Semantic Annotation of unstructured Wiki Knowledge according to Ontological Models

Roberto Boselli, Mirko Cesarini, Fabio Mercorio, Mario Mezzanzanica
Department of Statistics and Quantitative Methods / CRISP Research Centre
University of Milan Bicocca
Milan, Italy
Email:{roberto.boselli@unimib.it, mirko.cesarini@unimib.it, fabio.mercorio@unimib.it, mario.mezzanzanica@unimib.it}

Abstract— **The paper deals with the issue of supporting users to enrich unstructured wiki contents with semantic annotations. The authors present the development of a semantic wiki that provides semantic annotations compliant to ontological models. The semantic wiki developed, called WiWork, is presented with two related ontological models, the WiWork Core ontology and the Labour Market ontology. Finally, a methodology to define ontology concepts and properties is proposed by using Information Retrieval and statistical techniques.**

Keywords: Semantic Web; Semantic wikis; Ontology modeling; Unstructured data; Information Retrieval

I. INTRODUCTION

Wikis are collaborative web-based environments allowing users to create, share and reuse useful knowledge. Wiki's knowledge is mainly represented by an amount of unstructured textual contents. Unfortunately, the big volume and the lack of structure make often such contents inaccessible or hard to reuse.

A recent trend focuses on enhancing wikis with Semantic Web technologies and formal languages to access and reuse data, included the unstructured ones: namely semantic wiki. Semantic wikis are promising tools providing the users an easy way to manage machine-processable knowledge, allowing the creation of added-value services based on the semantics of Web pages. They support metadata insertions through semantic annotations and link relations between wiki pages. Annotations are required to refer to an ontological model defining concepts and properties that can be associated to pieces of wiki contents.

In the Semantic Web, ontologies are mainly developed and encoded with formal languages like RDF (Resource Description Format) [1] or OWL (Web Ontology Language) [2]. Moreover, ontology modeling and updating tasks are still hard to be accomplished for common users, and semantic wikis need ontologies as conceptual models to structure their contents.

The work described in this paper aims to provide to the users an easy way to annotate and to structure wiki knowledge without requiring the learning of formal ontology languages like RDF or OWL. The research questions investigated in this paper are, firstly, how to guide users in annotating semantic wiki contents according to ontological models, and secondly, how to build domain ontologies taking unstructured knowledge as source of concepts and relations. The proposed solution for the first question is the development of a semantic wiki, called WiWork, based on Semantic Mediawiki (SMW) [3]. The latter solution is addressed through the synergy between various research fields: ontology modeling, Information Retrieval (IR) and statistical methodologies. This synergy supported authors in defining two domain ontologies, the WiWork Core ontology and the Labour Market ontology.

The paper is organized as follows: in Section 2 the issue of annotating wiki's unstructured knowledge is discussed, and the semantic wiki WiWork is presented with some of its main functionalities; Section 3 describes the methodology used for ontology modeling based on Information Retrieval and statistical techniques and shows the produced ontologies; in Section 4 a brief survey of related works is provided; finally, some concluding remarks and the future works are outlined in Section 5.

II. UNSTRUCTURED WIKI KNOWLEDGE

Wikis are being confirmed as the most widespread collaborative technology supporting knowledge creation and sharing on the Web [4]. Mainly, wikis allow users to collaboratively create and maintain textual unstructured content by adding, changing, and sharing contents by means of a web browser and a simple markup language. A wiki limitation is that the contents are expressed using natural language text and its meaning is not directly accessible to automated semantic processing tasks. Knowledge in a traditional wiki is freely structured through pages, hyperlinks and user-generated tags (e.g., categories to label pages). Currently, the features available for searching and reusing wiki knowledge are based on full text keyword search. Therefore, wiki knowledge is not automatically accessible for functionalities such as querying, reasoning and semantic browsing. Semantic wikis add these functionalities (with respect to ordinary wikis) for managing knowledge in more formal, machine-processable ways [5]. To achieve this goal semantic wikis have to structure knowledge according to some conceptual models (i.e., ontologies) and to annotate texts with metadata (semantic links and properties). Ontologies written in RDF or in OWL are used in semantic wikis as a reference for concepts used in metadata and annotations.

Adding semantic annotations to wiki's textual content is a complex and laborious process, mainly for common users who have no competencies in formal languages or knowledge representation methodologies. One solution can be to provide tools facilitating the annotation tasks, a step further is to suggest the users the metadata to use during the creation of pages. Since metadata and semantic annotation are machine-computable in a semantic wiki, they can be either represented directly in RDF. RDF is a common standard to enrich content with formal semantic metadata, but is not easily understandable to common users. To avoid this issue it is necessary to transform RDF triples extracted from ontologies in user-friendly semantic wiki annotations. Various types of semantic wiki annotations can be provided, distinguishing between formal and semi-formal annotations. Tagging is a type of semi-formal annotation. Tags usually consist of keywords (i.e., categories and properties) that the users include in the text of wiki's pages. Semantic links are another type of semi-formal annotation, they consist in structured tags.

According to [6], structured tagging is a semi-formal annotation type in the form of *keyword:value* pairs. This annotation type allows for a simple representation of formal RDF triples: the resource to annotate is the subject, the keyword is the predicate and the value is the object of the triple. Such semi-formal annotation provides an intuitive and easy means to the users for transforming knowledge from human-only to machine-processable content and vice versa. In this way, they provide low barriers for user participation on wiki content enrichment with metadata. The methodology presented in this paper aims to help users to define semi-formal annotations derived from formal ones.

In the next sections, the authors present the main requirements to address the development of a semantic wiki, and, specifically, to enrich wiki content with semantic annotations according to well-defined ontologies.

A. WiWork, a semantic wiki for public services domain

In this section the authors describe how a semantic wiki for the CRISP (Interuniversity Research Centre on Public Services) [7] was developed. The research centre develops models, methodologies and tools for collecting, analyzing, and supporting data useful to define and improve services and policies for the public sector. Specifically, the centre designs, develops and uses information systems, Statistical Information Systems and portals for analyzing labour demand and supply [8]. The semantic wiki introduced in this paper, WiWork, was designed with the goal to provide documentations and manuals to domain experts and common users interested in CRISP activities and topics (WiWork is actually not accessible on the Web, but only on the CRISP intranet). Indeed, the wiki has actually about 14000 pages containing knowledge about the public services and focusing on the labour market, on the health, and on the education domains. WiWork has been developed to enrich labour market knowledge with semantic annotations to support tasks, such as matching between different texts, querying, and reasoning.

The content in WiWork is mainly unstructured text devoted to describe the taxonomy of occupations of the Italian labour market. The wiki pages on occupations were generated taking as primary source the Italian classification scheme of occupations (called CP2011) created by Italian National Institute of Statistics (ISTAT) [9]. This scheme is substantially based on the International Standard Classification of Occupation (ISCO08) criteria [10].

The classification scheme arranges all the occupations in four level groups: Major, Sub-major, Minor and Unit. Moreover, each occupation is classified according to the kind of work performed, the skill level, and the specialization required to fulfill tasks and duties of the job. The scheme provides, for each occupation, a textual description explaining: specific tasks, competencies, required field of knowledge, tools and machinery used, materials used, kind of goods and services produced, etc. The corpora of those textual descriptions contain the most of the domain terminology, mostly in an unstructured form. An example of textual description is the following, extracted from ISCO08: "*Information and communications technology service managers plan, direct, and coordinate the acquisition, development, maintenance and use of computer and telecommunication systems.*" Such a description contains some terms that can be processed only if properly structured, thus requiring the development and the implementation of various semantic technologies and the modeling of ontologies to enrich WiWork content with semantics.

B. Semantic annotations in WiWork

WiWork is based on Semantic MediaWiki (SMW) [3], an extension of the well-known MediaWiki, the software used for running Wikipedia. SMW allows one to add metadata to wiki pages and to perform queries over the metadata. Such extension makes the users able to enrich the text using semantic annotations in a way that is accessible both to the human readers and to the machines. The content can be created and maintained collaboratively, while the semantic insertion creates a flexible, extensible, and structured knowledge representation that can be automatically processed, enabling features such as semantic search and RDF data export.

In particular, SMW provides the means to develop a flexible, structured data schema consisting of properties and classes. Properties correspond to semantic links or tags inserted into the page text; classes correspond to categories grouping pages. Both classes and properties are the metadata that a user can manipulate in the wiki using the *keyword:value* form.

The following string *[[uses::computer]]* is an annotation example for the occupation description mentioned above, expressed in SMW syntax. This annotation can be transformed into a RDF triple where an instance of the *Occupation* concept is the subject having a semantic property *uses* as predicate, and an instance of *Tool* concept, *computer,* as object.

This kind of annotation in SMW can be used for semantically enriching both the link between two pages (if *computer* would be a page title) and the property value (if *computer* would be managed as a string of a property value). In both cases it is necessary to define the meaning of properties and concepts and make it available for wiki users.

What follows is the RDF code exported by WiWork of the example:

```
    <property:uses
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"> computer
    </property:uses>
```

Authors used the Pywikipediabot framework [11] for creating an initial draft of the pages containing the raw text descriptions of the occupations taxonomy. A bot is an application for creating, accessing, and modifying wiki pages via scripts. To this scope some templates were created: a first template to structure all the pages containing information about occupations, including the description, the related categories and the occupation position into the classification scheme (e.g., to which Major group the occupation belongs); a second template structures pages according to the economic sectors related to the occupations, and a third template structures all the information about professional skills linked to occupations. The bot manages the three templates to automatically create pages on the occupations.

Moreover, Semantic Forms, an extension of SMW for data editing [12], was used to facilitate the users to annotate wiki texts with metadata, linking the templates with the forms. The forms provide a graphical user interface allowing the users to create and exploit the templates without requiring the usage of wiki markups. Forms can be composed of several types of fields, e.g., title of occupation, name of economic sector etc. Each field can be related to a specific semantic annotation (e.g., *uses* property), by entering a value in a field (e.g., *computer*), the value is assigned to the corresponding property. All the information inserted via forms is automatically annotated using the properties specified during the template design time. Therefore, the combination of forms and templates allows the users to have a set of predefined semi-formal framework for structuring and annotating information.

Furthermore, the authors plan to develop a program that reuse Protégé-OWL API [13] and Pywikipediabot to automatically transform RDF and OWL triples created with the ontology editor Protègè [14] in SMW constructs (semi-formal annotations) to insert in WiWork pages. A list of categories and properties is necessary to enrich the fields of templates and forms for automatically create annotated pages. The list can be provided taking OWL classes and properties modeled in domain ontologies, as is described in the next Section.

III. Domain ontology modeling

In the following, the authors explain how two domain ontologies were built, the WiWork Core ontology and the Labour Market ontology, representing the conceptual models of the WiWork's content. Ontologies define a set of primitives, e.g., classes and properties, modeling a domain of knowledge [15]. Domain ontologies provide a shared and formal description of the concepts and relationships of a domain grounded on the knowledge and terminology in use in the domain [16].

The goal of this ontology modeling task is to enrich the semantic wiki WiWork with conceptual models that can be used to structure and to manage knowledge in a machine-processable way.

The adopted ontology development methodology was divided in two steps. First step was supported by the Protégé ontology editor to define the preliminary classes and properties in OWL [14], i.e., the CRISP core concepts and the categories used in WiWork pages. Second step enriched and expanded the ontologies with additional concepts and instances, i.e., the domain concepts. Both the ontology modeling tasks were performed with the support of Information Retrieval techniques to extract concepts from the texts. More precisely, the latter task was performed to identify a set of candidate words and relations to model the occupation terminology.

A. WiWork Core Ontology and Labour Market Ontology

Ontologies are designed to define concepts and relationships related to the WiWork knowledge and they are still under development. The methodology described in this paper helped designing and structuring two different OWL ontologies. The first is the WiWork Core ontology representing the conceptual model defining the main topics, research fields, activities, communities, users, and technologies managed by the CRISP.

The second is the Labour Market ontology, being the conceptual model of one specific WiWork topic, used to model the wiki pages about labour market knowledge. Specifically, this ontology has the specific aim to provide a set of predefined metadata that should facilitate users in annotating the occupation descriptions within the wiki. Although the ontologies are being developed separately, they can be seen as two integrated parts of a global conceptual model. In this model the WiWork Core concepts are in the upper levels, and the Labour Market concepts are in the lowest levels. The integration between the ontologies is obtained through the identification of relationships connecting similar or equivalent concepts, and through the identification of relationship types connecting the concepts and the instances extracted from the documents, as described in next Section B.

B. Modeling via IR techniques

Since textual documents have an important role in sharing knowledge and concepts about a domain, a corpus of textual documents related to labour market was selected to automatically extract domain knowledge. Documents are mainly unstructured texts focusing on occupation descriptions derived from the ISTAT classification scheme, and on textual documents produced by domain experts.

An experiment has been performed by focusing on a sub-set of occupation descriptions extracted from the whole corpora (about 400 descriptions). The subset has been identified in this way: data about worker transitions among the several occupations has been considered. The data of labour market are extracted from the administrative archives of an Italian Region. The term transition refers to a worker dismissing an employee and starting a new one having a different qualification. The transitions from an occupation to a different one have been counted, in a given observation period. A high transition number means that the two considered occupations are related. Then an occupation subset was selected performing a cluster analysis using the transition count as relatedness criterion [17][18]. This allows one to identify groups of related occupations representing consistent areas and coherent worker career paths. It is out of the scope of this paper to describe in detail the clustering algorithm, it is enough to report that several clusters were identified, and a cluster of 30 occupations related to ICT domain (in the following ICT cluster) has been selected for further analysis. A few operations performed on the ICT cluster are described below for extracting terms and expanding the ontology.

1) Cleansing and stopwords elimination

In this first operation, the documents containing the occupation descriptions are cleansed by removing the stop words, e.g., articles, pronouns, adjectives, prepositions and conjunctions, common adverbs and non-informative verbs (e.g., to be). The stop words carry on little informative content, indeed they get discarded. A list of stop words was built and checked in the documents contents to eliminate these words from it.

2) Keyword extraction, indexing and weighting

The second operation is extracting the keywords and measuring their degree of importance within the processed documents corpora. To this scope, an IR technique was used, namely the TF-IDF (Term Frequency, Inverse Document Frequency) weighting scheme to extract, and assign weights to distinguished terms in a document [19][20][21]. Each document is described by a set of representative keywords called index terms. An index term is simply a word whose semantics distinguishes the document's main themes. The assignment of numerical weights captures the ability of an index term to distinguish the document's main themes.

The intuition behind using the TF-IDF is that the best term candidates for inclusion in the ontology are those featured in certain individual documents, capable of distinguishing them from the remainder of the collection. This implies that the best terms should have high term frequencies (within the document), but low overall collection frequencies (i.e., it appears in very few other documents). The term importance is therefore obtained by multiplying the term frequency with the inverse document frequency (see for further details [22]). The result of the operation is a set of weighted keywords (i.e., nouns and verbs). The ones having the highest values identify the meaning and the terminology of documents. In the experiment performed 619 index terms were extracted from the ICT cluster.

3) Grouping terms in two groups, common and specific concepts

In this phase the set of index terms is analyzed to distinguish the more generic from the more specific. This distinction is just required to identify the position that each term could take within an ontology (i.e., in which class to include them). Some of the generic terms have been allocated to fill the WiWork Core ontology classes, while the specific terms have been used to fill the Labour Market ontology classes. On the basis of the results of previous operations, it has been observed by domain experts that the generic terms are those having the lowest TF-IDF values, while the specific ones have the highest values. Through few operations of further cleansing (e.g., all the repetitions arising when considering the terms of several documents) a set of 360 index terms was obtained. In this group, the 70 highest TF-IDF values were selected representing the specific terms of the ICT cluster (19%). These terms are the best candidates to be defined and included as instances in the Labour Market ontology. Some of the remaining terms are the candidates for the WiWork Core ontology.

4) Relationships identification

This task is still under development, during the experiment it was performed manually, but in the future the authors would to improve these steps with semi-automatic processes. This operation took as input the two groups of terms (i.e., the specific and the generic set) and started manually building the relationships connecting the ontology concepts and instances. Some relationships can be mainly induced by the verbs connecting relevant words in the occupation descriptions. The set of 70 highest TF-IDF values includes some verbs representing specific actions or tasks of only a few occupations. An example of this type is the verb *to write* extracted from the description of *Journalist* occupation. It can be easily put as instance of class *Task,* in relation with the instance *Journalist* of class *Role* (see below the ontology classes). While the most generic verbs, e.g., *to control*, *to perform*, *to use,* that more frequently appear in the occupation descriptions (and having the lowest TF-IDF values), describe common actions and tasks of many occupations. They can be managed as relations linking instances of, e.g., *Occupation* or *Role* classes with *Tool* or *Product* classes.

5) Concepts and relationships integration into ontologies

Nevertheless, ontologies modeled with index terms extracted from corpora of documents hardly represent an exhaustive terminology of a domain. An integration of domain concepts is required to provide a complete knowledge of that domain.

The approach planned for integrating other concepts and relationships into the ontologies is based on the comparison of terms and relations (identified during the previous operations) with classes and properties defined in domain dictionaries or specialized taxonomies as external data sources. In the Labour Market domain some specialized taxonomies, i.e., ILO (International Labour Organization) and O*Net [23][24], are internationally shared as the best semantic resources for the domain terminology. They provide the skeleton on which such ontologies can be built, mainly contributing to the classification of domain concepts.

Moreover, searching RDF or OWL ontologies existing on the Web that already defined and modeled those terms and relations could enrich such task. Thus is possible, for example, by using Semantic Web search engines (e.g., Sindice [25]), or by querying Linked Data [26] repositories via SPARQL (Simple Protocol and RDF Query Language) endpoints [27] (e.g., Dbpedia [28]), to retrieve useful semantic resources and to reuse concepts and relationships.

C. Some results

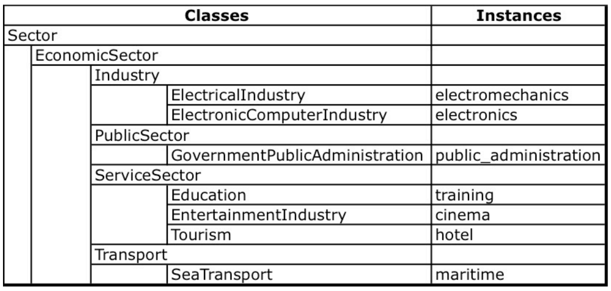
The experiment on the ICT cluster produced a set of candidate terms modeled as concepts and instances in the ontologies. The WiWork Core ontology concepts are mainly generic terms, including: *Person*, *Service*, *Institution*, *Technology*, *Knowledge Occupation*. The Labour Market ontology main concepts include *Economic Sector, Product, Material*, *Tool, Event, Role*. Within the latter ones several entities (i.e., concepts, relationships, and instances) could be defined from the set of candidate terms extracted with the IR techniques. Nevertheless, in many cases authors noted that some terms are very specific and appear in only one occupation description, these terms could be managed as instances of ontology classes.

In Table 1 some candidate terms are classified according to the Labour Market ontology classes.

TABLE I. CANDIDATE TERMS

Economic Sector	Product	Tool	Role	Event
cinema	calculation	airplane	client	congress
electromechanics	design	calculator	journalist	cruise
electronics	document	computer	personnel	exhibition
management	video/sound recording			fair
maritime	filming			manifestation
office management	performance			stage
public administration	planning			
theatre	receipt			
tourism	report			
training	survey			

All those terms derive from the occupation descriptions and belong to the group of specific terms with the highest TF-IDF values (see point 3 in the previous Section B). Ontology classes and relationships have to be refined from the initial classification taxonomy of candidate terms. By taking as main references the ILO and O*Net resources it is possible to specify some ontology parts, e.g., *Sector* class with its sub-classes and instances as represented in Figure 1.



Classes				Instances
Sector				
	EconomicSector			
		Industry		
			ElectricalIndustry	electromechanics
			ElectronicComputerIndustry	electronics
		PublicSector		
			GovernmentPublicAdministration	public_administration
		ServiceSector		
			Education	training
			EntertainmentIndustry	cinema
			Tourism	hotel
		Transport		
			SeaTransport	maritime

Figure 1. Classes and instances of Labour Market ontology

The candidate terms may be then annotated by users in WiWork pages using the property and category labels defined in the ontologies, and transformed in SMW constructs with the support of Protégé-OWL API and Pywikipediabot with an upcoming program described above. Moreover, users could propose new terms or lexical variations to include in the ontological models by exploiting the collaborative wiki functions.

IV. RELATED WORK

Several semantic wikis based on SMW extension exist, in various domains. To the best of the authors knowledge a similar semantic wiki on the labour market domain does not exist. One of the closest wikis to the one introduced in this paper is that of [29], that showed how difficult it is to find the right balance between structured and unstructured data in a semantic portal of a research centre. Another wiki based on ontological models is LinkedLab [30], a Linked Data solution for data management regarding research communities publishing and linking structured data on the Web using RDF.

Important for this paper is also the work of [31] giving an overview of how semantic wikis manage structured, semi-structured and unstructured data. In the field of ontology engineering the use of wikis is a widespread research topic, confirmed by works of [32][6].

Within the enormous literature related to ontology building some methodologies helped authors. In particular, the methodologies characterized by the synergy between various disciplines such as text mining, knowledge acquisition from texts, IR, corpus linguistics or even terminology [16]. In this context, authors found similarities with the research presented in this paper in the work of [33] that uses the TF-IDF to select the best candidate terms for ontologies, with the difference that their work is based on n-gram detection. It is also relevant the work of [34] that uses TF-IDF to identify different types of relationships in a specific step of his ontology building methodology.

V. CONCLUSION AND FUTURE WORK

The paper proposed a semantic wiki, WiWork, to enrich wiki unstructured contents using semantic annotations referring to ontological models. The paper showed also how to model domain ontologies required to structure the labour market domain knowledge managed by the wiki. The semantic wiki showed several advantages: it facilitates the task of structuring the knowledge according to ontologies, it guides users in annotating texts in an easy way and it provides means to process and reuse machine-readable knowledge.

Moreover, a particular methodology was presented based on IR and statistical techniques to support the ontology modeling tasks. The results of an initial experimentation in this direction are showed, and they represent part of a work that authors are pursuing within a multidisciplinary research team.

In the future, the authors plan to include into the WiWork architecture some reasoning and querying features and to experiment it with both structured data (e.g., user-generated annotations encoded in RDF) and unstructured one (e.g., wiki pages about health domain). Moreover, WiWork will be made available on the Web and linked with triple stores and SPARQL endpoints, with the aim to support publishing of semantic documentation for Linked Open Data sets about public sector domains.

REFERENCES

[1] D. Beckett, RDF/XML syntax specification (Revised) - W3C Recommendation 2004.

[2] D. McGuinness and F. van Harmelen, OWL Web Ontology Language - W3C Recommendation 2004.

[3] M. Krötzsch, D. Vrandecic, and M. Volkel, "Semantic Mediawiki," Proc. 5th International Semantic Web Conference (ISWC 06), vol. 4273, Springer, 2006, pp. 935-942.

[4] B. Leuf and W. Cunningham, The WIKI way quick collaboration of the Web, Addison-Wesley, 2001.

[5] S. Schaffert, F. Bry, J. Baumeister, and M. Kiesel, "Semantic wikis", Software, vol. 25, No. 4, 2008, pp. 8-11.

[6] F. Bry, M. Eckert, J. Kotowski, and K. Weiand, "What the user interacts with: reflections on conceptual models for semantic wikis," Proc. 4th Semantic Wiki Workshop (SemWiki 09), Hersonissos, Greece, 2009.

[7] CRISP Research Centre Web page, http://www.crisp-org.it.

[8] M. Martini and M. Mezzanzanica, "The federal observatory of the labour market in Lombardy: models and methods for the costruction of a statistical information system for data analysis," in Information systems for regional labour market monitoring - State of the art and prospectives, C. Larsen, M. Mevius, J. Kipper, and A. Schmid, Eds. Rainer Hampp Verlag, 2009.

[9] ISTAT Web page: http://www.istat.it/it/

[10] ISCO08 Web page: http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm

[11] Pywikipediabot Web page: http://www.mediawiki.org/wiki/Manual:Pywikipediabot

[12] Semantic Forms Web page: http://www.mediawiki.org/wiki/Extension:Semantic_Forms

[13] Protege OWL API Web page: http://protege.stanford.edu/plugins/owl/api/

[14] N. F. Noy and D. L. McGuinness, Ontology development 101: a guide to creating your first ontology, 2001, [Online]: http://protege.stanford.edu/publications/ontology_development/ontology101.html.

[15] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," International Journal of Human-Computer Studies, vol. 43, No. 5-6, 1995, pp. 907-928.

[16] N. Aussenac-Gilles and J. Mothe, "Ontologies as background knowledge to explore document collections," Proc. Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 04), Avignon, France, 2004, pp. 129-142.

[17] J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Norwell: Kluwer Academic Publishers, 1981.

[18] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Computational and applied mathematics, vol. 20, 1987, pp. 53–65.

[19] K. Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, vol. 28, No. 1, 1972, pp. 11–21.

[20] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval, New York: AddisonWesley, 1999.

[21] D. Jurafsky and J. H. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition, Pearson Prentice Hall, 2008.

[22] G. Salton and C. Buckley, "Term weighting approaches in automatic retrieval," Information Processing and Management, vol. 24, No. 5, 1988, pp. 513–523.

[23] ILO - International Labour Organization Web page: http://www.ilo.org.

[24] O*Net – The Occupational Information Network Web page: http://www.onetcenter.org.

[25] Sindice Web page: http://sindice.com/.

[26] T. Berners-Lee, Linked Data - design issues, 2006, [Online]: http://www.w3.org/DesignIssues/LinkedData.html.

[27] E. Prud'hommeaux and A. Seaborne, SPARQL query language for RDF. W3C recommendation, 2008. http://www.w3.org/TR/rdf-sparql-query/

[28] Dbpedia Web page: http://sindice.com/

[29] D. M. Herzig and B. Ell, "Semantic MediaWiki in operation: experiences with building a semantic portal," Proc. The Semantic Web (ISWC 10), vol. 6497, 2010, pp. 114-128.

[30] F. Darari and R. Manurung, "LinkedLab: a Linked Data platform for research communities," Proc. Advanced Computer Science and Information System (ICACSIS), Jakarta, Indonesia, 2011, pp. 253-258.

[31] R. Sint, S. Schaffert, S. Stroka, and R. Ferst, "Combining unstructured, fully structured and semi-structured information in semantic wikis," Proc. 4th Semantic Wiki Workshop (SemWiki 2009), Hersonissos, Greece. 2009.

[32] S. Auer, S. Dietzold, T. Riechert, and T. Riechert, "OntoWiki - a tool for social, semantic collaboration," Proc. 5th International Semantic Web Conference (ISWC 06), Springer, 2006, pp. 736-749.

[33] K. Englmeier, F. Murtagh, and J. Mothe, "Domain ontology: automatically extracting and structuring language community from texts," Proc. IADIS International Conference Applied Computing, Salamanca, Spain, 2007, pp. 59-66.

[34] Y. Rezgui, "Text-based domain ontology building using tf-idf and metric clusters techniques," The Knowledge Engineering Review, vol. 22, No. 4, 2007, pp. 379-403.