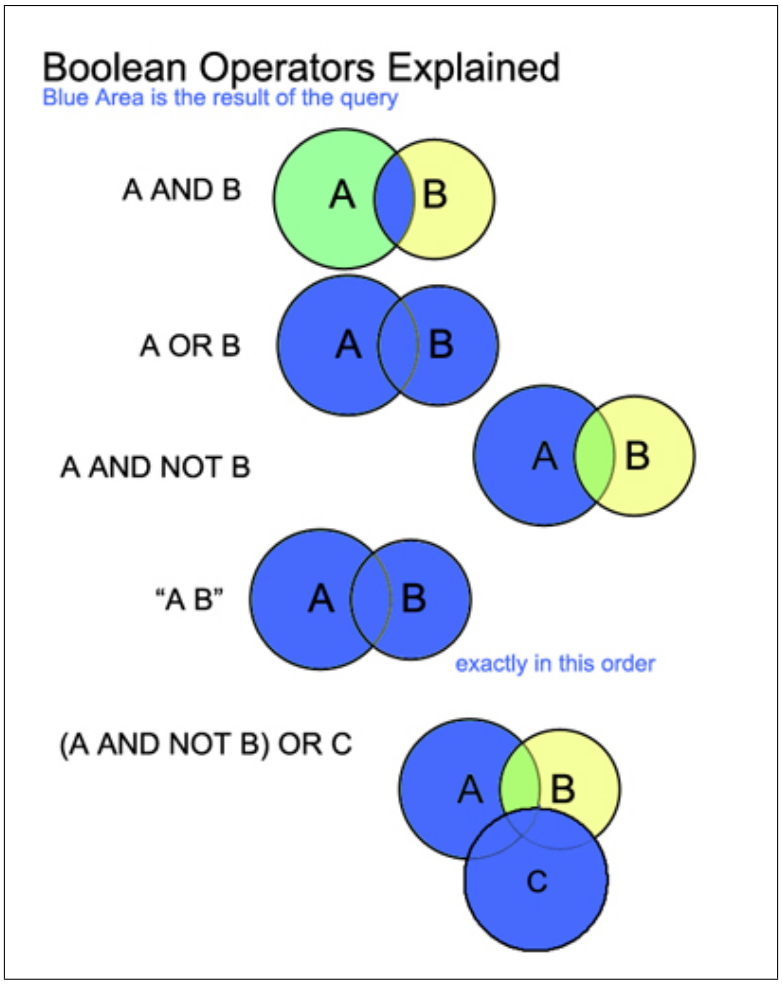


Figure 5: Boolean Operators Explained

On the contrary from AND operator, the more search terms separated by the operator OR we are having, the more results will be retrieved. On practice this operator is frequently used when trying to find a concept which has several synonyms;

- NOT (AND NOT) - since it usually doesn't make sense trying to search for the documents that don't contain the query (since it will be "all but"), on practice joined operators "AND NOT" are used, e.g., A AND NOT B;
- Phrasal Search: all words of the query are mandatory to be found in the document (consider AND operator between the terms), but also in the exact given order. This type of query usually is represented by quotes, "search query";
- Including/excluding words to/from the query results, e.g., (search query +/-specificResults). When searching Web, including specific words can be particularly helpful when dealing with frequently used words (common words such as "which", "that", etc);
- Nesting: parentheses group Boolean expressions together, showing the order in which they should be processed, e.g., "(A AND NOT B) OR ("C")".

Specific notions:

- When Classical Boolean Search method is employed with certain preprocessing (lemmatization), searching for one specific word would imply searching for all its other wordforms as well (example could be searching for the word "decompose" and retrieving results such as "decomposition", "decomposing", "decomposed", etc);
- Sometimes there can be added proximity operator [27], which tells on what distance from one another the words in the query should be (how many words or sentences should be between the terms, etc). This operator can also specify exact order in which the words in the query should appear.

Evaluation of the IR Quality

In order to be able to evaluate the quality of retrieval, specific measures of precision and recall are used. Precision, which can also be called "positive predictive value" defines the fraction of retrieved results that are relevant. Recall (or "sensitivity"), on the other hand, defines the fraction of relevant instances that are retrieved. The higher values precision and recall have, the higher is the quality of the retrieval. High levels of recall indicate that the majority of relevant results were retrieved, whereas high levels of precision show that the majority of results that were retrieved are relevant.

The concepts of precision and recall emerged as a realization that during any search process there will be four different groups of results:

1. Relevant results that have been retrieved (true positives);
2. Relevant results that have not been retrieved (false negatives);
3. Irrelevant results that were retrieved (fasle positives);
4. Irrelevant results that haven't been retrieved (true negatives).

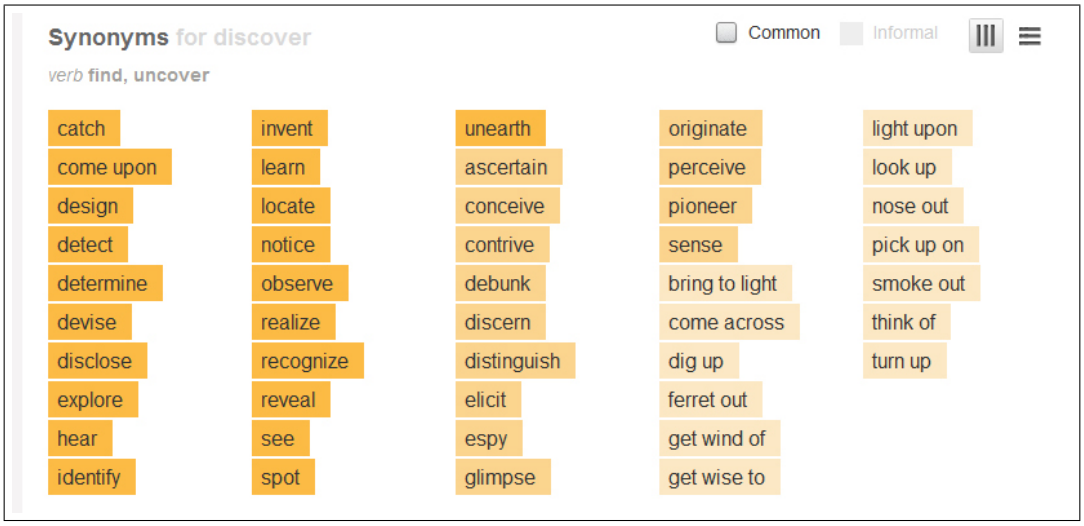


Figure 6: Example of synonyms for the word "discover"

Drawbacks of the Boolean Search

Boolean search is a classical straight-forward approach, it's easy to implement since it's based on the Boolean logic, nevertheless, when it comes to dealing with large amounts of unstructured text, performance can be rather poor. Results fetched by the Boolean Search principle are not accurate, containing many false positives (results that are retrieved but are not relevant) as well as false negatives (results that were not retrieved, even though they are relevant). The reason for that lies in the Natural Language. Boolean Retrieval Model would perform better if every particular word had just one unique meaning, but on practice that is not the case, which is why the concepts of synonymy and polysemy are well-known.

Synonymy, or, in other words, dealing with several words of the same meaning, can cause the system to fail in retrieving results relevant to the search because of not knowing that certain concepts is equal to another one. Example of the synonyms for the word "Discover" is shown on the Figure 6. Synonyms are taken from the English Thesaurus website[2]. Due to synonyms false negatives can appear. On the other hand, polysemy, or having equally spelled words having different meanings, leads to the appearance of false positives.

Polysemy is a major problem which Boolean Search is not able to overcome, since for 200 most polysemous terms in English, the typical verb has more than twelve common senses, and typical noun - more than eight, [28]. For 2000 most polysemous words in English those numbers are eight for verbs and five for nouns. Example of the polysemic word "Degree" is shown on the Figure 7.

Apart from synonymy and polysemy, Boolean search method can include misspelled query terms, or just certain words which are spelled differently in British and American English ("color" and "colour", "grey" and "gray"). Misspelled query terms can be caused either by direct human manipulations, or as a side effect of the scanning and recognition of the text documents.

[2]http://www.thesaurus.com

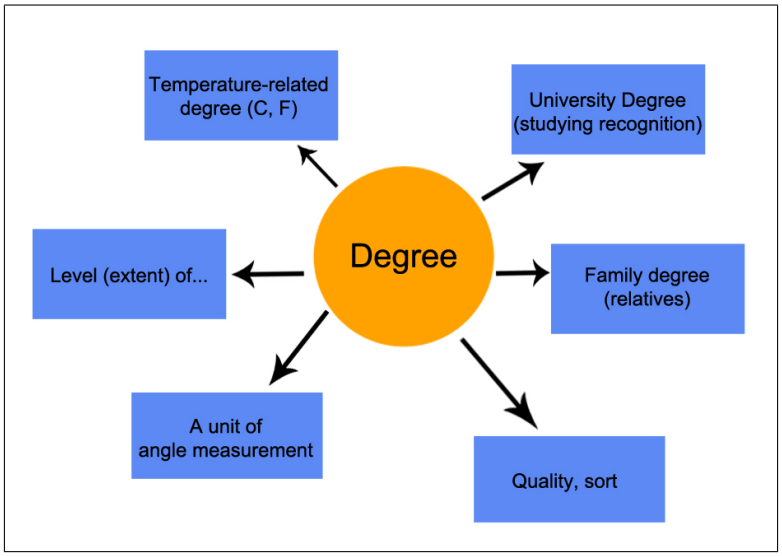


Figure 7: Example of polysemy: the word "Degree" and some of its meanings

Classical vs. Extended Boolean Search

In Classical Boolean Search the documents are searched on a subject of them containing the search query, and there can be only two resulting options: either there is a match of a query (1) or there is no match (0). It means that in case of complex query with AND/OR operators, we will see no difference in results fetched, either there are many terms in one document that match query or just single one; correspondingly, when we will be searching for the set of terms, we will not be able to see if there is all but one terms matched that caused failure of the whole process and mismatch in the output - or there is no similarity between search query and the contents of a document whatsoever, [29].

Extended Boolean Search mechanism aims to fix this major drawback: Extended Boolean operator evaluates query arguments on the scale from 0 to 1 (not just 0 or 1), correspondingly to the extent to which certain expression matches the query.

2.2.2 Vector Space-based Search

In Vector Space-based approach we view document as a "document space", and each document is represented as a set of weights of its terms, [22]. The weights are assigned to the terms according to the frequency of their appearance in the document, e.g., if certain term is not present in the given document, its frequency will equal zero. The purpose of this approach is to determine, which term will be a better descriptor of the contents of the document.

The method is called "Vector-space-based", since it's possible to interpret assigned terms' weights for certain document as coordinates of the document's space. Therefore, we can infer that in vector space approach each document is defined by its terms' weights. In some cases, the term "collection space" can also be used - when the weights of the terms for the whole collection of documents are defined.

It's important to mention that in this approach the query could be given both in usual form

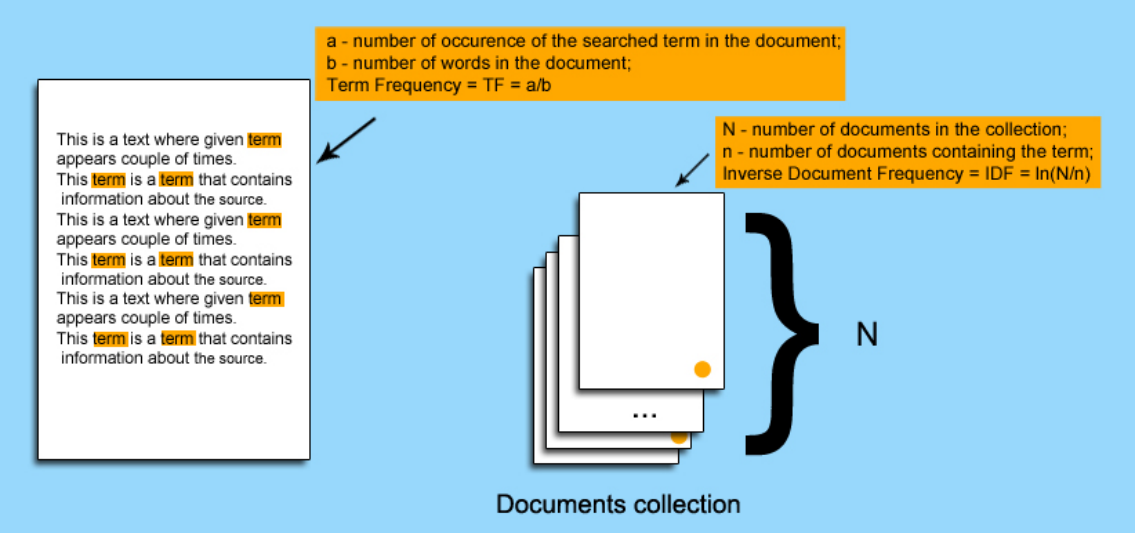


Figure 8: Term Frequency and Inverse Term Frequency Principle

of the set of keywords (terms) or the query could even be a document itself. As in the case of Boolean Retrieval Models, the query may be preprocessed with stemming, stop-words lists, etc. Due to all of the above, the query could be viewed also as just another document in a document space [22]. If there are terms in the query that are not present in the documents, it indicates that these terms form additional dimensions in document space.

What is the principle by which weights are assigned to the terms? Specific weighting scheme, "TF*IDF" is usually used for that. TF/IDF refer to Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency is a frequency of occurrence of the given term inside the given document, therefore, it varies from one document to another, and measures the extent of "importance" of the term inside the document. IDF, or Inverse Document Frequency, on the other hand, is a global measure of the frequency of the term occurrence within the whole document collection, and it indicates the distribution rate of the term in all documents, predicting how likely it is that the term will appear in the document by chance.

IDF =ln(N/n), where N refers to the overall number of documents in the collection, and n - the number of those of them that contain the given term. Figure 8 explains this principle. Therefore, IDF increases when less documents contain specific term (which also usually means that the term is more important). If all documents in the collection contain given term, then N/n=1, and IDF=0, which is logical, since the term is intuitively not particularly important if it is present in every single document.

TF*IDF formula implies that the most "descriptional" terms in specific document are the ones that occur a lot within this document, but only few times in other documents, it means that such terms will have moderate TF*IDF values. The terms with high values will be the ones that occur in most of the documents, and the terms that appear rarely in any of the documents will have the lowest TF*IDF values.

2.2.3 Probabilistic Approach

Probabilistic Approach to Information Retrieval is quite alike in methods to the Vector-Space based one, and rather often it produces results of the same quality. Nevertheless, the major difference between these two methods could be formulated as following, [22]: Vector Space-based approach ranks documents by the similarity measure (which doesn't directly correspond to probability of the term occurrence), whereas Probabilistic approach makes use of clean probability measuring techniques, identifying probability values for each term in the query.

2.2.4 Specific Projects

Automatic Tag Recommendation

P. Alexopoulos et al. in their paper "Exploiting Ontological Relations for Automatic Semantic Tag Recommendation" [30] are looking into the problem of automatic generation and recommendation of semantic tags for text documents by the use of domain ontologies, making use of the internal relations between the concepts. The authors identify the term "tagging" as used with the following goals, [31]:

- to match specific terms and phrases from the document to the concepts in the ontology;
- to find out the topic of the document (by stating whether it refers to particular topic);
- to characterize and summarize the document's content.

The work that has been done by P. Alexopoulos et al. is focusing on two major points: tagging context model and tag recommendation process. Tagging context model calculates the relative importance of the ontological properties for tag identification. Tag recommendation process, on the other hand, determines the concepts of the ontology that potentially can serve as tags for the certain parts of the text (terms, phrases).

Experimental setting was implemented on the example of movie review, taken from Internet Movie DataBase IMDB[3]. The goal was to identify the name of the movie the review was about and, with the highest "confidence score" of 0.084, the movie "Steel" was identified.

Another work is done by Erik Schlyter in his Master's Thesis on the topic "Structured Data Extraction" [32], which concentrates on the implementation and evaluation of the system for Product Information Extraction and Monitor Environment (PIEME). Nevertheless, although the author briefly describes semantically-related part of the problem, developed system focuses mainly on the information extraction part.

2.3 How preprocessing can improve the results of the information extraction in the scholarship-related domain?

2.3.1 Natural Language Processing Methods

Natural Language Processing Methods for Information Retrieval (IR) refer to the specific IR-methods that are based on the knowledge of Natural Language, its structures, terms and words - and the way they are built and used. Natural Language Processing can easily be a preprocessing step for Statistical IR-methods (rarely it is used without them), and can be implemented on seven different levels, [22]:

[3] http://www.imdb.com

1. Phonological - the level of sounds, phonemes. It is used in speech recognition algorithms, and is not useful to text-based information retrieval techniques;
2. Morphological - one of the most commonly used levels of Natural Language Processing. It makes use of the knowledge of the elements of the word: roots, prefixes and suffices. Therefore, example of Morphological IR is stemming, when different forms of word are stemmed to the root, which extends the set of query terms;
3. Lexical - is second most common level of NLP techniques that are used for the Textual Information Retrieval. This level regards words as the smallest elements for analysis. Stop-lists method for eliminating the words of less importance is one of Lexical NLP examples; another one is the use of thesauri and dictionaries of a different kind that aim to help to boost the relevance in retrieved results by, among other methods, adding synonyms to the query. Other, more advanced examples may relate to part-of-speech tagging and proper noun identification techniques;
4. Syntactic - syntactic level NLP refers to the analysis of the structure of the sentences, how they are built and what are their elements;
5. Semantic - aims to analyze sentences and their sets from the point of semantic meaning. Disambiguation of the word sense is also an issue presented on the semantic level, as the sense of the word can't be identified without the context it appears in;
6. Discourse - analysis of text on the level of paragraphs;
7. Pragmatic - analysis of the text by means of external knowledge (it can be general knowledge of the world, data from particular documents, etc.).

J. Pomikálek et al. in their paper "The Influence of preprocessing parameters on text categorization" [33] evaluate performance of different preprocessing parameters, among which - tokenizers, stemmers, stop-lists and others. Experimental datasets included newsgroups as well as conference proceedings. Results showed that Krovetz stemmer, which is considered to be rather "light" one, has slightly outperformed other stemmers that were considered to be more advanced. The authors also concluded that stemming used on unigrams (single n-gram, n=1) in some cases are less effective than when used on bigrams. Nevertheless, it was also concluded that b-gram tokenization works better on longer documents, where it causes significant improvement in categorization.

S. Abels and A. Hahn were looking into the compound words problem in their paper "Preprocessing text for web information retrieval purposes by splitting compounds into their morphemes", [34]. Compounds is the term to describe words that are formed as a combination of several words (usually two), morphemes. Examples could be such words as "afternoon", "rainfall", etc. For the area of web IR, splitting compounds can help to understand the meaning of the text, and can be particularly useful when trying to find synonyms for the word: since it's usually easier to find synonyms to the parts of the compound word than to it itself.

Another positive impact of recognizing and splitting compounds can be seen in direct increase of the search effectiveness: some compounds have different ways of writing them, e.g., one can write "flowerpot" as "flower-pot" or "flower pot", Google, for example, as seen on the Figure 9, has

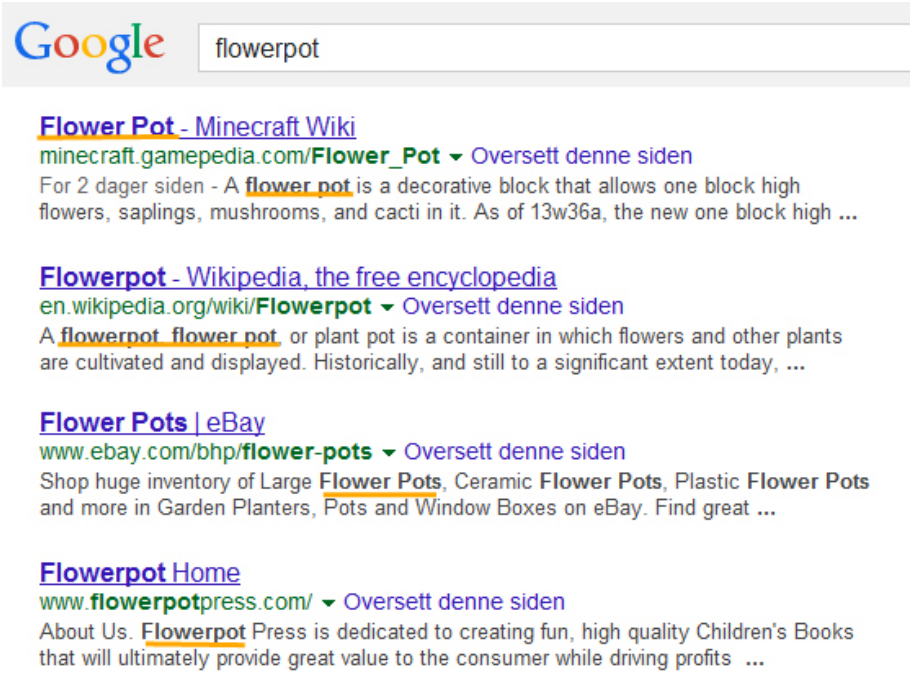


Figure 9: Google Search results for "flower pot"

the algorithm to recognize that both cases refer to the same concept, therefore, when searching for one particular case of writing, it retrieves all possible cases.

The authors suggest the following algorithm for decomposing compounds:

1. Direct decomposition: using "findTupel" recursive method that looks for the morphemes and returns the list of the retrieved results;
2. Left-to-right word truncation: if direct match is not found, the characters are being removed from left to right one by one;
3. Right-to-left word truncation: one-by-one character removal but in right-to-left direction.

One of the problems with the introduced method that has appeared on practice while testing was connected with the length of the words that are searched. When testing for the German language, the word "Laserdrucker" (laser printer) was decomposed as "Las", "Er", "Druck", "Er" ("read", "he", "print", "he"), [34]. To solve this problem, the authors suggested to set restriction on the word length, for it to be no less than four characters.

In general, approach showed rather high levels of performance: for the words consisting of 5 or more morphemes, the system that was implemented in Java, jWordSplitter, has demonstrated the speed of splitting 5.0000 morphemes/minute. For the compounds that consist of one or two morphemes the speed was 150.000 words/minute.

For the purposes of testing, 200 random compounds were chosen that included 456 morphemes. 89% of morphemes were recognized correctly, and 5% were also decomposed but not

completely. Therefore, the system could not recognize only 6% of the compounds.

3 Choice Of Methods

3.1 What features should ontology for scholarships have, and how to evaluate and modify existing ones?

The first thing that needs to be considered while evaluating ontology is why, in fact, it was created? What was its purpose? What main questions it should be able to give answers for? Perhaps, that is why expressiveness is probably the most important characteristics to analyze: after all, the core aim of any ontology is to be able to describe certain knowledge domain, and its efficiency depends on how well it is able to do it.

Comparing different kinds of evaluation, we can conclude that human-based evaluation is probably the most promising one, since it partly includes characteristics of other methods, but under strong supervision of a specialist. After all, thorough evaluation of expressiveness cannot really be done well enough without human component, and expressiveness analysis is something that is hard to automate, mostly because it requires certain level of creativity.

But human-only evaluation can also be a subject of a bias: without having an adequate support of data, some important points could be neglected. For example, for the case of scholarships, asking people to write possible queries to the system won't guarantee "Oceanography" or "Osteopathic Medicine" to be mentioned.

So what is the solution then? The combination of methods, with human expertise being in the center of it.

Different methods can be used for implementation of the human-based approach. To name some that would definitely serve as a benefit to our subject of study - scholarships, those are:

- Individual evaluation;
- Technical evaluation;
- Quantitative surveys;
- Qualitative interviews.

3.1.1 Individual Evaluation

For the case of scholarships individual evaluation of the ontology could include:

1. Choosing the ontology to evaluate;
2. Choosing the appropriate dataset;
3. Testing ontology by the means of a dataset: can ontology express the most important concepts that text describes? What could be possible queries of a user who would want to find the information in this text? How ontological concepts could be formulated/renamed so that it would be easier to express texts by them? What concepts ontology cannot describe?
4. The problems that ontology has should be identified;

5. Choice of the methods of how to solve the problems should be presented;
6. Ontology should be modified according to the chosen methods;
7. Conclusions should be made.

All the actions described above could be performed by one person, based on the overall level of expertise in the area - and extensive dataset (scholarship announcements). The drawback which can be seen in such approach lies exclusively in a fact that evaluation of the ontology is always of an iterative nature, which means that finding and fixing problems after evaluating ontology once with a certain dataset won't guarantee that, if choosing another dataset, new problems won't be discovered. But - due to the limited amount of time allocated for the Master Thesis writing, individual evaluation will be performed only once.

3.1.2 Technical Evaluation

Technical evaluation can be seen as an attempt to find certain web resources with an extensive mechanism of scholarships search. By investigating concepts that are identified as the most important ones, it will be easier to draw conclusions on what should be included/modified in the ontology.

This kind of evaluation can take quite a lot of time, when in its nature it doesn't differ from the Individual Evaluation significantly: it also requires an expert analysis of certain web sources. Ideally it could be nice to implement it too, but, again, due to the timing concerns, it could be referred to the Future Works.

3.1.3 Quantitative and Qualitative Evaluation

Both quantitative (in a form of surveys) and qualitative (in a form of one-on-one interviews) can be seen as a good way to get people's opinion on what concepts are believed to be most important for the ontology and what queries would people choose if searching for scholarships. Nevertheless, it might be more useful on the later stage of research, since on the early stages (and considering short period of time allocated for writing Master's Thesis) it is possible to perform primary evaluation by the means of extensive data analysis alone. Later on, though, it could be also beneficial to set up the survey with well-designed questions and ask people to spend their time on answering them.

3.1.4 Method Decision

Based on the information provided above, it was chosen to use human-based approach, Individual Evaluation in particular. There will be chosen 20 representative scholarship announcements to be used as dataset, and their contents will be analyzed. Can the ontology describe the most important concepts that scholarship announcements have? The list of issues in performance of the ontology will be made and the ways to solve them will be discussed, in accordance to which ontology will be modified afterwards. Modified ontology will be presented.

3.2 What kinds of information retrieval methods can be used for extracting knowledge in the scholarship content by the use of domain ontology?

3.2.1 Choice of the Environment

For the purposes of the given project, we can concentrate on one particular website in an attempt to determine how the necessary descriptional information could be extracted from it. The website that is to be analyzed - Scholarships For Development[1], the one example scholarships announcements had been taken from. The website provides information in quite structural way, which will make the data extraction process a bit easier. Example of a scholarship announcement provided on a website, is shown on the Fig. 10 and Fig. 11.

First of all, there can be distinguished three ways for information extraction, particularly for the given website - we can get the data either by analyzing the structure of sections (using HTML tags, scrapers) or by applying Information Retrieval techniques - or the combination of both methods (we are identifying section to apply Information Retrieval for).

3.2.2 Webpage Information Retrieval

In order to extract information from the webpage, different methods can be used which are commonly addressed by the term web scraping or web harvesting. The goal of web scraping, in contrast to the simple indexing and crawling, is to transform and organize unstructured data on the Web, usually for its further storage in the database.

Different techniques can be used for web scraping, and many vendors offer services either in a form of online or offline commercial (sometimes also free of charge) software that allows to scrape information from the website in a specific for every webpage way. Such scrapers show the best results when dealing with repeatable information such as tables, and do not demonstrate good performance when dealing with a regular less structured text. Examples of such scrapers are Mozenda [2], Convextra [3], etc.

XPath, or XML Path Language, is a language for selecting nodes inside XML documents that can be used for extracting the content from within specific tags inside HTML files. Specification for the first version of the language, xPath 1.0, [35] was introduced in 1999, and in 2007 the second version, xPath 2.0, was released, [36]. According to the description of the xPath capabilities, it will be able to help to extract necessary information from the webpage by manually setting the path to the desired element in XHTML.

3.2.3 Analysis of the Environment

Our first task is to determine what kinds of information we need to describe using Scholarship Ontology, and what parts of it could be extracted either by direct web-scraping or/and by applying IR-techniques. Properties of the ontology that can be directly extracted based on their structural location on the page, are framed in orange on the Fig. 10 and Fig. 11, and are also

[1] http://www.scholars4dev.com/
[2] http://www.mozenda.com/
[3] http://convextra.com/

Figure 10: Example of the scholarship announcement, p.1

Figure 11: Example of the scholarship announcement, p.2

listed below:

- hasSponsor;
- hasScholarshipName;
- hasDeadlinee;
- hasOfficialURL;
- scholarshipForLocation;
- ScholarshipProvidesDegree;
- isOfferedForInstituition

Hence, all properties described above can be filled in with information from the webpage, extracted using, for example, xPath. But what to do with the properties which are not so trivial to extract? For example, *Scholarship Value/Inclusions* section on a website contains the information that can fill in properties hasTypeOfScholarship, hasFrequencyOfScholarship, hasAmountOf, hasCurrency, costsCover; *Target Group* section includes information that can be used to describe properties requiresCountryOfOrigin, hasGroupOfLocationsName, hasCountriesMembers; and *Eligibility* section (if it exists) - requiresLevelOfEducation, scholarshipForTuitionLanguage, requiresAge. *Number of awards* section could get the information hasNumberOfScholarships (again, if it exists), whereas *Field(s) of study* contains information to fill in the value of the property IsGivenForDiscipline.

Obtaining the data to fill in the properties' values from the plain text is rather complex task, which can be divided in two steps:

1. Identify text segments that contain information to fill in the properties' values;
2. Match extracted text to the ontological terms (classes, subclasses, instances) - to retrieve results.

The task of the given project is to look in depth into certain problem, rather than trying to discuss many points superficially. Therefore, *Subject of the study* (discipline) has been chosen as the one to study in more detail, as it is one of the most important aspects of scholarships querying process (usually people are searching for scholarships in certain country for the certain field of study, other information can be not so essential to the possible applicant).

3.2.4 Choice of IR-Methods

Our goal is to "scrape" the "free text" from the "Field of study" section, and then extract the names of the disciplines from it. XPath can handle the scraping process, since the website "Scholarships for development" has rather good structure of sections, therefore, it is possible to get the contents of elements by their tags. When it comes to the further discipline retrieval methods, different approaches must be considered.

Classical Boolean Search, as trivial as it may sound, really meets the requirements of the IR task, since we can perform "full phrase" search on all levels of the ontology, trying to match ontological concepts (superclasses, subclasses and instances) to the words in the text.

Probably, just Boolean method alone will not provide us results that will be good enough, that

is why preprocessing will be an important part of the Discipline Retrieval Process - that is to be discussed further.

When it comes to the Vector Search, it cannot be helpful enough in our case, since it is intended to find in which of the searched documents there will be found information that contains most occurrences of searched terms - and then it ranks them by relevance. Our task, on the other hand, is to find out whether certain query can be found in one document, and if not - we are no longer interested. Vector Search can be a good solution when we are dealing with large amounts of free text with full sentences, when there is much of a context. But with the "Field of Study" area extraction, full sentences are quite rare, and even if they occur, they don't have much of semantical cohesion: they often are just sets of lists without much of sentential structure.

Some other well-known methods, for example, EditDistance, is concerned about the similarity between the documents, so, again, it's not much of a helper in our task of matching the queries.

Specific projects, like the one mentioned by P. Alexopoulos et al., "Exploiting Ontological Relations for Automatic Semantic Tag Recommendation" [30], are also concerned more about getting the major concepts out of the context, rather than matching the set of predefined ones - that is why the authors' experience won't be useful to our research problem.

Judging from all of the above, it was chosen to use Boolean Search for automatic extraction of the discipline names - with the further evaluation of its results.

3.3 How preprocessing can improve the results of the information extraction in the scholarship-related domain?

There are different kinds of preprocessing methods that can help to extract information from the text. The ones appropriate for the given context are:

- Changing the word order inside the phrase;
- Breaking query into pieces;
- The use of stop-lists;
- Stemming/lemmatization;
- Decomposition of compounds;
- Query extension by the means of synonyms;
- Other preprocessing methods that can be proposed in a specific context.

Changing the word order inside the query can be a good practice, since sometimes specific disciplines are mentioned together, like "Arts and Humanities". By changing the word order inside the phrase-based query, we will be able to find "Humanities and Arts" in the text as well.

Breaking query into pieces can be rather beneficial, if we define specific "splitting rules", for example, regarding "and", slashes "/" and commas "," as separators. In this case we will be dealing with two queries "Social" and "Behavioral Sciences" instead of one "Social and Behavioral Sciences" which is, quite understandable, is rather hard to find. Stop-lists can help in this process even more, since we can neglect the word "Sciences" in the query, since "Behaviorial" will be

easier to find than "Behavioral Sciences".

Stemming and lemmatization can be seen as a good consistent approach to the ontological disciplines' names preprocessing, since same concepts can have a bit different names. Examples could be: "Agriculture" and "agricultural", "Resource" and "Resources", etc. So, performing stemming could really improve discipline retrieval rates.

Synonyms could be also an interesting point of research, but for the case of disciplines they can be neglected for two reasons:

- Discipline names do not represent synonymically rich domain of knowledge, and after closer consideration it was noticed that each concept has relatively unique meaning;
- The taxonomy of disciplines consists of 472 distinct discipline names, so it is safe to say that the majority of distinct discipline names are covered in it.

Compound names of the disciplines are not a usual case for English, that is why compound-based IR was not performed in the current project. However, if we had ontology for the disciplines in, for example, German language, decomposition of compounds would be highly appropriate and needed, since in German the names of the majority of disciplines are formed by combining several words together, e.g., "Medienwirtschaft" (Media Economics).

4 Results

4.1 What features should ontology for scholarships have, and how to evaluate and modify existing ones?

4.1.1 The Choice Of Ontology

At the time of writing current MSc Thesis there was found only one ontology that satisfies core requirements of expressiveness of the ontology in the field of scholarships - Scholarship Ontology, [13]. For the purpose of search for scholarships, following engines were used:

- Swoogle[1] [37]. This semantic search engine is outdated, as it was maintained during 2004-2007 as it is stated on the frontpage of the project. Nevertheless, it still returns certain results, Fig. 12. Although, when trying to open the results provided by the engine, it turns out that the majority of links are not working anymore, probably they are still kept in the "memory" of Swoogle engine. From the obtained results one can also see that retrieved ontologies are not devoted to scholarships exclusively (even if "Scholarship" word is used and described there);
- Schemapedia[2] - didn't bring any results, currently only beta-version is available;
- Falcons[3] - another semantic search engine, established in 2011. Results are retrieved not in the form of links, but in the form of graphs, Fig. 13. Named engine retrieved only two results. After closer examination of the classes in retrieved ontologies, it was found that they contain only part of the information about scholarship knowledge domain, therefore, retrieved ontologies cannot be used for testing the hypothesis of the MSc Thesis;

Other search engines that were used are Watson[4] [38] and The Semantic Web Sindice[5]. No relevant results have been found by their use.

A lot of other semantic search engines that were suggested by different users and specialists on forums and specific websites, also had outdated links, and therefore could not be researched.

Therefore, ontology that is proposed for the analysis in the given research, is called "Scholarship Ontology", and it has been developed in the course of Advanced project Work in March-May 2013, [13]. Ontology covers the domain of scholarships by defining both essential concepts and relationships between them. Scholarship Ontology, as any other ontology, is based on the taxonomy, it has 78 classes (including "Thing"), 33 properties (23 Object Properties and 9 Data Properties) and many individuals (instances) of the classes. Visual representation of the ontology's taxonomy is shown on the Fig. 14.

So, we have an ontology for testing MSc Thesis hypothesis upon, now it's time to choose methodology we will be using to evaluate it.

[1] Swoogle - semantic search engine: http://swoogle.umbc.edu/.
[2] Rdf schema compendium: http://schemapedia.com/
[3] Falcons - semantic search engine. http://ws.nju.edu.cn/falcons/ontologysearch/ index.jsp
[4] Watson search engine: http://watson.kmi.open.ac.uk/WatsonWUI/
[5] Sindice search engine: http://sindice.com/

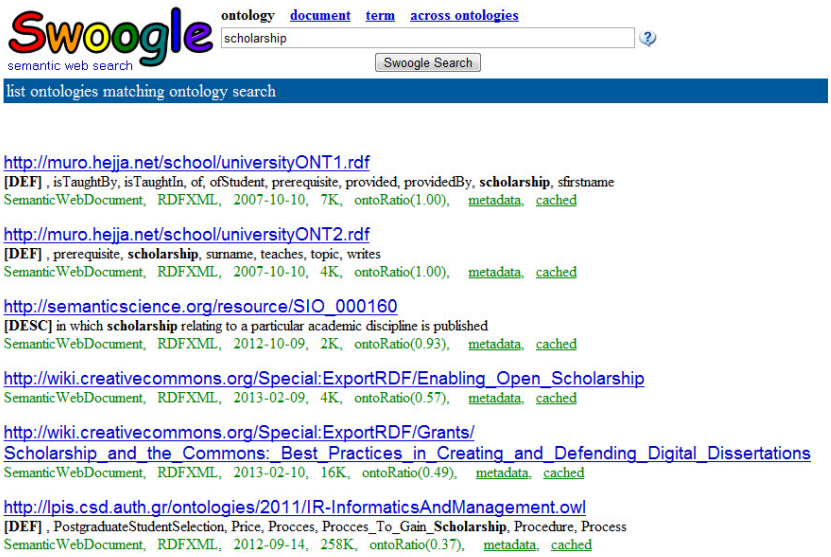


Figure 12: Semantic Scholarship Search Results in Swoogle

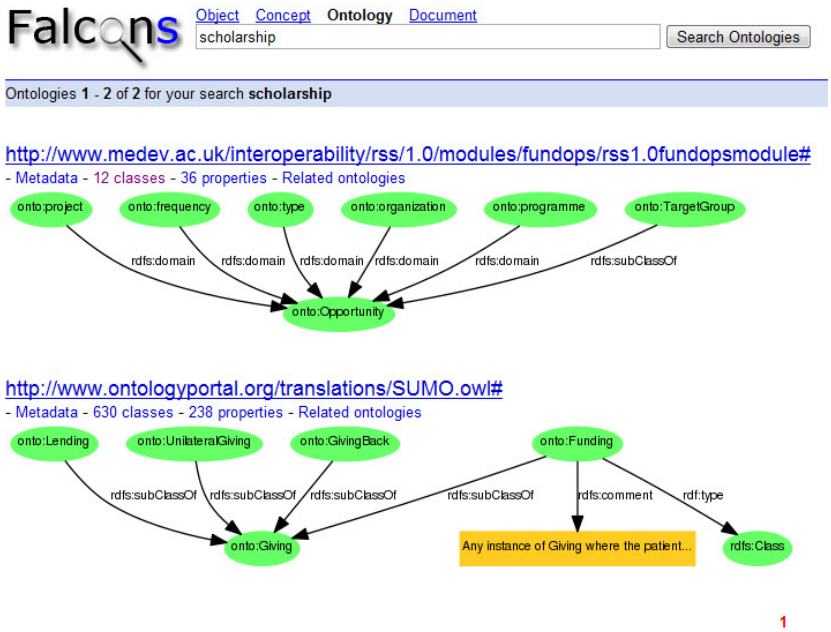


Figure 13: Semantic Scholarship Search Results in Falcons

Figure 14: Visualization of Initial Scholarship Ontology Model in OwlViz

4.1.2 Evaluation Of Scholarship Ontology

Individual evaluation has been performed in the following steps:

1. 20 scholarship announcements have been found and analysed: can the Scholarship ontology express their most important messages? What about the possible queries users can have, is it possible to both express them via ontology and to find/match them to the certain scholarship announcement? Graphical representation of the scholarship announcement data that can be encoded via ontology will be presented;
2. Difficulties that Scholarship ontology and possible search queries have with expression of semantic data will be documented;
3. The ways to solve discovered problems will be discussed;
4. Ontology will be modified;
5. Results will be documented and overall conclusions about individual evaluation of Scholarship Ontology will be drawn.

There was found rather decent website that provides up-to-date information about scholarship opportunities, "Scholarships for Development", [39]. Information about different kinds of scholarships can be found easily on the website, therefore twenty announcements about scholarship were taken from it.

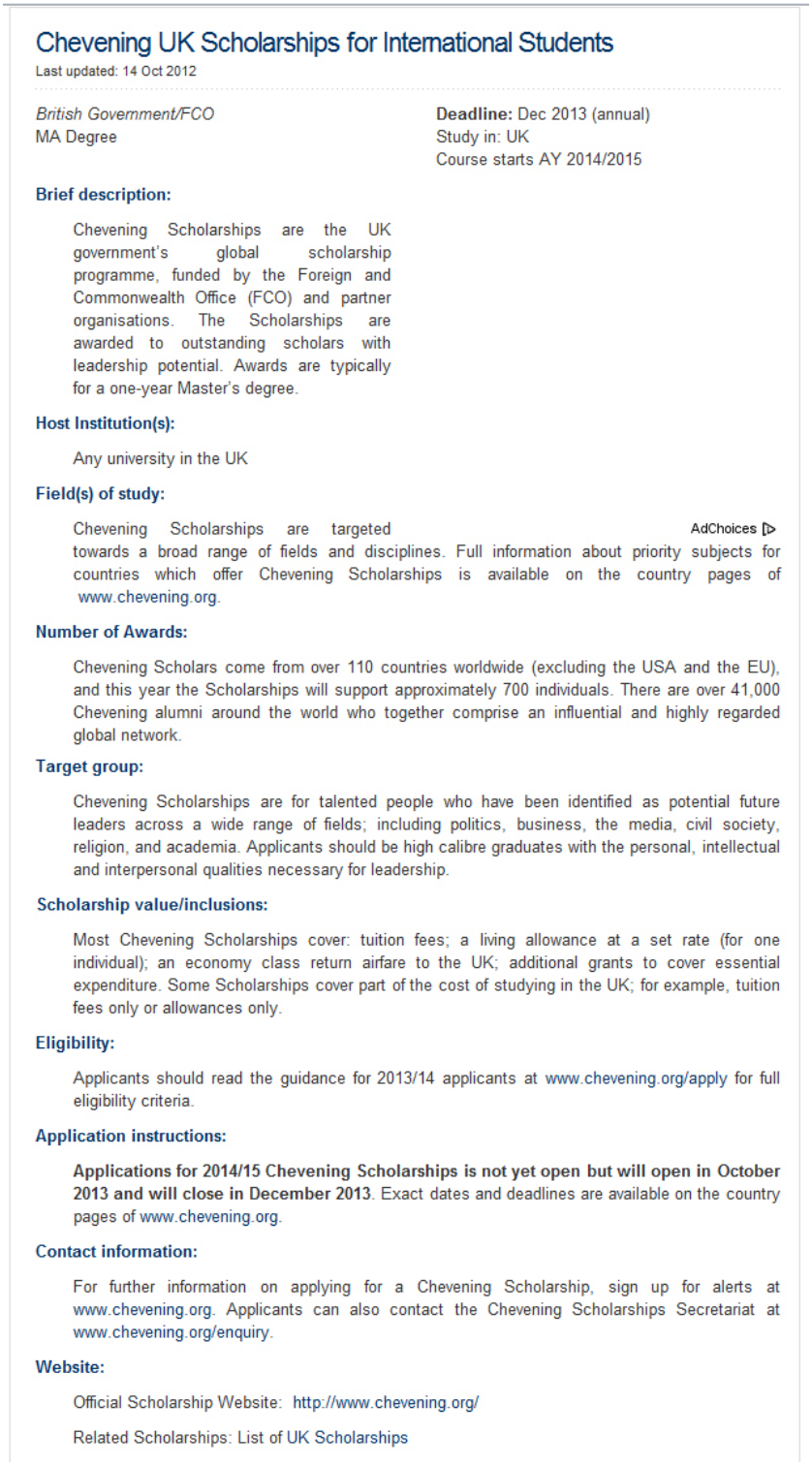
Dataset has been chosen in a way that scholarships given by different countries were represented in it, as one country usually has more or less the same requirements for the perspective applicants. Therefore, announcements to study in UK, US, Germany, Norway, Australia, Netherlands, etc were selected. Some of them provided funding for citizens of particular geographical or economical area, some - only under conditions of certain GPA scores (US), some were offered for non-degree short courses, whereas even some - for Online degree. The idea was to get wider range of distinguished funding options in order to determine whether it's possible to describe them by the use of Scholarship Ontology, and if not, what alterations need to be implemented to the ontology to accomplish that. The full list of selected scholarships announcements could be found in Appendices, A.

Analysis from the scholarship announcement point of view

Example of scholarship announcement is presented on the Fig. 15.

Let's see what information we can extract from this announcement by the use of ontology. In order to do this, for the given announcement N3-notations were generated, converted to RDF, and then visualised using one of the available tools, [40], [41]. Visual representation of the information obtained from the announcement is shown on the Fig. 16.

The figure shows that some essential information could not be represented by the current ontology: for example, we need a mechanism to assign certain level of the degree program to the scholarship announcement and not to the educational institution (it is obvious that educational institutions provide numerous study programmes, but are they covered by the certain scholarship? We also need a mechanism to assign Tuition language to scholarship and not to the study programme (subject of studies). Other problems faced with when describing scholarship



Chevening UK Scholarships for International Students

Last updated: 14 Oct 2012

British Government/FCO
MA Degree

Deadline: Dec 2013 (annual)
Study in: UK
Course starts AY 2014/2015

Brief description:

Chevening Scholarships are the UK government's global scholarship programme, funded by the Foreign and Commonwealth Office (FCO) and partner organisations. The Scholarships are awarded to outstanding scholars with leadership potential. Awards are typically for a one-year Master's degree.

Host Institution(s):

Any university in the UK

Field(s) of study:

Chevening Scholarships are targeted towards a broad range of fields and disciplines. Full information about priority subjects for countries which offer Chevening Scholarships is available on the country pages of www.chevening.org.

AdChoices

Number of Awards:

Chevening Scholars come from over 110 countries worldwide (excluding the USA and the EU), and this year the Scholarships will support approximately 700 individuals. There are over 41,000 Chevening alumni around the world who together comprise an influential and highly regarded global network.

Target group:

Chevening Scholarships are for talented people who have been identified as potential future leaders across a wide range of fields; including politics, business, the media, civil society, religion, and academia. Applicants should be high calibre graduates with the personal, intellectual and interpersonal qualities necessary for leadership.

Scholarship value/inclusions:

Most Chevening Scholarships cover: tuition fees; a living allowance at a set rate (for one individual); an economy class return airfare to the UK; additional grants to cover essential expenditure. Some Scholarships cover part of the cost of studying in the UK; for example, tuition fees only or allowances only.

Eligibility:

Applicants should read the guidance for 2013/14 applicants at www.chevening.org/apply for full eligibility criteria.

Application instructions:

Applications for 2014/15 Chevening Scholarships is not yet open but will open in October 2013 and will close in December 2013. Exact dates and deadlines are available on the country pages of www.chevening.org.

Contact information:

For further information on applying for a Chevening Scholarship, sign up for alerts at www.chevening.org. Applicants can also contact the Chevening Scholarships Secretariat at www.chevening.org/enquiry.

Website:

Official Scholarship Website: http://www.chevening.org/

Related Scholarships: List of UK Scholarships

Figure 15: Example of a Scholarship Announcement for Description

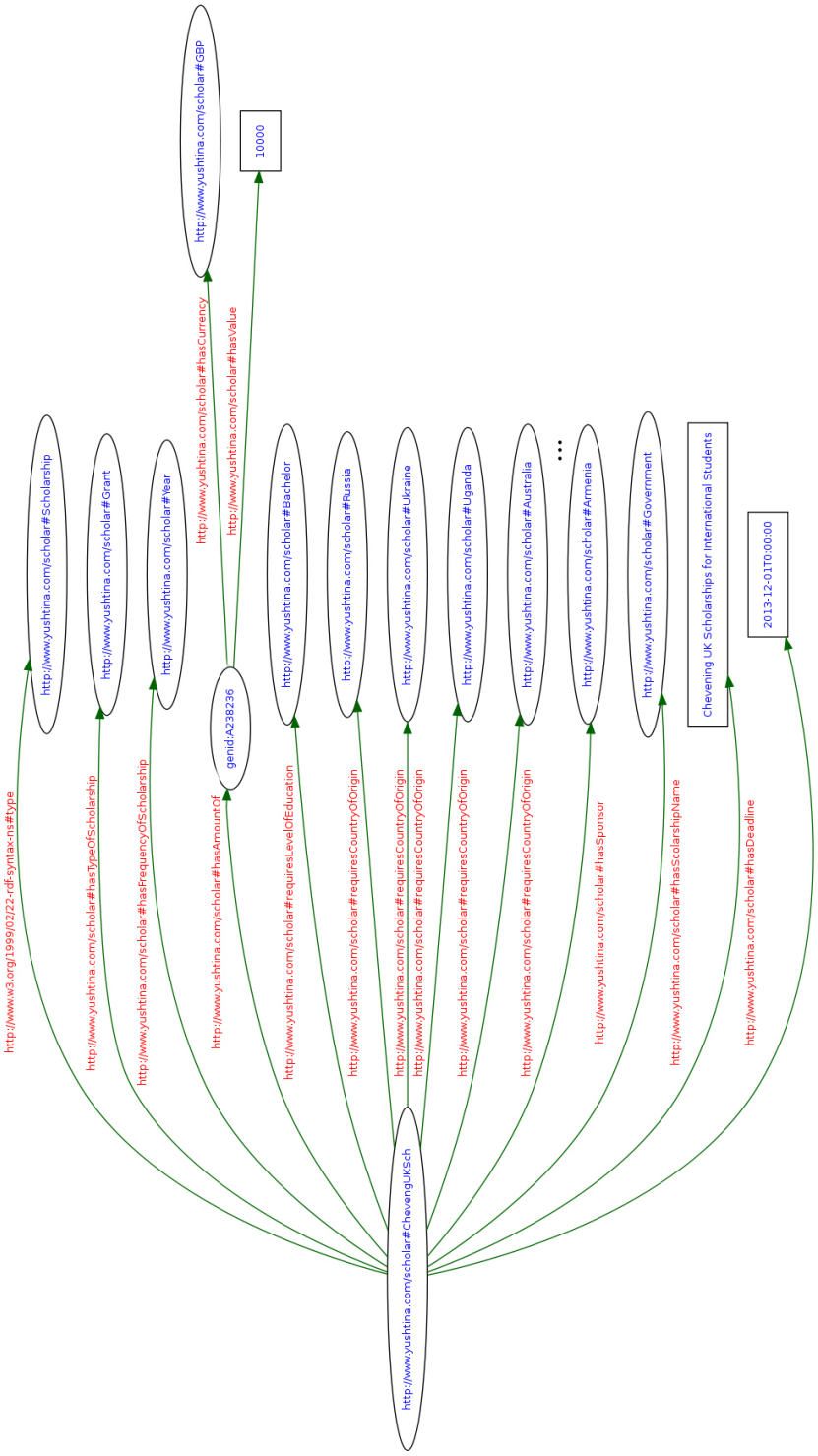


Figure 16: Chevening Scholarship Description

announcement appear because of the lack of initial concepts defined by ontology.

Hence, following problems of the ontology were identified when performing individual evaluation of given and other scholarship offers:

1. Problems related to the lack of introduced concepts:
   - Currently ontology does not present the opportunity to show what kind of costs does the scholarship cover (providing that this information is available). Examples can be costs for the accommodation, travel costs, insurance, tuition fees, living costs, books, visa, etc;
   - For every scholarship announcement there must be stated its official web address where users can obtain more information about the program;
   - Number of offered scholarships should also be stated.
2. Problems related to the lack of introduced relationships:
   - Scholarship announcement must be directly related to the Degree Programme;
   - Scholarship announcement must be directly related to the Tuition Language;
   - Scholarship announcement must be linked to the location of the study programmes for which those scholarships are given (for now Location is linked only to the educational institution).
3. Other concerns:
   - There must be found a solution for selecting "all countries BUT", "EU" and other possible groupings of countries (it is also possible that the query itself proposed by the user will sound like "select scholarships for EU-origins" or "select scholarships for studying in EU". Therefore, a mechanism for encoding this kind of queries should be thought through both from the point of scholarship announcement (available information) and possible query. Other possible queries can concern "Eastern Europe", "Africa", etc. Similar problem appears when scholarship can be given for "all courses BUT" (for example, all courses in all American universities except from medicine (Fulbright scholarship);
   - Some scholarship announcements have information about the type of Degree students will get, either Bachelor/Master of Arts, Science, Business Administration, etc. Should these options be able to be described by the ontology? The choice must be well argued;
   - Some scholarship announcements include information about the required language tests (TOEFL, ILES, TestDaf, etc) and the minimum scores one is expected to get to be eligible to apply for scholarship - should this information be encoded into the ontology also? Same applies to the necessary scores in GPA and other metrics (GPA for the case of studying in American universities);
   - Some scholarships are offered for non-degree programs (short courses), so there must be found a mechanism to describe that using ontology;
   - There are also universities that provide Online Degree. There must be found a way to describe these kinds of scholarships;

- Other requirements of some of the scholarship sponsors are that applicants should have a certain work experience, want to pursue certain degree program in the field of their previous studies (or related fields), some scholarships state that recipient should return back to their countries after studies (or there are some requirements for that, e.g., in Quota Scholarship Programme for studying in Norway, scholarship is given in a form of both grant (30 percent) and loan (70 percent), which means that if recipient will stay in Norway after graduation, he or she will need to return that part of scholarship back to the Norwegian government. However, if they choose to go back to their countries, the loan is being waived). So the question is, do we need to be able to express this information?
- A lot of other eligibility restrictions could be found in every particular scholarship offer, nevertheless, possible applicant probably wouldn't start their query by searching for them - which means that they could be neglected during ontology revision.

Analysis from user query point of view

According to the found scholarship announcements, following queries could be expected to be seen:

- Scholarships to study in EU for EU-origins;
- Scholarships for Masters in Computer Science for developing countries;
- Scholarships for studying MBA;
- Scholarships for studying Nuclear Physics Bachelor in Germany in English language;
- Scholarship to study Bachelor in Czech Republic in English;
- Scholarship to study in Norway for non-EU citizens;
- Scholarships for studying in UK;
- Scholarships for Master Programme in International Business in English in Europe;
- Full-covered Bachelor level scholarships for non-EU citizens;
- Scholarship for chemistry studies for Master level in English.

Statements that ontology currently can't express:

- Again, computer needs to understand what "Europe", "EU", "European Union" etc means. All those terms need to be enabled to be described by the ontology, and also there must be found a way to describe complementary concepts, e.g., "non-EU", "excluding US", etc.,
- We need to find a way to encode "full-covered" type of scholarship into ontology.

The majority of discovered expressiveness issues can be solved rather straight forward by adding extra concepts and relationships to the ontology, it doesn't require any planning and doesn't involve any dilemmas. But some of expressiveness aspects need to be handled differently, and specific approach has to be invented for them:

1. Type of degree programs (BA, BBA, BSc, MA, MBA, MSc, etc). Do we need to enable their description in the ontology?

2. Extended location description: "European Union", "Europe", "all countries BUT", etc - how to express all that using ontology?

Type of Degree Programs

Many scholarship announcements contain information about the type of the Degree they provide, not just Bachelor's or Master's, but also their field (Arts, Science, Business Administration, etc). Currently ontology doesn't provide the tools to deal with this kind of information, therefore, it doesn't distinguish between different types of Bachelor's and Master's. The question is, do we need to add this functionality to the ontology?

From one point of view, this information helps to describe educational program better, and it might be tempting to add ability to express this information to the ontology. On the other hand, though, what are the chances that actual user will query for scholarships in MSc or MA? Usually person is interested in a specific educational program (Chemistry, Physics, Maths or Marketing, etc), and does not really care which field the chosen speciality refers to. It may be nice to know in the future, but in the process of querying (which was and is the core motivation for ontology development and modification in the first place), it is not that crucial for the user - at least on the initial evaluation step, when the time for the research is limited. However, such functionality can be added in the future and can be considered referring to the Future Works.

Same reasoning can be applied for the discussion on whether to include language tests requirements and GPA scores. Since requirements for the scholarships very much differ throughout announcements, it is not that important to include that many details into their descriptions - at least on the initial step of writing Master's Thesis when the time is limited.

Extended Location Querying Concerns

The question whether to include information about countries' "groupings" like European Union, etc is not a question - from the point of scholarship announcement, it's possible, of course, just to encode countries included in a certain group manually. But when we are talking about proper evaluation of ontology, we need to perform it from the end-user point of view. So let's say, if user enters a query of a kind "scholarships to study in EU", what results will he or she get if there is no statement about what is EU either in ontology or in the particular RDF-description of a particular scholarship? There will be no results fetched - which can be crucial, as there may be found a lot of countries from European Union that provide scholarship offers to the perspective students, but they won't be retrieved because there was not stated anywhere what is EU. Therefore, certain solution to handle such situations should be found.

For the purposes of encoding geopolitical information specific ontology can be imported and reused. Such ontology has been found, Geopolitical Ontology "geo", [42]. But it will provide only "straight forward" solutions, what if we search for "non-EU" location, for example, how is it possible to match such user's query?

Two different options can be exploited here. We can provide such functionality on the level of ontology or on the level of particular scholarship announcement. For better understanding, it will be convenient to show how ontology and actual specific scholarship description connect with each other, Fig. 17.

The first method to enable usage of the statements of a kind "non-EU", etc (providing we have

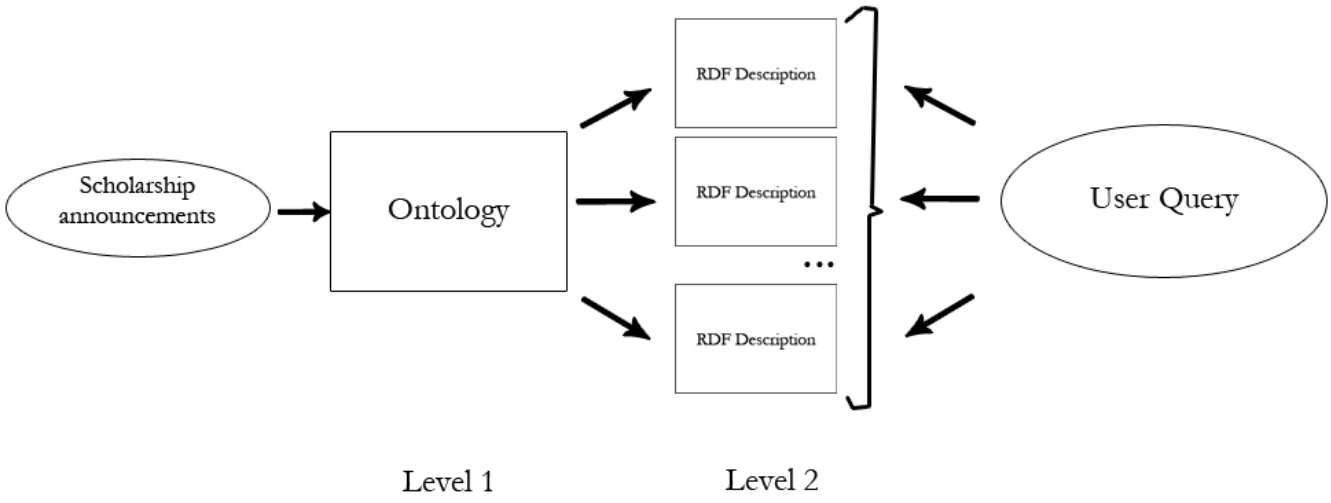


Figure 17: Process of forming scholarship announcement descriptions and querying

"EU" previously defined) is to change ontology itself (first level of scheme), creating new classes which are complementary to existing ones. This can cause several concerns though: first, to make complementary class that doesn't include members of a certain class can be rather tricky and complicated. But even if it's achieved, how can we predict what kinds of geopolitical locations will the user actually query? There can be so many groups of countries, and it is impossible to encode all of them into the ontology.

Another option is to make changes on Level 2, when actual scholarship announcement descriptions are being formed, simply by calling annotated group of countries by certain name. To do that, also several changes on the level of ontology (level 1) should take place: we need to develop new subclass of Location that will allow to describe group of countries and to give this group a name.

This approach is really much simpler than the first that was proposed before: we don't need to predict anything, creating complicated relationships between concepts and making the whole ontology much more complex than it actually needs to be. This approach is easier, and it will ensure that all groups of countries are encoded the way they need to be encoded.

Hence, second approach has been chosen to be used on practice.

4.1.3 Modification of the Ontology

According to the results of individual evaluation, several modifications need to take place. Namely:

1. New class of the ontology Expenses was introduced with several instances: TuitionFee, Accommodation, TravelCosts, Books, Insurance, FullCover. Object Property costsCover was created with Domain: Amount of Scholarship; Range: Expenses (not functional). Another Object Property isCoveredByScholarship is created, which is Inverse to the costsCover property.;
2. New Data Property hasOfficialURL was created, Domain is Scholarship, Range: datatype:string;

3. New Data Property hasNumberOfScholarships was created, Domain is Scholarship, Range: datatype:integer;
4. New Object Property scholarshipForTuitionLanguage was introduced, with Domain Scholarship and Range TuitionLanguage. Another object property was created, isOfferedForScholarship, which has been made inverse to scholarshipForTuitionLanguage;
5. New Object Property scholarshipProvidesDegree was introduced, with Domain Scholarship and Range HighEducation. Another object property was created, degreeIsProvidedByScholarship, which has been made inverse to scholarshipProvidesDegree;
6. New Object Property scholarshipForLocation was introduced, with Domain Scholarship and Range Location. Another object property was created, hereAreScholarships, which has been made inverse to scholarshipForLocation;
7. To the HighEducation class another subclass OtherEducation was added with its members being "ShortCourses" and "OnlineDegree";
8. New class AdditionalRequirements was introduced. Therefore, new Object Properties requiresAdditionalRequirements (Domain: Scholarship, Range: AdditionalRequirements) and isAdditionallyRequiredByScholarship (as inverse to the previous one) were created.

Geopolitical ontology has been imported to the current Scholarship Ontology. It has different economic and geographical regions encoded in it, which means that it enables to express such terms as "Southern Africa", "European Union", etc.

Geopolitical ontology consists of superclass "Area" with one of the subclasses "Group" with its subclasses economic region, geographical region, organization and special group. Currently given ontology is considered to be the most comprehensive in expressing political and geographical groups of countries. Therefore, class "Area" of the geopolitical ontology has been set as a subclass of class "Location" of the Scholarship Ontology.

Also, another modification should take place: we had Object Property "requiresCountryOfOrigin", which pointed on the specific country possible applicants could apply from. But now, since we have expanded our Location class by adding to it subclass "area" from the Geopolitical ontology, it will make sense to change the Range for the Object Property "requiresCountryOfOrigin" from Location-Country to just "Location", enabling the user to describe countries whether one by one as before, or by pointing on the groups of the countries from the geopolitical ontology. Hence, the range of object property "requiresCountryOfOrigin" was changed to "Location".

Additionally, new subclass of Location was introduced: "GroupOfLocations". Data Property "hasGroupOfLocationsName" with type "string" is added, and Object Property "hasCountriesMembers" is created with the Domain "GroupOfLocations" and Range "Location" itself (because countries can be selected there).

4.1.4 Modified Ontology

After modification of the ontology scholarship announcement presented on the Fig. 15 can be visualized as follows: Fig. 18.

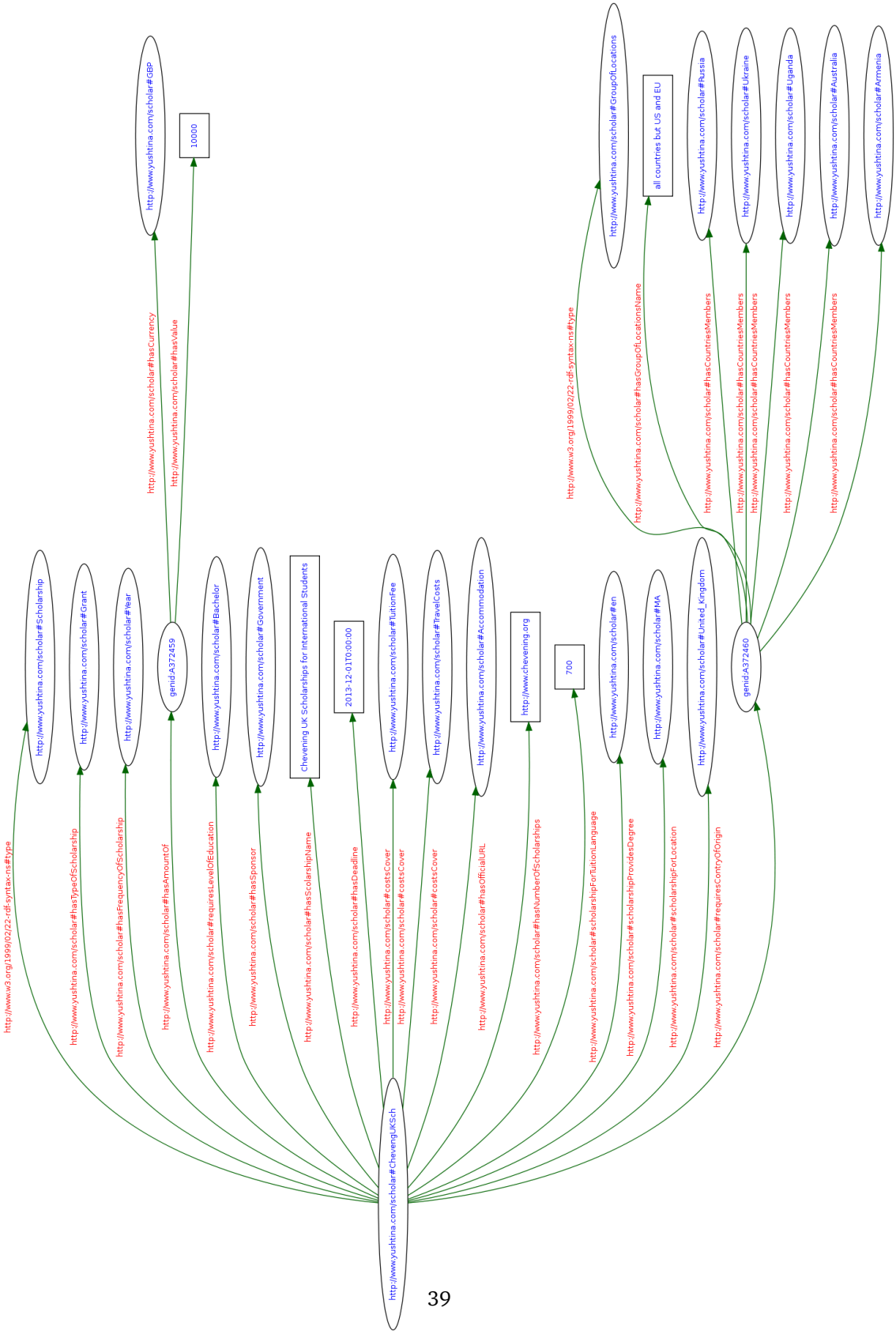


Figure 18: Ontology Expression of Chevening Scholarship (modified)

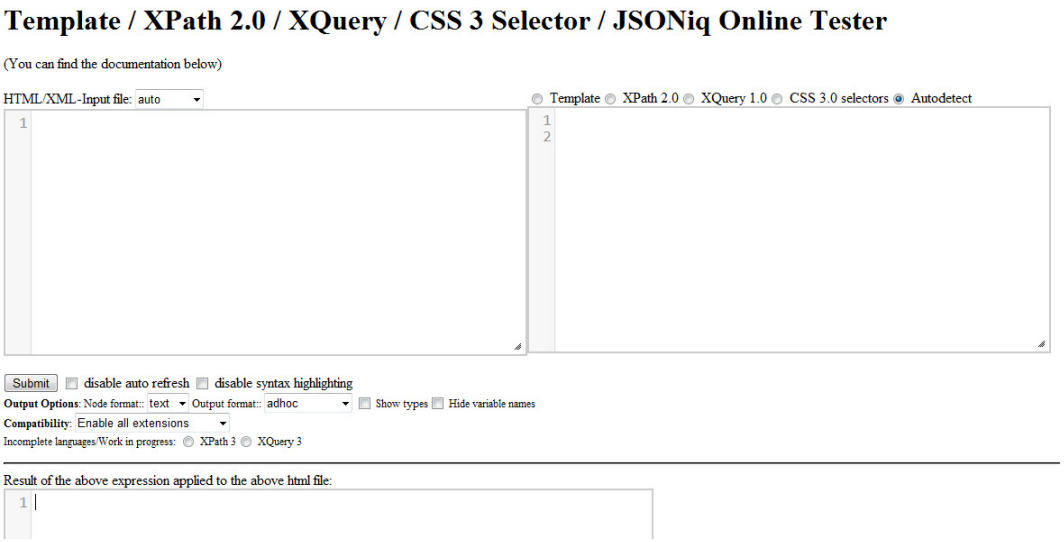


Figure 19: xPath Scraping Graphical Interface

4.2 What kinds of information retrieval methods can be used for extracting knowledge in the scholarship content by the use of domain ontology?

In order to test performance of the IR methods, 20 scholarship announcements that were used for the evaluation of ontology were considered. The section "Field of Study" has been analyzed.

4.2.1 Field Of Study Content Scraping

The content of the Field of Study section can be extracted by the use of xPath Query Language. For the testing purposes "Template / XPath 2.0 / XQuery / CSS 3 Selector / JSONiq Online Tester" [6] project was chosen, since it provides good functionality, allowing to enter HTML code in one of the windows, xPath code in another one, and immediately see the result in the third one - without even refreshing the page, Figure 19.

In order to be able to write an xPath for the content of the "Field of study" section, the structure of the scholarship announcements was thoroughly analyzed. Example of the scholarship announcement with quite a lot of data inside "Field of study" section is shown on the Figure 20 (part one) and Figure 21 (part two).

After analyzing the page, specific features of "Field of study"'s content were identified:

- The target content is located after the paragraph <p> "Field(s) of study:" which is written in bold characters;
- The target content ends when new paragraph <p> "Target Group" starts, written in bold again.

According to the given announcement, we could extract target information by writing the

[6] http://videlibri.sourceforge.net/cgi-bin/xidelcgi

Greek Government Scholarships for Foreign Students

Last updated: 18 Jun 2013

Greek Government
Masters/PhD Degree

Deadline: 26 July 2013
Study in: Greece
Course starts 2014

Brief description:

The Greek State Scholarships Foundation (I.K.Y.) announces that it will offer up to thirty (30) scholarships for nationals from selected countries who wants to purse further education/postgraduate studies/ postdoctoral research in Greece beginning in the academic year 2013-2014.

Host Institution(s):

Universities in Greece

Field(s) of study:

a. Postgraduate studies in combination with Modern Greek Language courses in a school of Modern Greek Language, at a state Greek University only during the first year of the scholarship.

i) Master's Degree (one (1) year up to two (2) years).
ii) Doctorate (PhD) (one (1) year up to three (3) years).

b. Postdoctoral research (six (6) months up to one (1) year) in combination with optional Modern Greek Language courses in a School of Modern Greek Language, at a state Greek University.

c. Further education in the following subject areas: Greek Language, Literature, Philosophy, History and Art aimed for professors of Greek studies at Universities abroad (six (6) months up to one (1) year).

d. Specialisation in Fine Arts (attendance of seminars) – one (1) year in combination with optional Modern Greek Language courses in a School of Modern Greek Language, at a state Greek University.

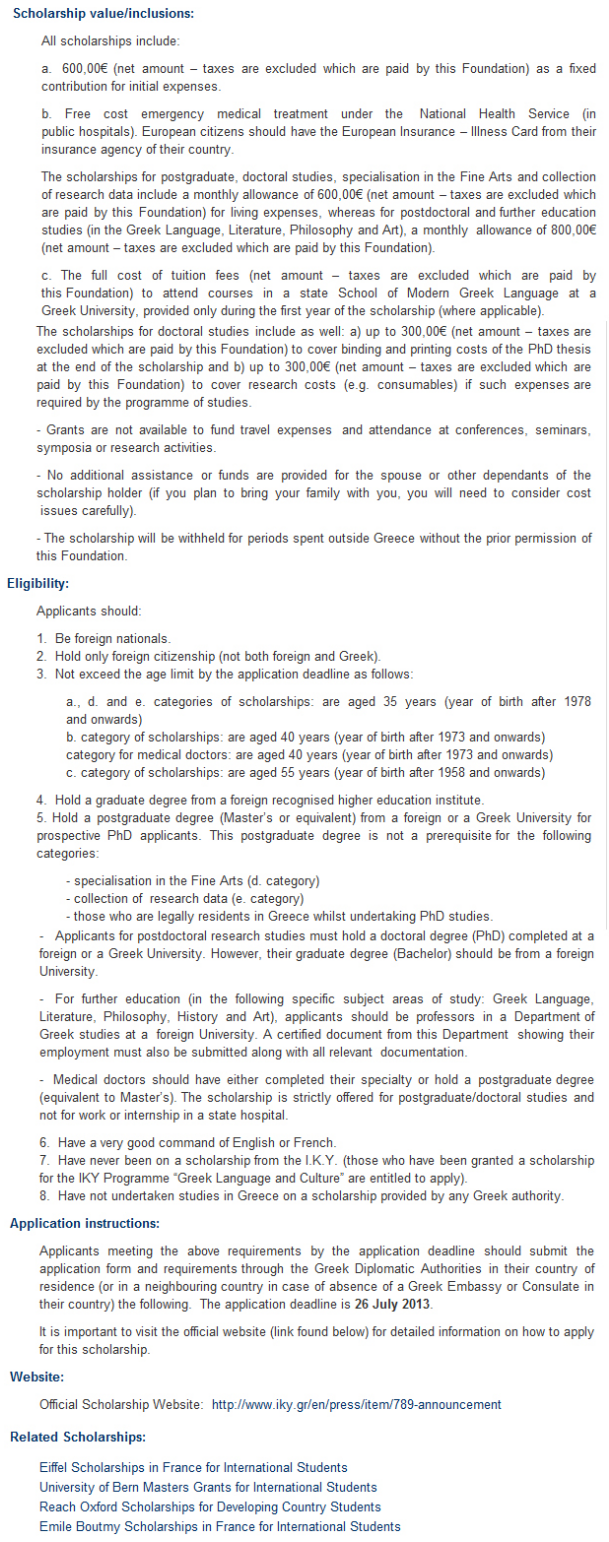
e. Collection of research data for applicants who are conducting PhD studies in their country, (one (1) year).

Target Group:

Fifteen (15) scholarships to nationals (foreigners or of Greek origin) of Balkan and Eastern European countries (non-member states of the European Union), Asia, Africa and Latin America and

Fifteen (15) scholarships to nationals (foreigners or/of Greek origin) of the European Union Member States, Iceland, Norway, Switzerland, U.S.A., Canada, Japan and Oceania, for studies.

Figure 20: Snippet of a scholarship announcement (part one)



Scholarship value/inclusions:

All scholarships include:

a. 600,00€ (net amount – taxes are excluded which are paid by this Foundation) as a fixed contribution for initial expenses.

b. Free cost emergency medical treatment under the National Health Service (in public hospitals). European citizens should have the European Insurance – Illness Card from their insurance agency of their country.

The scholarships for postgraduate, doctoral studies, specialisation in the Fine Arts and collection of research data include a monthly allowance of 600,00€ (net amount – taxes are excluded which are paid by this Foundation) for living expenses, whereas for postdoctoral and further education studies (in the Greek Language, Literature, Philosophy and Art), a monthly allowance of 800,00€ (net amount – taxes are excluded which are paid by this Foundation).

c. The full cost of tuition fees (net amount – taxes are excluded which are paid by this Foundation) to attend courses in a state School of Modern Greek Language at a Greek University, provided only during the first year of the scholarship (where applicable).

The scholarships for doctoral studies include as well: a) up to 300,00€ (net amount – taxes are excluded which are paid by this Foundation) to cover binding and printing costs of the PhD thesis at the end of the scholarship and b) up to 300,00€ (net amount – taxes are excluded which are paid by this Foundation) to cover research costs (e.g. consumables) if such expenses are required by the programme of studies.

- Grants are not available to fund travel expenses and attendance at conferences, seminars, symposia or research activities.

- No additional assistance or funds are provided for the spouse or other dependants of the scholarship holder (if you plan to bring your family with you, you will need to consider cost issues carefully).

- The scholarship will be withheld for periods spent outside Greece without the prior permission of this Foundation.

Eligibility:

Applicants should:

1. Be foreign nationals.
2. Hold only foreign citizenship (not both foreign and Greek).
3. Not exceed the age limit by the application deadline as follows:

   a., d. and e. categories of scholarships: are aged 35 years (year of birth after 1978 and onwards)
   b. category of scholarships: are aged 40 years (year of birth after 1973 and onwards)
   category for medical doctors: are aged 40 years (year of birth after 1973 and onwards)
   c. category of scholarships: are aged 55 years (year of birth after 1958 and onwards)

4. Hold a graduate degree from a foreign recognised higher education institute.
5. Hold a postgraduate degree (Master's or equivalent) from a foreign or a Greek University for prospective PhD applicants. This postgraduate degree is not a prerequisite for the following categories:

   - specialisation in the Fine Arts (d. category)
   - collection of research data (e. category)
   - those who are legally residents in Greece whilst undertaking PhD studies.

- Applicants for postdoctoral research studies must hold a doctoral degree (PhD) completed at a foreign or a Greek University. However, their graduate degree (Bachelor) should be from a foreign University.

- For further education (in the following specific subject areas of study: Greek Language, Literature, Philosophy, History and Art), applicants should be professors in a Department of Greek studies at a foreign University. A certified document from this Department showing their employment must also be submitted along with all relevant documentation.

- Medical doctors should have either completed their specialty or hold a postgraduate degree (equivalent to Master's). The scholarship is strictly offered for postgraduate/doctoral studies and not for work or internship in a state hospital.

6. Have a very good command of English or French.
7. Have never been on a scholarship from the I.K.Y. (those who have been granted a scholarship for the IKY Programme "Greek Language and Culture" are entitled to apply).
8. Have not undertaken studies in Greece on a scholarship provided by any Greek authority.

Application instructions:

Applicants meeting the above requirements by the application deadline should submit the application form and requirements through the Greek Diplomatic Authorities in their country of residence (or in a neighbouring country in case of absence of a Greek Embassy or Consulate in their country) the following. The application deadline is **26 July 2013**.

It is important to visit the official website (link found below) for detailed information on how to apply for this scholarship.

Website:

Official Scholarship Website: http://www.iky.gr/en/press/item/789-announcement

Related Scholarships:

Eiffel Scholarships in France for International Students
University of Bern Masters Grants for International Students
Reach Oxford Scholarships for Developing Country Students
Emile Boutmy Scholarships in France for International Students

Figure 21: Snippet of a scholarship announcement (part two)

following xPath code:

"*//text()[preceding::p[contains(., 'Field(s) of study')]/strong] [following::p/strong[contains(., 'Target Group')]]*". The code could be understood as following:

- "//" indicates a relative path, so we are looking for the expression in the whole document;
- "text()" means that we are looking for the text;
- "Preceding" addresses the starting point of information extraction, and "following" - the end point;
- "p" indicates that we are looking for the tag <p> in the HTML file with the characteristics described in the square brackets "[]";
- "contains()" is a command that says that we are looking for an expression inside the paragraph;
- "strong" indicates that inside the paragraph <p> there will be tag (turns normal text into bold one).

This code might work for one scholarship announcement, but we need certain unified algorithm that can work for the majority of scholarships announcements. Hence, after analyzing all other 19 scholarships announcements, following differences have been identified:

- The section "Field(s) of Study" has multiple variations of the name. In particular, across 20 announcements, there were discovered following variations: "Fields of study", "Field of study", "Field and Level of study", "Field and Level of study", "Fields of study/Programmes", "Program of study";
- After the section "Field of study" there goes section "Target Group" or - "Number of scholarships" ("Number of awards", etc) and only then "Target Group". Hence, it was discovered that in some cases the section "Number of awards" is missing.

In order to deal with such differences, we can say that we want to find a paragraph that contains only "of study" which is written in bold manner, and the ending point of extraction can be either paragraph <p> containing "Number" in bold (we can't write "Number of awards", because there are also "Number of disciplines", etc) or the paragraph containing the word "Target", depending on the existence of "Number of awards" section:

- In case we have "Number of awards" section:
  "*//text()[preceding::p[contains(., 'of study')]/strong] [following::p/strong[contains(., 'Number')]]*";
- In case "Number of awards" section is missing:
  "*//text()[preceding::p[contains(., 'of study')]/strong] [following::p/strong[contains(., 'Target')]]*"

Figure 22 shows resulting retrieved text for the example scholarship announcement. Extracted information is accurate and precise.

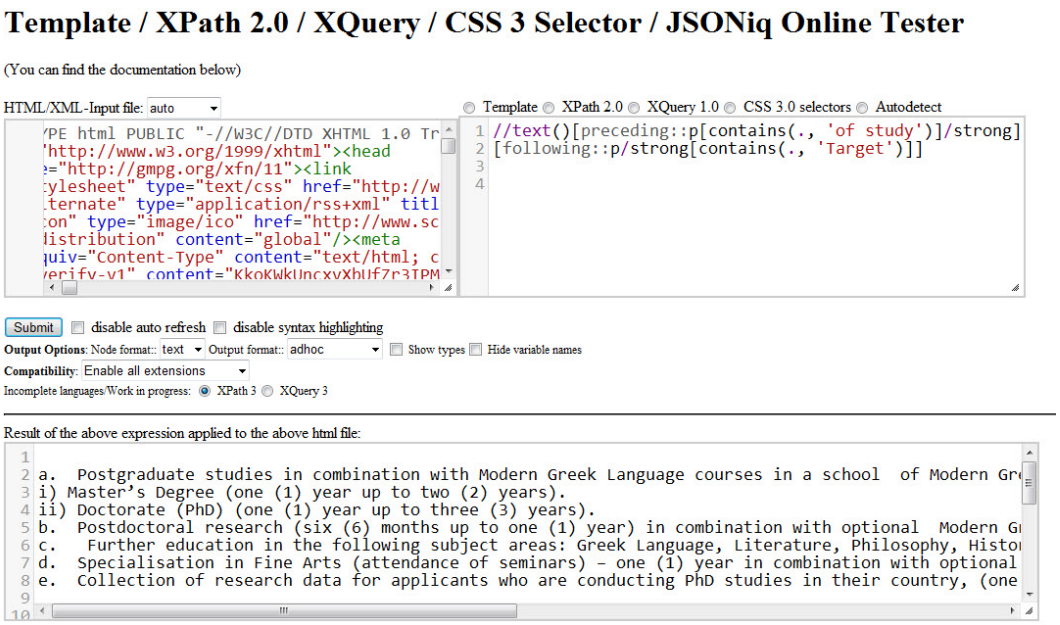


Figure 22: xPath Subject Of Study Web Scraping

4.2.2 How well a Boolean phrase-search will work for the discipline name extraction?

So, 20 "field of study" sections from tested announcements have been gathered in one place and alanyzed. How can we extract the names of the disciplines from the retrieved texts? Some of the scholarships announcements' snippets are shown on the Fig 23.

Text from the section "Field of Study" from 20 scholarship announcements has been searched for the names of disciplines. All 20 sections of "Field of study" contents are presented in the Appendices B. Search queries are disciplines' names that are encoded inside the Scholarship Ontology. In this section following experiments have been performed:

1. Full "phrasal" search on all levels of the ontology;
2. Breaking the query into parts, and performing search first on the superclass, then on the subclass, and, finally, on the instance.

In order to estimate, what percentage of the actual number of disciplines this method retrieved, we need to know how many disciplines are there in these 20 scholarship announcements. It is hard even for human to say clearly, when the name of discipline starts and when it ends (since by phrase-matching we can extract only a part of the discipline name). That is why, to avoid confusion, we will count disciplines in a way that "," and "and" will be considered as discipline separators. In this case the whole list of disciplines can be seen in Appendices C. In general, we can distinguish 82 disciplines in 20 documents, 75 of which are unique.

1. University of the People Online Tuition Free Degrees:

Fields of study:

UoPeople offers Associate and Bachelor Degree Programs in Business Administration and Computer Science.

2. Macquarie University International Scholarships (MUIS)

Field of study:

Available across most courses *except* Master of Business Administration (MBA), Master of Applied Finance (MAF), Master of Advanced Surgery and Master of Surgery.

3. Australia Awards Scholarships;

Fields of study:

Study programs must relate to your country's priority areas for development. These are listed on the participating country profiles.

4. Australian Leadership Awards Scholarships;

FIELDS AND LEVEL OF STUDY (CHANGE OF THE NAME)

Field and Level of study:

Only applicants applying to undertake a postgraduate course (Masters or PhD) are eligible to be considered for the supplementary ALA.

Study programs must relate to your country's priority areas for development. These are listed on the participating country profiles. AusAID Development Awards are not available for training in areas related to flying aircrafts, nuclear technology or military training.

5. Sydney Achievers International Scholarships;

Field(s) of study:

Any undergraduate or postgraduate coursework program offered at the University

6. SGU Commonwealth Jubilee Scholarship Program;

Field(s) of study:

The scholarships cover both graduate and undergraduate degree programs:

- Doctor of Medicine
- Doctor of Veterinary Medicine
- Master of Public Health
- Master of Business Administration
- School of Arts and Sciences

7. Netherlands Fellowship Program for Short Courses;

Field(s) of study:

See the NFP course list for 2013-14 for Short courses. Please note that this is a provisional list and that information is subject to change. Please regularly check the Nuffic website for the latest information.

Special Announcement: One can apply for NFP fellowships for 4 training courses offered at the The Hague Academy:

- Decentralization, Democratization and Development (March 11-22, 2013)
- Leadership and Municipal Management (April 8-19, 2013)
- Peacebuilding and Local Governance (May 27 to June 7, 2013)
- Citizen Participation and Accountability (June 17-28, 2013)

Figure 23: Snippets from the Field Of Study sections from different scholarships announcements, [scholars4dev.com]

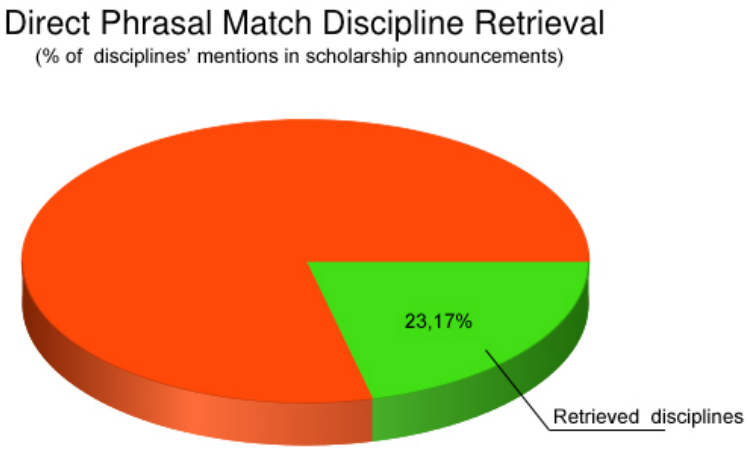


Figure 24: Direct Phrasal Match Discipline Retrieval

Direct Phrasal Search

First of all, we will try to make direct phrasal search, trying to find a match for ontological class/subclass/instance in the text.

By direct matching it was found 19 disciplines, with 13 of them being unique:

1. Computer Science
2. Economics
3. Educational Administration
4. Engineering
5. Higher Education Administration
6. International relations
7. Mathematics
8. Natural sciences
9. Public administration
10. Public Health
11. Sociology
12. Veterinary Medicine
13. Philosophy

During the matching process certain confusion has taken place: since certain ontological queries were written in the same form, but related to the different concepts and instances in the hierarchy. For example, there is a concept "Economics" inside the "Social and Behaviorial

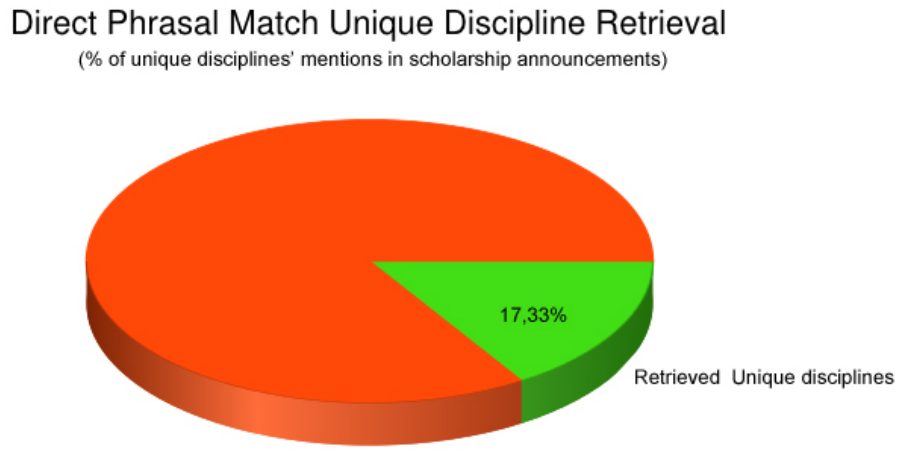


Figure 25: Direct Phrasal Match Unique Discipline Retrieval

Sciences" class which is a subclass, and then it also has "Economics" as one of the instances. Similar situation happened to the "Engineering", "Public Administration" and "Sociology" terms. The question is, whether to assign to discovered in the document disciplines the values as class (subclass) or instance?

In the case of scholarships, there need to be decided that we should use the higher level of hierarchy first when matching. That is why on practice it would be better to start the matching process by first trying to match higher levels of hierarchy (classes), then - subclasses, and only afterwards - instances. In this case we can neglect the matches on the lower levels if the ones on the higher level have already been discovered.

We can conclude that 23,17% of all disciplines have been successfully retrieved by direct "phrasal match" approach, Figure 24, which makes it 17,33% rate for unique disciplines, Figure 25.

4.3 How preprocessing can improve the results of the information extraction in the scholarship-related domain?

In this section different kinds of preprocessing will be implemented and discussed:

1. Change of the word order inside the query;
2. Breaking query into pieces;
3. Stemming of the words inside the query.

Also, after performing all kinds of preprocessing mentioned above, ontology evaluation process will be revisited and results will be analyzed.

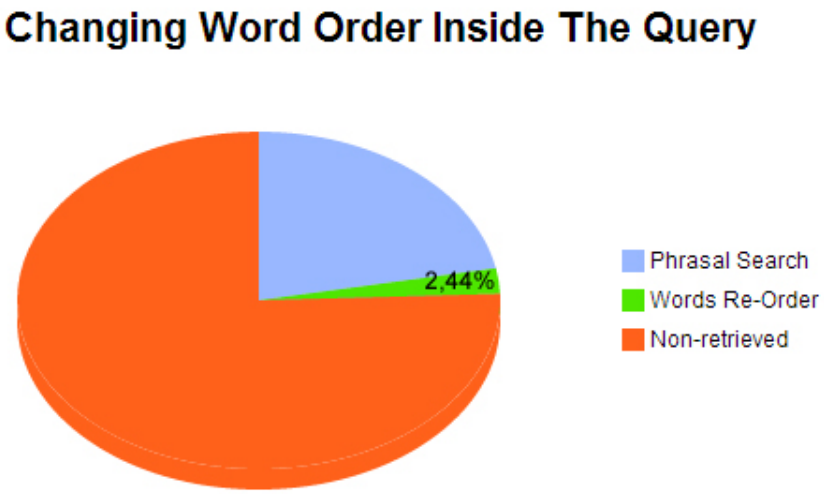


Figure 26: Changing The Word Order Inside The Query IR-retrieval

4.3.1 Change of the word order inside the query

1. For every element of the Discipline class in the Scholarship Ontology there have to be developed alternative versions, when the order of terms is changed, e.g., for the term "Chemical and Biomolecular Engineering" there can be used "Chemical and Engineering Biomilecular", "Biomolecular and Chemical Engineering", etc. Some versions could be not really human-readable, but some (like "Biomolecular and Chenical Engineering") could actually be useful. It is possible to automate the process of creating those alternative names (that have to be removed right after the search is performed) by means of combinatorics;
2. Next step would be to implement full phrasal search again, but on all those alternative names. Again, search should be implemented in top-down hierarchy fashion, searching for matches inside the higher classes first, and only afterwards - lower levels;
3. Results should be documented.

After the implementation of the experiment it was discovered two new matches in the discipline names:

1. Finance and Banking (we have originally "Banking and Finance" in our ontology);
2. Humanities and Arts (we have originally "Arts and Humanities" in our ontology).

For 82 disciplines that we have mentioned in our 20 scholarship announcements, extracting two new ones is sure a step forward. It means that extraction improved in 2,44%, Figure 26, and for unique disciplines (since all two are present in announcements just once and therefore are unique) - for 2,67%.

Breaking Query Into Pieces

After applying direct "phrasal" search, regarding ontological concepts as full queries, the next step is to break them into pieces. This can be done by regarding "and", slashes "/" and commas "," as separators, and that might be helpful, since certain concepts have rather complicated names, such as "Wildlife and Wildlands Science and Management", so it does not come as a surprise that direct match was not found. On the other hand, if we split the following query into pieces, we will get three options for the match: "Wildlife", Wildlands Science", "Management", which gives us more opportunities to find the match. There is one problem though: Management itself cannot be considered a discipline (it has to be Management of something), and many examples of disciplines' parts can be found that on their own cannot be considered a discipline - that is why specific list of such words - the words that we do not look for in text - should be made. It can be considered to be a kind of postprocessing: if we find match for "Management", we just ignore it. The words which should be neglected at the search, are:

- Management;
- Science (Sciences);
- Technology;
- Policy;
- Other;
- General;
- Conservation;
- Theory;
- Related;
- Instruction (Instructions);
- Talented;
- Services;
- Research;
- Language;
- Human;
- Leadership;
- Public;
- Administration;
- Securities;
- Development.

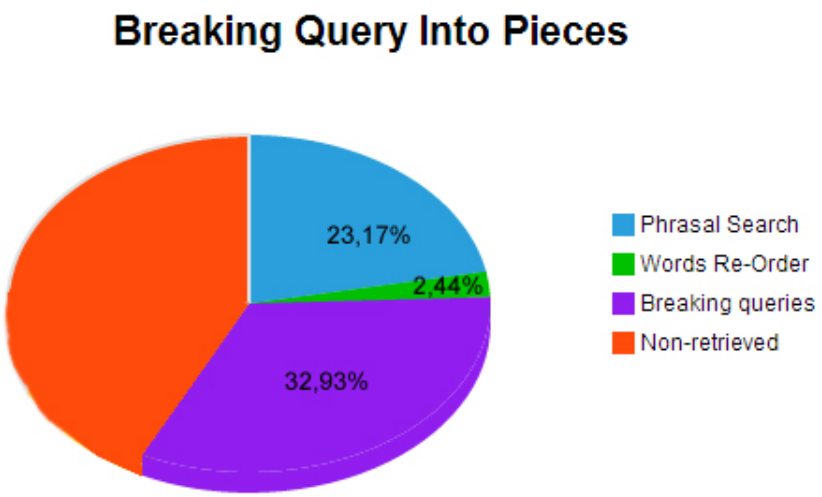


Figure 27: Breaking Query Into Pieces IR-retrieval

When implementing this strategy, we also need to make up another rule: we should start to look for disciplines' matches from the highest to the lowest level of hierarchy. That is important because we can have a lot of queries "Engineering" as a part of the original instance, and we will not know to what class to assign the found result. Nevertheless, when such situation occurs, and all found matches belong to the different instances in one subclass, we should assign to the found text the name of the whole subclass.

On the next iteration, it is better to remove these stop-words from inside the query when searching, for example, "Health and Medical Sciences" we turn into "Health", "Medical Sciences" first, and then into "Health", "Medical".

After adding to the list of full-phrasal search-based retrieved disciplines the ones found by described "breaking query" method, whole 48 disciplines could be retrieved, which is 58,53 % of all, Figure 27 and Figure 28.

The rates are quite high, nevertheless, certain issues encountered: disciplines which were considered by experts to be a single discipline, "Agriculture and Veterinary", "Business and Law", "History and Art" and "Sociology and Education " could be associated with different classes of the ontology. Basically, it is quite understandable, since in the "Agriculture and Veterinary" could be interested those who want to study Agriculture, as well as those who want to study "Veterinary". When such situation occurs, it'd be better to assign to this "one" discipline two labels, as it can add semantic meaning.

4.3.2 Stemming of the words inside the query

This preprocessing method includes several steps:

1. Ontology terms are exploded word by word;
2. Each word is then stemmed and all possible wordforms are set;

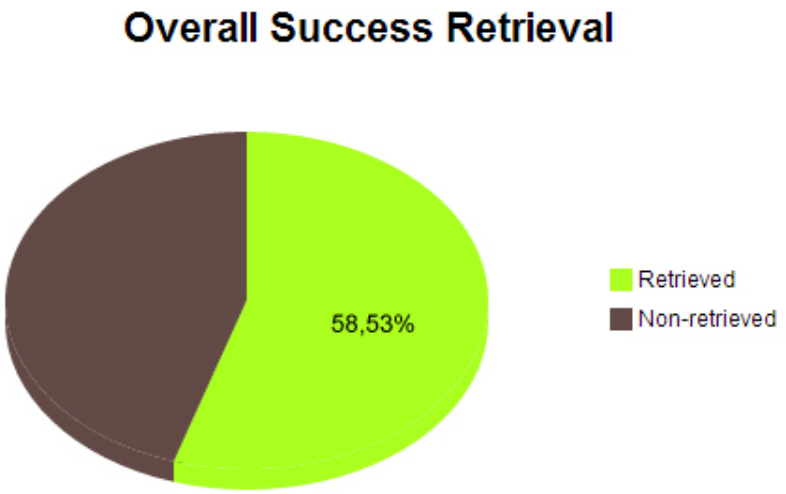


Figure 28: Overall Success Retrieval After Breaking Queries

3. By the means of combinatorics all possible variants of phrase (which is split already on the previous step) are then created;
4. Usual search is performed.

After performing this preprocessing method, following results were retrieved:

1. Agricultural and Rural Development (assigned to subclass Agriculture);
2. Economic Development (assigned to subclass Economics);
3. Economic Sciences (assigned to subclass Economics) ;
4. Human Resource Management (assigned to the instance Human ResourceS Management);
5. Medicine (assigned to subclass Medical).

These results form 9 out of 82 = 10,98% of retrieved disciplines, Figure 29, which overall means that by this point overall 69,51%, Figure 30, disciplines were retrieved.

Another approach to preprocessing includes the use of synonyms. The thing is, though, that topic of scholarships is not that much synonyms-rich: the majority of disciplines have unique names, and even if it will be possible to find several options for their terms, it still will be much less than in more general fields of knowledge, therefore the benefits of their usage is quite questionable.

4.3.3 Revisiting Ontology Evaluation

So, as we can see, the percentage of disciplines retrieved by all methods described above (Boolean phrasal search, Breaking query into pieces, Stemming and Stop-lists) is 69,51%. So what about other 30,49%? Is there a way to retrieve them?

Non-retrieved disciplines could be divided into two groups:

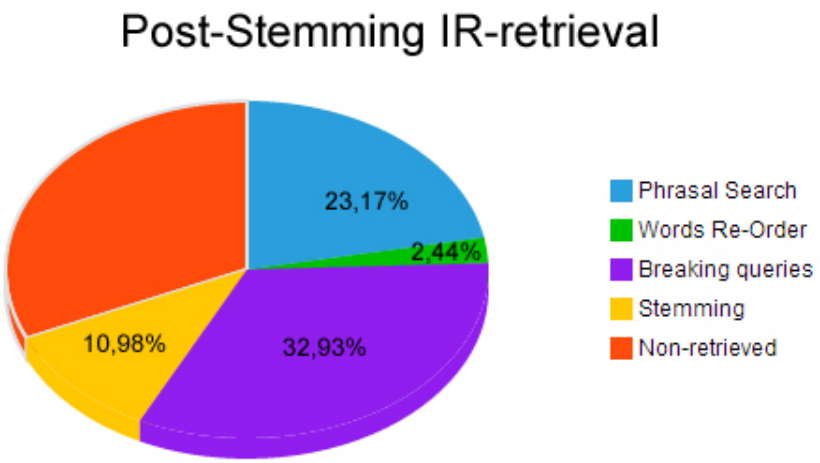


Figure 29: Post-Stemming IR-retrieval

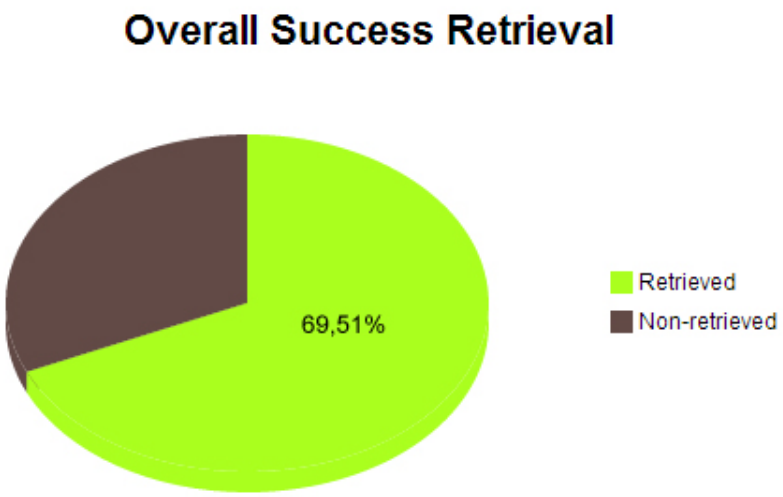


Figure 30: Overall Success Retrieval After Stemming

- Those that were not retrieved because they are not in the ontology (and they are even not the synonyms of those that are), e.g., "Surgery", "Conflict Resolution", etc;
- Those that are in the list of stop-words (or the combination of them), e.g., "Leadership and Municipal Management" (Leadership and Management are both stop-words).

Whereas it's quite hard to distinguish stop-words from the actual disciplines, the absence of certain concepts in the ontology is an easier issue to solve. Since on the previous steps of the ontology evaluation it was meant that experts will be the ones writing RDF-descriptions manually, therefore, it could be possible for them to assign, for example, "Surgery" to the "Health and Medical Sciences", for the computer it is quite hard to do. That is why only on this step it was determined that ontology should be extended and populated with the missing concepts.

Concepts that should be further added:

1. Surgery (inside Health and Medical Sciences subclass);
2. Citizen Participation (inside Political Science subclass);
3. Climate (inside Earth, Atmospheric, and Marine Sciences subclass);
4. Conflict Resolution (inside Political Science subclass);
5. Decentralization (inside Political Science subclass);
6. Democratization (inside Political Science subclass);
7. Law and Human Rights (new subclass inside Social and Behavioral Sciences superclass);
8. Peace studies (inside Political Science subclass).

Adding these concepts to the ontology will help to retrieve 9 more disciplines out of 82, which is 10,98%. It means that overall rate of retrieved disciplines will be 80,49%, Figure 31 and Figure 32 . This rate is quite high for the automatic discipline retrieval, and the disciplines which won't be retrieved in this case are either very specific and rare - or consist of primarily "stop-words", so their meaning is not completely identified.

4.3.4 Results Analysis

While describing the results obtained in the process of research, we were focusing more on recall rates (counting how many percent of relevant instances are retrieved), and didn't discuss much the precision rates (the percentage of retrieved instances that are relevant). In order to do that, we need to investigate what kinds of false positives we get during retrieval (if we get any).

We do have false positives retrieved in our experiment, and those can be divided into two major groups:

1. False positives which are not really "false" discipline names, but they just need to be added to another category in the ontology (which also should be created). Those are the disciplines for which scholarships are NOT provided and that are explicitly mentioned in the scholarship announcement. Technically, if we are saying that our task was to retrieve the names of the disciplines, they still fall under the category. Such disciplines were "Business Administration" (one of the times), "Applied Finance", "Surgery", "Advanced Surgery".

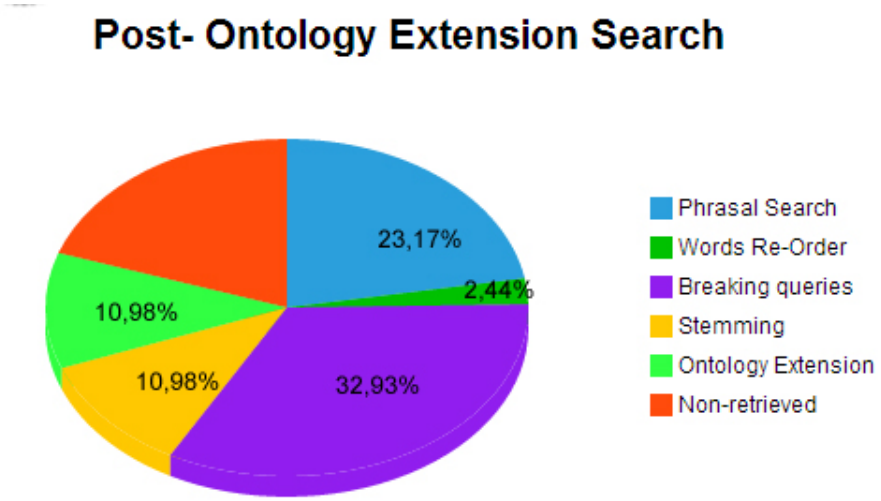


Figure 31: Post-Ontology Extension Search

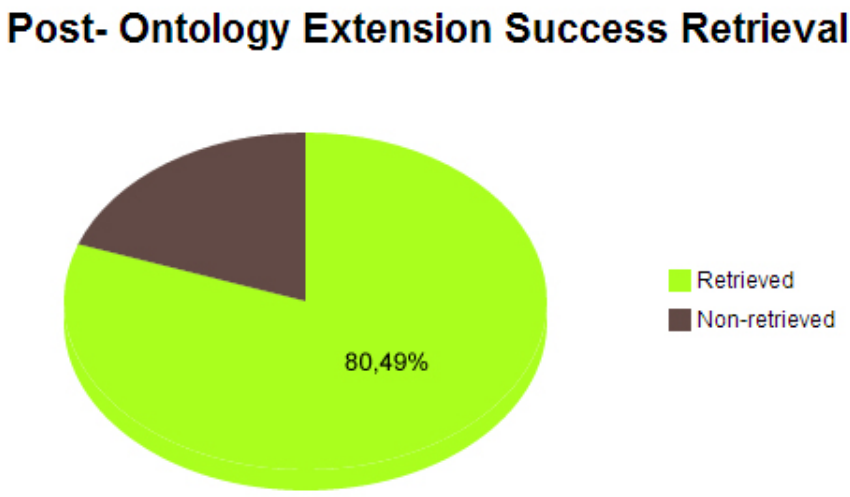


Figure 32: Overall Success Retrieval After Ontology Extension

So how do we deal with such cases? Further research on the subject should be made, but judging from the available cases, we can say that if there is a word "except" near the discipline name or the phrase "not available for", those disciplines should be assigned to the "notProvidedForDiscipline" (or some similar name) ontological property. However, it is just a hypothesis that should be tested in the Future Works.

2. The real "false positives" are the words that can be understood as a discipline but are not a discipline on practice. Due to the elimination of common-used words in a form of the stop-list, it was possible to avoid big numbers of false positives (for example, the word "development" can be found outside "disciplinary" scope, but it was eliminated). However, some words did fall into the category on false positives, those are two cases, "educational" and "medical".

Hence, if we want to calculate precision and recall (and we will mind all retrieved words that are disciplines as true positives in our calculations), we will have 70% recall (80% if extending the ontology), and 57/59=96,61% precision for the case without extending ontology, and 66/68=97,06% in case of its extension. Such high results in precision can be explained due to the lack of much additional context inside the "Field of study" section on a website and due to the use of stop-word list. If we were working with another website, the results could be much worse.

5 Discussions and Implications

5.1 What features should ontology for scholarships have, and how to evaluate and modify existing ones?

In the course of the research, it was determined that the most important feature of ontology is its expressiveness, and for its evaluation 20 scholarship announcements were selected and alanyzed. Ontology could not originally describe concepts of the set of countries ("Southern Africa", "European Union", etc), for that Geopolitical ontology has been imported. Also, new properties were added for describing full URL of the scholarship program and the number of available scholarships, etc.

To answer this research question we were using Individual Evaluation method, which has certain limitations, since it is based on the individual perception and doesn't take into consideration ideas of other people. If for writing Master's Thesis there was allocated more time, it would be beneficial to set up surveys and conduct in-depth one-to-one interviews with people in order to gain more insights on what possible queries people may have when/if searching for scholarships. Nevertheless, Individual evaluation was chosen as one of the most promising methods since it involves extensive analysis of the available real-life examples of scholarship programmes. The biggest drawback in such approach is a number of scholarships selected for the dataset, but because of limited timing it was hard to perform qualitative analysis on more than 20 scholarship announcements.

5.2 What kinds of information retrieval methods can be used for extracting knowledge in the scholarship content by the use of domain ontology?

For answering this question 20 scholarships announcements have been analyzed, the information inside the section "Field of study" was extracted by web scraping, and then Boolean phrasal search was performed to find a match of the ontological concept of discipline (superclasses, subclasses, instances) with a term/phrase in the section.

Used research methods were chosen based on the literature. For scraping xPath technique was chosen, simply because it performs well in the case of relatively structured content that we had on a tested website, so its capabilities were sufficient for our purposes. However, for the less-structured websites xPath is unlikely to give high levels of performance and, depending on the complexity of structure, other web scraping techniques could be used.

Searching for the "ontological" discipline names in the retrieved text were made by straightforward Boolean phrasal search. Even though there are many different search techniques available out there, in order to get direct match and to be able to assign certain term (or set of terms) to the certain class in the ontology, we don't need bigger functionality that just "phrasal" match, which alone brought 23,17% of matches, retrieving 19 out of 82 disciplines in the documents,

with 13 of them being unique.

5.3 How preprocessing can improve the results of the information extraction in the scholarship-related domain?

For the preprocessing several methods were chosen: change of the word order inside the query, breaking query into pieces and stemming. The decision to select these approaches came out of investigation of the disciplines that we wanted to retrieve, their morphological analysis and structure of ontological taxonomy.

It is in our clear realization that it was and is possible to use other methods of preprocessing, and more approaches would be used and compared if there were more time available. However, in the realms of rather short period of allocated for Master's Thesis time, it was decided to investigate the performance of the methods above, which, however, demonstrated good results. Overall retrieval was about 70%.

After implementing the methodology, the disciplines which were not retrieved were analyzed. It was found that some of them, surprisingly, were not present in the ontology in any form (although ontology has 472 disciplines). So it was suggested to extend ontology by populating it with new terms - which would increase discipline names retrieval levels for 10,98%, overall making for about 80% of retrieval.

Some of the disciplines that were found had to be assigned to the different ontological class, since they represented the subjects of study for which the scholarships were not provided. However, we cannot consider them as fully "false" positives, since they are disciplines, and we will need only to form a set of rules for assigning them to another category in the ontology (that also needs to be created). Real "false positives" though are not many, there were found only two of them ("medical" and "educational"), which gives us very high precision rate of 96,61% (for the case that ontology is not extended) and 97,06% for the case it is. Such high level of precision can be explained by the lack of much context in the "Field of study" section on this particular website, and by the use of stop-words that decreased possibility of retrieving non-relevant words such as "development", "management", "science", etc. It also means that for other websites the rates of precision and recall can be much lower.

We can also explain such high retrieval rates of discipline names because they represent entities that are rarely to be met across non-specific "disciplinary" texts and they don't have many synonyms. We can predict that data for another ontological concepts would be harder to retrieve, e.g., "hasNumberOfScholarships" - usually it's a number, and in the section "Number of awards" on the website sometimes there can be found several numbers, so identifying which one of them is related to the actual number of scholarships can be rather non-trivial task. Another examples are the concept "requiresLevelOfEducation", "hasFrequencyOfScholarship" - such information is usually written in the words from "general vocabulary", and it could be very hard to say if found word is related to our ontological concept or not. That is why we can predict that for the other fields retrieved rates will be much lower, and the amount of false positives will be also much higher.

5.4 Overall Remarks

Results obtained in the experiment show that around 70% of disciplines can be found in a text on a website and assigned to the certain classes or instances of the ontology. Achieving such high number of retrieved results was possible due to the combination of different IR-approaches, namely, Boolean "phrasal" search, breaking query into pieces, and specific preprocessing techniques (stop-words lists and stemming). Additional 10% in retrieval can be guaranteed when extending ontology by populating it with certain concepts that were not introduced before. The reason why ontology was not modified at the first step of the current research is because it was evaluated from the point of view of an expert who was going to assign disciplines to the classes manually, and not automatically.

Direct phrasal search brought only 23,15% of results - which was quite expected, since it is rare that the names of disciplines that consist of several words had exactly the same form as in the ontology. That is why it was necessary to add preprocessing - which overall increased retrieval efficiency in 46,36%, up to 69,51%.

Are obtained results representative? The dataset that was chosen for the purposes of the Thesis consists of 20 scholarship announcements, and, therefore, of course, in order to be able to make more reliable results, we would need bigger dataset - but for the time allocated for the research it was not possible to consider bigger dataset. We can predict that by taking bigger dataset into consideration, we would find more distinct disciplines that probably were not already in the ontology - therefore, it would be preferable to extend ontology again. Also, the name of the section "Field Of study" could have even more variants of notation, so suggested web scraping rules will need to be revisited. Also, if we were working with another website with poorer structure, it is very likely that xPath wouldn't have had sufficient performance, and we would need to search for another ways to scrape information from the section.

6 Conclusions and Future Work

The first research question, "What features should ontology for scholarships have, and how to evaluate and modify existing ones?" was answered by means of individual evaluation of Scholarship Ontology. 20 distinct scholarship announcements were chosen as a form of a representative dataset, and analyzed. Most important concepts that were described in the announcements were identified, and evaluated ontology was tested on a subject of its ability to describe them. According to the results of the evaluation, ontology was modified: Geopolitical ontology was imported to be able to describe geopolitical locations ("European Union", "Northern Africa", etc), the properties to describe number of scholarships and programme URL were added. Also, the links between concepts were "fastened" by adding more additional properties, "scholarshipForFuitionLanguage", "scholarshipProvidesDegree", etc. Modified active ontology has 85 classes and 50 properties (34 Object Properties and 16 Data properties).

The second research question, "What kinds of information retrieval methods can be used for extracting knowledge in the scholarship content by the use of domain ontology?" was answered by means of surveying literature and finding the most appropriate methods for extracting the ontological names of the disciplines from the free text. The method that was found and proved efficient is simple Boolean "phrasal" search - which, judging from the content of the text where the search was implemented, was sufficient enough to be used as a core "match finder", and retrieved 19 disciplines names out of 82, which makes 23,17%.

The third question, "How preprocessing can improve the results of the information extraction in the scholarship-related domain?" The methods that were chosen for preprocessing included: change of the word order inside the phrase query (+2,44%), breaking query into pieces (+32,93%) and stemming (+10,98%). Additional 10,98% could be retrieved by extending ontology by populating it with new concepts. Overall, Boolean search with preprocessing can retrieve 69,51% discipline names, and adding new concepts to the ontology increases this number up to 80,49%. Other 19,51% of disciplines' names are hard to retrieve due to their not completely defined meaning. That is why, our suggested automated retrieval method should better be supervised by the human expert. Arranging such an environment, where automated processes can be checked by the humans, would decrease the costs of the necessary human experts, or, in cases when it is affordable to lose around 20% of the information, the process could be implemented in an unsupervised fashion.

When it comes to the Future Works, it is possible to identify the following points:

1. Quantitative and qualitative evaluation of the ontology should be performed;
2. Ontology can be populated with more detailed concepts (test scores requirements, Master of Science/Master of Arts distinction, etc);
3. The rules for determining that scholarship is NOT provided for the discipline should be tested and formulated, and corresponding property and class should be added to the ontology;

4. The number of the scholarships announcements in the dataset should be extended;
5. Scholarships announcements should be taken from different websites;
6. Different ways of web scraping should be researched, and it should be taken into account that for different websites the scraping method can modify in complexity;
7. Performance of other methods and systems for Information Retrieval should be tested;
8. Semantic information from the other sections of the website should be retrieved based on the ontology;
9. Specific system could be developed which would combine all IR methods "under one roof" in order to extract information from the section more easily - and also from other sections of the scholarship announcement document. The way to automate generation of RDF-descriptions should also be developed.

Bibliography

[1] Brank, J., Grobelnik, M., & Mladenić, D. 2005. A survey of ontology evaluation techniques.

[2] Porzel, R. & Malaka, R. 2004. A task-based approach for ontology evaluation. In *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer.

[3] Lozano-Tello, A. & Gómez-Pérez, A. 2004. Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, 2(15), 1–18.

[4] Berners-Lee, T., Hendler, J., Lassila, O., et al. 2001. The semantic web. *Scientific american*, 284(5), 28–37.

[5] Hendler, J. 2009. Web 3.0 emerging. *Computer*, 42(1), 111–113.

[6] Cormode, G. & Krishnamurthy, B. 2008. Key differences between web 1.0 and web 2.0. *First Monday*, 13(6).

[7] Graham, P. 2005. Web 2.0. *Consultado (21/12/2008) en: http://www. nosolousabilidad. com/articulos/Web20. htm*.

[8] O'reilly, T. 2007. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1), 17.

[9] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. 2008. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, 1265–1266. ACM.

[10] 2013. W3c semantic web activity. `http://www.w3.org/2001/sw/`.

[11] Studer, R., Benjamins, V. R., & Fensel, D. 1998. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1), 161–197.

[12] Shadbolt, N., Hall, W., & Berners-Lee, T. 2006. The semantic web revisited. *Intelligent Systems, IEEE*, 21(3), 96–101.

[13] Yushtina, A. 2013. Advanced project work course.

[14] Yushtina, A. 2013. Research project planning course.

[15] Maedche, A. & Staab, S. 2002. Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web*, 251–263. Springer.

[16] Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. 2004. Data driven ontology evaluation.

[17] Tartir, S., Arpinar, I. B., & Sheth, A. P. 2010. Ontological evaluation and validation. In *Theory and Applications of Ontology: Computer Applications*, 115–130. Springer.

[18] Noy, N. F. & Klein, M. 2004. Ontology evolution: Not the same as schema evolution. *Knowledge and information systems*, 6(4), 428–440.

[19] Plessers, P. & De Troyer, O. 2005. Ontology change detection using a version log. In *The Semantic Web–ISWC 2005*, 578–592. Springer.

[20] Haase, P., Van Harmelen, F., Huang, Z., Stuckenschmidt, H., & Sure, Y. 2005. A framework for handling inconsistency in changing ontologies. In *The Semantic Web–ISWC 2005*, 353–367. Springer.

[21] Alani, H., Brewster, C., & Shadbolt, N. 2006. Ranking ontologies with aktiverank. In *The Semantic Web-ISWC 2006*, 1–15. Springer.

[22] Greengrass, E. 2000. Information retrieval: A survey.

[23] Porter, M. F. 1997. An algorithm for suffix stripping. In *Readings in information retrieval*, 313–316. Morgan Kaufmann Publishers Inc.

[24] Damashek, M. et al. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199), 843–848.

[25] Manning, C. D., Raghavan, P., & Schütze, H. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

[26] Barker, J. Basic search tips and advanced boolean explained. teaching library, university of california, berkeley. `http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Boolean.pdf`.

[27] Lynch, C. A. 1997. The z39. 50 information retrieval standard. *D-lib Magazine*, 3(4).

[28] Bradford, R. B. et al. August 19 2008. Word sense disambiguation. US Patent 7,415,462.

[29] Salton, G. & Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513–523.

[30] Alexopoulos, P., Pavlopoulos, J., Wallace, M., & Kafentzis, K. 2011. Exploiting ontological relations for automatic semantic tag recommendation. In *Proceedings of the 7th International Conference on Semantic Systems*, 105–110. ACM.

[31] Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., & Tasso, C. 2010. Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems*, 25(12), 1158–1186.

[32] Schlyter, E. 2007. Structured data extraction.

[33] Pomikálek, J. & Rehurek, R. 2007. The influence of preprocessing parameters on text categorization. *International Journal of Applied Science, Engineering and Technology*, 1, 430–434.

[34] Abels, S. & Hahn, A. 2005. Pre-processing text for web information retrieval purposes by splitting compounds into their morphemes. *Open Source Web Information Retrieval*, 7.

[35] Clark, J., DeRose, S., et al. 1999. Xml path language (xpath).

[36] Berglund, A., Boag, S., Chamberlin, D., Fernandez, M. F., Kay, M., Robie, J., & Siméon, J. 2007. Xml path language (xpath) 2.0. *W3C recommendation*, 23.

[37] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., & Sachs, J. 2004. Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 652–659. ACM.

[38] 2013. Watson search engine. `http://watson.kmi.open.ac.uk/WatsonWUI/index.html`.

[39] 2013. Scholarship for development. `http://www.scholars4dev.com/`.

[40] 2013. Rdf validator and converter. `http://www.rdfabout.com/demo/validator/`.

[41] 2013. W3 validation service. `http://www.w3.org/RDF/Validator/`.

[42] 2013. Geopolitical ontology - homepage. `http://www.fao.org/countryprofiles/geoinfo/en/`.

A List of scholarship announcements selected for testing

All scholarship announcements were taken from th website "Scholarships for Development", www.scholars4dev.com.

1. University of the People Online Tuition Free Degrees;
2. Macquarie University International Scholarships (MUIS);
3. Australia Awards Scholarships;
4. Australian Leadership Awards Scholarships;
5. Sydney Achievers International Scholarships;
6. SGU Commonwealth Jubilee Scholarship Program;
7. Netherlands Fellowship Program for Short Courses;
8. DAAD Scholarships with Special Relevance for Developing Countries;
9. USA Fulbright Scholarships for International Students;
10. Rotary International Peace Fellowships;
11. MasterCard Foundation Scholarship Program for Africans;
12. Hubert Humphrey Fellowships in USA for International Students;
13. Greek Government Scholarships for Foreign Students;
14. Denmark Government Scholarships for International Students;
15. Quota Scholarships in Norway for Developing Countries;
16. VLIR-UOS Scholarships for Developing Countries;
17. Aga-Khan Foundation International Scholarship Programme;
18. Joint Japan/World Bank Scholarships in Development for International Students;
19. La Trobe Academic Excellence Scholarships for International Students;
20. Erasmus Mundus Scholarships for Developing Countries.

B Contents of the Field Of Study section for 20 scholarship announcements

1. UoPeople offers Associate and Bachelor Degree Programs in Business Administration and Computer Science.
2. Available across most courses except Master of Business Administration (MBA), Master of Applied Finance (MAF), Master of Advanced Surgery and Master of Surgery.
3. Study programs must relate to your country's priority areas for development. These are listed on the participating country profiles.
4. Only applicants applying to undertake a postgraduate course (Masters or PhD) are eligible to be considered for the supplementary ALA.
   Study programs must relate to your country's priority areas for development. These are listed on the participating country profiles. AusAID Development Awards are not available for training in areas related to flying aircrafts, nuclear technology or military training.
5. Any undergraduate or postgraduate coursework program offered at the University.
6. The scholarships cover both graduate and undergraduate degree programs:
   - Doctor of Medicine
   - Doctor of Veterinary Medicine
   - Master of Public Health
   - Master of Business Administration
   - Master of Business Administration
7. See the NFP course list for 2013-14 for Short courses. Please note that this is a provisional list and that information is subject to change. Please regularly check the Nuffic website for the latest information.
   Special Announcement: One can apply for NFP fellowships for 4 training courses offered at the The Hague Academy:
   - Decentralization, Democratization and Development (March 11-22, 2013)
   - Leadership and Municipal Management (April 8-19, 2013)
   - Peacebuilding and Local Governance (May 27 to June 7, 2013)
   - Citizen Participation and Accountability (June 17-28, 2013)
8. Postgraduate courses are offered in the following fields:
   - Economic Sciences / Business Administration / Political Economics
   - Development Co-operatio

- Engineering and Related Sciences
- Mathematics
- Regional Planning
- Agriculture and Forest Sciences
- Environmental Sciences
- Medicine and Public Health
- Veterinary Medicine
- Sociology and Education

View the list of selected programmes for 2013/2014.

9. Fulbright grants are available for a variety of disciplines and fields, including the performing and visual arts, the natural sciences, mathematics, engineering and technology. Fulbright encourages applications from all fields, including interdisciplinary ones.
   The Fulbright Program will not fund applicants seeking to enroll in a medical degree program nor does it offer grants to those who wish to conduct clinical medical research or training involving patient care and/or contact but it supports the fields of public health and global health.
   See the country-specific websites (link found in contact information) for updated information on approved field of studies.
10. Fellows can study either a master's degree in international relations, public administration, sustainable development, peace studies, conflict resolution, or a related field, or a professional development certificate in peace and conflict resolution.
11. Approved fields of studies offered by participating Universities/Institutions. See the MasterCard Foundation Scholarship pages of the Universities/Institutions above.
12. - Agricultural and Rural Development
    - Communications/Journalism
    - Economic Development
    - Educational Administration, Planning and Policy
    - Finance and Banking
    - Higher Education Administration
    - HIV/AIDS Policy and Prevention
    - Human Resource Management
    - Law and Human Rights
    - Natural Resources, Environmental Policy, and Climate Change

- Public Health Policy and Management

Public Policy Analysis and Public Administration
Substance Abuse Education, Treatment and Prevention
Teaching of English as a Foreign Language
Technology Policy and Management
Trafficking in Persons Policy and Prevention
Urban and Regional Planning

13. a. Postgraduate studies in combination with Modern Greek Language courses in a school of Modern Greek Language, at a state Greek University only during the first year of the scholarship.
    i) Master's Degree (one (1) year up to two (2) years).
    ii) Doctorate (PhD) (one (1) year up to three (3) years).
    b. Postdoctoral research (six (6) months up to one (1) year) in combination with optional Modern Greek Language courses in a School of Modern Greek Language, at a state Greek University.
    c. Further education in the following subject areas: Greek Language, Literature, Philosophy, History and Art aimed for professors of Greek studies at Universities abroad (six (6) months up to one (1) year).
    d. Specialisation in Fine Arts (attendance of seminars) - one (1) year in combination with optional Modern Greek Language courses in a School of Modern Greek Language, at a state Greek University.
    e. Collection of research data for applicants who are conducting PhD studies in their country, (one (1) year).
14. Approved full-time postgraduate programmes.
15. Any approved academic program offered by selected colleges, universities, and institutions in Norway. Most of the Norwegian institutions offer courses and educational programmes in English. Please refer to the websites of participating Norwegian institutions for more information on programmes they offer under the Quota Scheme.
16. The scholarships support development-related fields of study. The list of supported Training and Master's Programme for 2014/2015 and its description is found at this page.
17. Any; not specified.
18. Eligible applicants should propose a program of study related to development at the master's level, in fields such as economics, health, education, agriculture, environment, natural resource management, or other development-related subject.
    The proposed program of study should start during the academic year 2013/2014 for a maximum duration of two years. The JJ/WBGSP does not support applicants who are already enrolled (i.e., taking classes) in graduate degree programs. Applicants should submit evidence of current unconditional admission to at least one development-related university master's degree program and are encouraged to submit application to a second such program. Applicants are encouraged to apply to one of the Preferred Universities which, other things equal,

will have priority in the scholarship award. The Program does not support studies in the applicant's home country.
The Program does not support applicants for MBA, MDs, M.Phil. or Ph.D. degrees.
The Program does not support legal studies such as J.D., L.L.M. or S.J.D. except for L.L.M.'s related to human rights, environment, or good governance.
The scholarship program does not sponsor undergraduate studies, distance learning programs, short-term training, conferences, seminars, thesis writing, research projects, and fields of studies not related to development. All these requests will not be considered.
The Program does not support certain other fields of study.

19. Any full-time undergraduate or postgraduate coursework offered by the University.
20. For 2014-2015, about 138 Masters courses and 42 Joint Doctorate courses are supported by scholarships. The field(s) of study covered are: Agriculture and Veterinary, Engineering, Manufacture and Construction, Health and Welfare, Humanities and Arts, Science, Mathematics and Computing, and Social Sciences, Business and Law.

C List of separated discipline names from 20 scholarship announcements

Full list of disciplines (considering "and" and "," to be discipline names' separators when they are outside lists) is shown below (82 in a whole, and 75 of those are unique).

1. Business Administration
2. Computer Science
3. Business Administration
4. Applied Finance
5. Advanced Surgery
6. Surgery
7. Medicine
8. Veterinary Medicine
9. Public Health
10. Business Administration
11. Arts and Sciences
12. Decentralization
13. Democratization
14. Development
15. Leadership and Municipal Management
16. Peacebuilding and Local Governance
17. Citizen Participation and Accountability
18. Economic Sciences
19. Political Economics
20. Development Co-operation
21. Engineering and Related Sciences
22. Mathematics
23. Regional Planning
24. Agriculture and Forest Sciences
25. Environmental Sciences

26. Medicine and Public Health
27. Veterinary Medicine
28. Sociology and Education
29. Performing and Visual arts
30. Natural sciences
31. Mathematics
32. Engineering
33. Technology
34. Medical
35. Clinical medical research
36. Public Health
37. Global Health
38. International Relations
39. Public Administration
40. Sustainable Development
41. Peace studies
42. Conflict Resolution
43. Agricultural and Rural Development
44. Communications and Journalism
45. Economic Development
46. Educational Administration
47. Finance and Banking
48. Higher Education Administration
49. HIV and AIDS Policy and Prevention
50. Human Resource Management
51. Law and Human Rights
52. Natural Resources
53. Environmental Policy
54. Climate Change
55. Public Health Policy and Management
56. Public Policy Analysis and Public Administration

57. Substance Abuse Education
58. Treatment
59. Prevention
60. Teaching English as a Foreign Language
61. Technology Policy and Management
62. Trafficking in Persons Policy and Prevention
63. Greek Language
64. Literature
65. Philosophy
66. History and Art
67. Urban and Regional Planning
68. Economics
69. Health
70. Education
71. Agriculture
72. Environment
73. Natural Resource Management
74. Agriculture and Veterinary
75. Engineering
76. Manufacture and Construction
77. Health and Welfare
78. Humanities and Arts
79. Science
80. Mathematics and Computing
81. Social Sciences
82. Business and Law