Transinformação



REFERÊNCIA

SCHIESSL, Marcelo; BRÄSCHER, Marisa. Ontology lexicalization: relationship between content and meaning in the context of Information Retrieval. **Transinformação**, Campinas, v. 29, n. 1, p. 57-72, jan./abr. 2017. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862017000100057&lng=en&nrm=iso>. Acesso em: 7 fev. 2018. doi: http://dx.doi.org/10.1590/2318-08892017000100006.

# Ontology lexicalization: Relationship between content and meaning in the context of Information Retrieval[1]

# *Lexicalização de ontologias*: *o relacionamento entre conteúdo e significado no contexto da* Recuperação da Informação

Marcelo SCHIESSL[2]
Marisa BRÄSCHER[3]

## Abstract

The proposal presented in this study seeks to properly represent natural language to ontologies and vice-versa. Therefore, the semi-automatic creation of a lexical database in Brazilian Portuguese containing morphological, syntactic, and semantic information that can be read by machines was proposed, allowing the link between structured and unstructured data and its integration into an information retrieval model to improve precision. The results obtained demonstrated that the methodology can be used in the risco financeiro (financial risk) domain in Portuguese for the construction of an ontology and the lexical-semantic database and the proposal of a semantic information retrieval model. In order to evaluate the performance of the proposed model, documents containing the main definitions of the financial risk domain were selected and indexed with and without semantic annotation. To enable the comparison between the approaches, two databases were created based on the texts with the semantic annotations to represent the semantic search. The first one represents the traditional search and the second contained the index built based on the texts with the semantic annotations to represent the semantic search. The evaluation of the proposal was based on recall and precision. The queries submitted to the model showed that the semantic search outperforms the traditional search and validates the methodology used. Although more complex, the procedure proposed can be used in all kinds of domains.

**Keywords**: Information Science. Ontology. Information retrieval. Representation of information. Semantic Web.

## Resumo

*Esta proposta visa representar a linguagem natural na forma adequada às ontologias e vice-versa. Para tanto, propõe-se à criação semiautomática de base de léxicos em português brasileiro, contendo informações morfológicas, sintáticas e semânticas apropriadas para a leitura por máquinas, permitindo vincular dados estruturados e não estruturados, bem como integrar a leitura em modelo de recuperação da informação para aumentar a precisão. Os resultados alcançados demonstram a utilização da metodologia, no domínio de risco financeiro em português, para a elaboração da ontologia, da base léxico-semântica e da proposta do modelo de recuperação da informação semântica. Para avaliar a performance do modelo proposto, foram selecionados documentos contendo as principais definições do domínio de risco financeiro. Esses foram indexados com e sem anotação semântica. Para possibilitar a comparação entre as abordagens, foram criadas duas bases, a primeira representando a busca tradicional, e a segunda contendo o índice construído, a partir dos textos com as anotações semânticas para representar a busca semântica. A avaliação da proposta é baseada na revocação*

[1] Article based on the doctoral dissertation of SCHIESSL, M. entitled "*Lexicalização de Ontologias: o relacionamento entre conteúdo e significado no contexto da Recuperação da Informação*". Universidade de Brasília, 2015.

[2] Universidade de Brasília. Faculdade de Ciência da Informação. Programa de Pós-Graduação em Ciência da Informação. *Campus* Universitário Darcy Ribeiro, Edifício da Biblioteca Central, 70 910-900, Brasília, DF, Brasil. *Correspondência para/*Correspondence to: M. SCHIESSL. *E-mail*: <marcelo.schiessl@gmail.com>.

[3] Universidade Federal de Santa Catarina, Departamento de Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação, Florianópolis, SC, Brasil.

Received on 19/10/2015, resubmitted on 23/6/2016 and appoved in 7/7/2016.

*e na precisão. As consultas submetidas ao modelo mostram que a busca semântica supera o desempenho da tradicional e validam a metodologia empregada. O procedimento, embora adicione complexidade em sua elaboração, pode ser reproduzido em qualquer outro domínio.*

***Palavras-chave***: *Ciência da informação. Ontologia. Recuperação da informação. Representação da informação. Web semântica.*

## Introduction

The Web revolution has led to widespread access to information. Another revolution, still in progress, is the Semantic Web revolution, which is based on the principle that electronic information will not be ambiguous, data will be readily available, reusable, and interoperable, and the devices will be ubiquitous. The idea is to bring Web ubiquity to the everyday lives of users with documents enriched with semantic information of Web pages, thus creating an environment in which agents, in the form of computer software programs, can surf through the Internet, collect information, and perform complex tasks on behalf of users.

Thus, even if complex systems, such as ontologies, ambitiously aim for semantic information processing, current technologies are restricted to the ability of computers to run only syntactic processing, *i.e.*, search for patterns. In this case, the initial and unique proposal of ontologies to interact with both man and machine can be affected and human involvement in the preparation, organization, and content indexing cannot be waived.

Despite the semantic web promise to establish a relationship between people and machines, Wilks and Brewster (2009) argue that knowledge representation – ontological in this case – must be combined with any natural language to be justified. The authors add that a language is a system of rare events, but it is a complete model. Therefore, they quote Spärck Jones, who claimed that the words are self-representing and no other symbol can substitute or codify them with similar meaning. Charniak (1973) and Wilks (1977) corroborate this statement in different ways, but they state that words retain essential information that is not present in any other representation.

It is evident that the world of semantic web and natural language need to be connected. In order to use knowledge, it is necessary to create a bridge between the components of an ontology – classes, properties, and individuals – and their correspondents in natural language. Therefore, to capture linguistically rich information about verbalizations of simple and complex elements of an ontology, lexical knowledge is needed, that is, knowledge of the set of words related to the domain of interest. Furthermore, this knowledge should be made accessible in machines and should be published to facilitate its reuse. The effective bridge between these two worlds would allow queries submitted in natural language to seek semantics available in the semantic web and provide alternatives to address a central problem of interest in Information Science: Ambiguity.

It seems natural that the exploitation of resources and technology is the way to create the balance between lexical elements present in the documents and ontologies that are at the level of knowledge representation. Therefore, the proposal presented in this study aims to properly represent the natural language to ontologies and vice versa. The semi-automatic creation of a lexical database in Brazilian Portuguese containing morphological, syntactic, and semantic information that can be read by machines was proposed, allowing the link between structured and unstructured data and its integration into an information retrieval model to improve precision. The inclusion of language resources in a natural language processing system can provide better interaction with the user and improve the quality of information retrieval systems.

## Methodological procedures

The use of a natural language interface with any language processing and information retrieval systems allows direct interaction and therefore allows raising questions that accurately reflect the user's needs. However, this influences the complexity and the characteristic of the representation of document contents associated with the quality of information recovery (BRÄSCHER, 1999).



 https://doi.org/10.1590/2318-08892017000100006

## Semantic Web

According to Guarino, Oberle and Staab (2009), in the context of Semantic Web, semantics conveys meaning. This enables more effective use of underlying data because the human reader has to interpret the gaps and relationships present in the texts. The available sources usually have only keywords visible in search engines, which can be seen as a limited semantics. However, if the keywords are related to other defined links, the context is formed revealing the semantics. For example, the word bank alone is ambiguous, but if it is combined with other words, such as *agência*, *caixa eletrônico, saque*, and *depósito* (agency, Automatic Teller Machinen, withdrawal, and deposit), it falls within the context of financial institution and reveals its semantics.

Guarino (1998) argues that ontologies capture knowledge but fail to capture the structure and use of terms that are objects of Terminology and Lexicology. The structure and use of terms are essential to express and refer to the same knowledge in natural language. Paradoxically, researchers have given less attention to issues related to the lexicon and linguistics in the fields of knowledge organization and information retrieval. Therefore, the solution of this problem requires a formal knowledge representation model that encompasses the semantics of ontology, the terminology used to express this knowledge in natural language, and linguistic information about the terms and their lexical units. This model allows the participation of machines in the translation and inference process. Therefore, in addition to semantic and terminological levels, representing the lexical level is also necessary for proper use of ontologies in language processing and as a way to integrate the terminological and ontological levels.

In human communication, people use contextual knowledge, world knowledge, and personal experiences to facilitate utterance interpretation. On the other hand, the communication between machines is established using artificial and standardized methods developed for this purpose. The Web is based on the HyperText Markup Language (HTML), which cannot explain the real meaning of information. Consequently, the machines deal only with syntax in order for information to be exchanged between them, but they cannot understand the meaning of these messages.

If knowledge becomes explicit through Web technologies, Semantic Web is created:

> […] is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation (BERNERS-LEE; HENDLER; LASSILA, 2001, p.3).

The major goal of the Semantic Web is to help machines to "read" and use the web. According to Berners-Lee *et al*. (1998), this will transform the current Web, a giant global book, into a giant global database. Such technology does not provide intelligence or transform machines into conscious human beings, but it provides tools for them to find, exchange, and interpret information.

All information added to the Web should be named to be identifiable and retrievable. This can be done using the Uniform Resource Identifier, which refers to a string of characters used to identify resources that are built on standards. According to Heath and Bizer (2011), the uniform resource identifier provides a simple and extensible means for identifying a resource. Furthermore, it is intended to distinguish and identify anything that can be represented via URI, such as texts, images, videos, sounds, and concrete (car, moon) or abstract (love, divinity) concepts.

The uniform resource identifier concept is widespread in the Information Science such as in the specification the location of Web pages via Uniform Resource Locator (URL) and in the identification of books via International Standard Book Number (ISBN), serial publications via International Standard Serial Number (ISSN), and digital contents via Digital Object Identifier (DOI). All of them correspond to the standard mechanism that identifies and individualizes objects.

Resource Description Framework (RDF), as its name suggests, provides a framework for describing resources using a simple mechanism to express facts or statements. The idea behind the RDF is clear, the whole concept is represented by the triple: subject, property (or predicate), and object. In fact, this combination is familiar to all speakers of Western languages because it is the intuitive way to form simple sentences. Subject

refers to the concept to be described; property refers to the attributes related to the subject; and object refers to property. Anything can be described using this simple triple.

Allemang and Hendler (2008) state that the Resource Description Framework Schema (RDFS) is a language that defines the vocabulary to be used in RDF. It allows the definition of classes of entities that have something in common. Moreover, it enables defining properties and their restrictions, as well as the hierarchy of classes (subclasses and superclasses) and properties (subproperties and superproperties).

According to Nardi and Brachman (2003), RDFS and RDF combine two types of knowledge: 1) intensional knowledge (general), which remains at the conceptual abstract level and deals with the actual data model, *i.e.* the relationship between general entities, such as classes and properties; and 2) extensional (specific) knowledge that deals with the specification of the entities or class instantiation. As a result, the relationships between entities in the specialization layers are reflected in generalization layer which forms a RDF (S) knowledge base.

Figure 1 shows the representation of specialization and generalization layers. The first one is commonly referred to as ABox or Assertional Box. For example, the representation of the sentence "*Heitor Villa-Lobos was born in Rio de Janeiro*" in Japanese would use the particularization, instantiation, or specification of the class 'person'. The second is called TBox or Terminological Box, which contains the domain abstractions that enable inferences about the data model. Thus, in this layer, the relationship between classes and properties introduces the semantics in the data model, which leads to an ontology and translates into the computer world the ideas of Dahlberg (1978) about extension and intension of the concepts that are the basis of ontologies in Information Science.

The semantics of the elements of the RDF (S) knowledge is based on their properties and values, *i.e.*, it is possible to make inferences about the hierarchical relationships between classes and properties and based on restrictions connected to the properties, such as domain and range. Therefore, RDF and RDFS provide sufficient semantics to represent knowledge, although at a superficial level only. With these Semantic Web tools, information systems can go a long way with a little semantics.

However, the difficulty in predicting relationships involving conflict or incompatibility still remains. For example, disjoint classes: if we consider the classes Man and Woman, we know that no individual can be an instance of both classes. This means that, in RDFS, it is impossible to determine whether there are inconsistencies. On the other hand, the Open World (OWA) assumption is the view that what is stated in the database is what is known; everything else is unknown. Similarly, there is no assumption of single names, *i.e.*, it should be explicitly expressed that person A is not person B. Finally, there should be a comprehensive specification of entities and relationships unless they add inference rules in a more abstract layer that can set limits and introduce generalized restrictions to the database.

Web Ontology Language (OWL) was designed for more complex class structures and properties. It extends RDF and RDFS and adds more vocabulary for describing groups of things, such as classes, facts about these classes, relationships between classes and instances, and characteristics of these relationships. It is focused on the processing of the Web content and is intended to be read by computer applications. Moreover, it enables the creation of rules, axioms, and inferences to enable deductions using logical tools (W3C, 2014).

**Information retrieval**

There have been undeniable advances in information retrieval in recent years due to the Web, the popularization of Graphical Use Interfaces (GUI), and inexpensive mass storage devices. In addition, the continuous optimization of search engines, which improves users' experience, has made the Web the standard and preferred source of information, especially after the launch of the Google search engine by Brin and Page (1998), which tries to respond to the challenges of designing a system that gathers Web documents and keeps them updated, according to the rate of growth of the Web.

Baeza-Yates and Ribeiro-Neto (1999) propose the distinction between the user task and the logical view of the document. Both directly influence the effectiveness

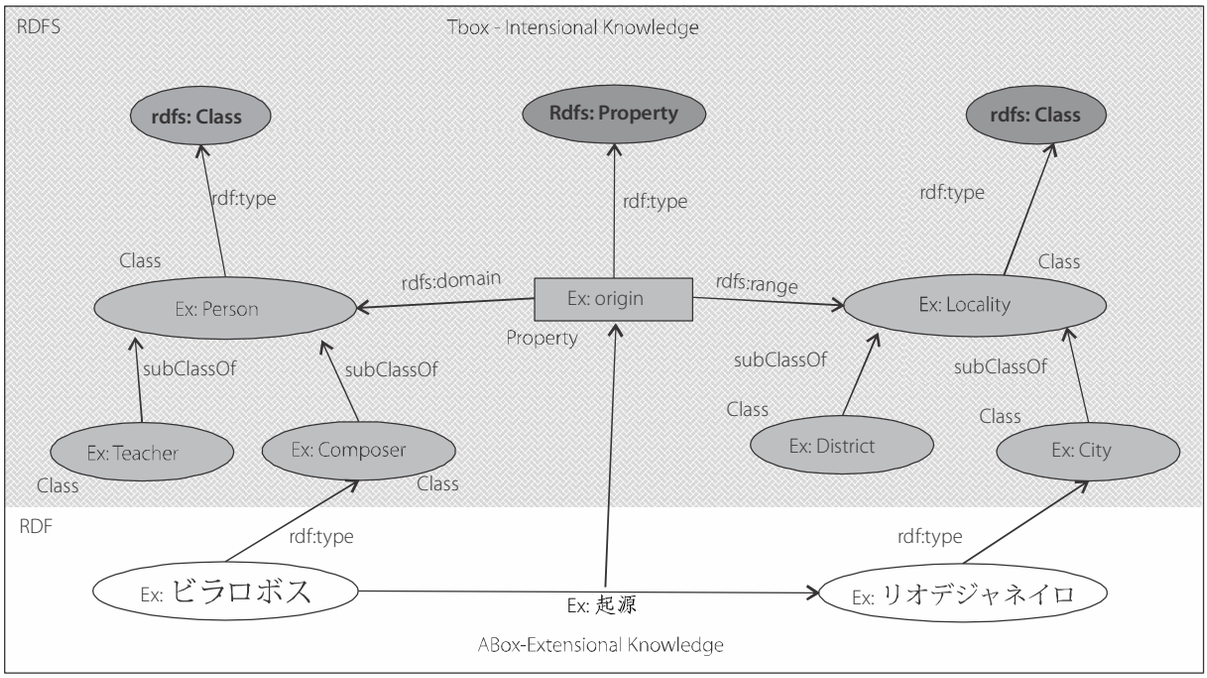


**Figure 1**. Intensional x Extensional Knowledge.
Source: Created by the Author.

of the Information Retrieval System (IRS). User task implies specifying terms which convey the semantics of the user need and that meet the user information needs when browsing retrieved documents. Logical view of the document refers to a sequence of transformations aimed at representing documents through a set of index terms or keywords, which is justified because although full texts are the most complete logical view of a document, their usage implies high computational cost. On the other hand, a small set of categories provides the most concise logical view of a document, but its usage leads to poor quality retrieval.

From the traditional information retrieval to the Web today, there has been a significant change in the Web user profile. Professionals trained to perform queries on well-structured and well-known collections have been replaced with ordinary people who tend to ignore or disregard the heterogeneity of the contents, query languages, or any conceptual foundation about Information Systems. This has led to increased complexity in the infrastructure involved in the entire information management process.

## Semantic search

Activities involving the Semantic Web have been widely studied and many proposals have been made in an attempt to create a Web of distributable, machine-readable data. Since the concept of semantic web has been introduced, many problems have been solved but more complex ones are still approached differently by different researchers that contribute to a more generalized view of semantic web, which is discussed below.

Semantic portals discussed by Maedche *et al*. (2001), Castells *et al*. (2004a; 2004b) and Contreras *et al*. (2004) essentially provide simple search functionalities that are characterized as semantic data retrieval. Searches return ontology instances rather than documents and no relevance ranking is provided. In some systems, links to documents that reference the instances are added in the user interface next to each returned instance in the query answer according to Contreras *et al*. (2004), but neither the instances nor the documents are ranked.

The relevance ranking issue was addressed by Rocha *et al*. (2004), who suggested a solution that

provides a ranked list in response to user queries. The authors proposed a semantic network in which the relation instances have semantic labels and numerical weights. The query terms are mapped to the semantic network nodes, and the order of the search results is determined according to the relevance provided by the associated weights.

Guha and McCool (2003) and Guha *et al*. (2003) assumed that semantic web data are modeled as a directed and labeled graph, in which each node corresponds to a resource and each arc is labeled with a property type like a RDFS data model.

Popov *et al*. (2004) believe that the combination of information retrieval techniques, semantically lightweight ontologies, knowledge representation, and information retrieval can address the problem in annotation and automatic semantic retrieval.

The study by Vallet *et al*. (2005) and Castells *et al*. (2007) complements the studies carried out by Guha and McCool (2003), Guha *et al*. (2003) and Popov *et al*. (2004) by introducing a ranking algorithm especially designed for an ontology-based retrieval model using a semantic indexing scheme based on annotation weighting techniques.

Seeking to overcome the limitations of specific organizational ontologies, Fernandez *et al*. (2008) investigated the combination and the range of information spaces provided by semantic web and WWW. Their study represents an important step towards the design of semantic retrieval technologies to the open Web by: (1) bridging the gap between the users and semantic data and (2) bridging the gap between the semantic web data and unstructured textual information available on the Web.

Exploring the Linked Open Data (LOD) potential, Hogan *et al*. (2011) proposed a Semantic Web Search Engine (SWSE) for searching and browsing RDF Web data. Given the flexibility of the semantic web, retrieved objects can represent people, companies, cities, proteins, or anything that has been published without predefined categorization such as that in traditional search engines. Moreover, this system must scale to large amounts of data and must be robust enough to deal with heterogeneity, noise, unreliability, and possible conflicts of data collected from a large number of sources.

The exploitation of metadata associated with semantic web documents can increase the precision of information retrieval systems. Silva *et al*. (2009) introduced a generic information retrieval model for the semantic web using metadata in all stages of the process: representation, matching, and similarity measure. The model uses semantic representation rather than keywords. The documents are described through concepts and instances clustered in "semantic cases" that represent the user interest. In order to achieve more precise results, the matching and similarity models compare the same "semantic cases" of queries and documents.

In an attempt to interconnect semantic web with WWW, various processes have been proposed, especially lately. Despite the growth of structured databases to levels that enable various searches, Heath and Bizer (2011) mention that the gap between text and structured data remains a barrier to the popularization of semantic web and to the use of tools designed for this environment.

Some initiatives such as those introduced in the studies by Navigli *et al*. (2003) and Reymonet *et al*. (2007) include ontology lexicalization models without integrating the lexical and ontological levels. The model proposed by Buitelaar *et al*. (2011) and improved by McCrae *et al*. (2012), Unger *et al*. (2013) and Cimiano *et al*. (2014) reflects the urgent need to establish a connection between the knowledge of the world of concepts and the world of terms, accurately describing the difference between them.

Given the large volume of Web content, it is impossible to develop solutions without the help of machines. Therefore, in order to automate the lexicon construction, Walter *et al*. (2013) and Walter *et al*. (2014) used structured databases to provide the semantics and the corpus to find lexical and morphological variants. The aim is to induce the creation of a lexicon from the knowledge represented in ontologies to feed the originally proposed model.

Finally, the integration between semantic web and WWW will make it possible to obtain appropriate structured and unstructured information about the user

profile. The new generation of IRS will be able to indistinctly search either in databases of formal knowledge containing ontological structures that are not understandable to people or in textual databases that are not understandable to intelligent computer programs and provide good quality results. Thus, only with this free and unrestricted communication between the two worlds, the claimed potential of semantic web will be available to the common user.

## Semantic information retrieval model proposal

In the present study, we propose the semi-automatic construction of a lexical database in Portuguese for the Financial Risk domain, which, for the purpose of this study, was called RiscoLex. This database was created based on ontology of risk and its corresponding corpus, as described below.

Figure 2 shows the top level view of the financial risk domain. The various concepts are linked by relationships that contradict the forces between the threat and protection of the assets of an entity. Each dimension of this diagram gives rise to increasingly specific concepts. The set of concepts must be interpreted following the arrows that establish the type of relationship between one concept and another. For example, IE/LE (IE- Individual Entity; LE- Legal Entity) is a defense agent that imposes defense measures to mitigate the asset risk.

The collection of texts about financial risk contained 2,978 documents in Portuguese, which are in various formats known by most users. The formats are: Microsoft Word (.doc and .docx) and PowerPoint (.ppt), Portable Document Format (.pdf), and HyperText Markup Language (HTML). In addition, Wikipedia in Portuguese was also used containing 1,385,451 documents in

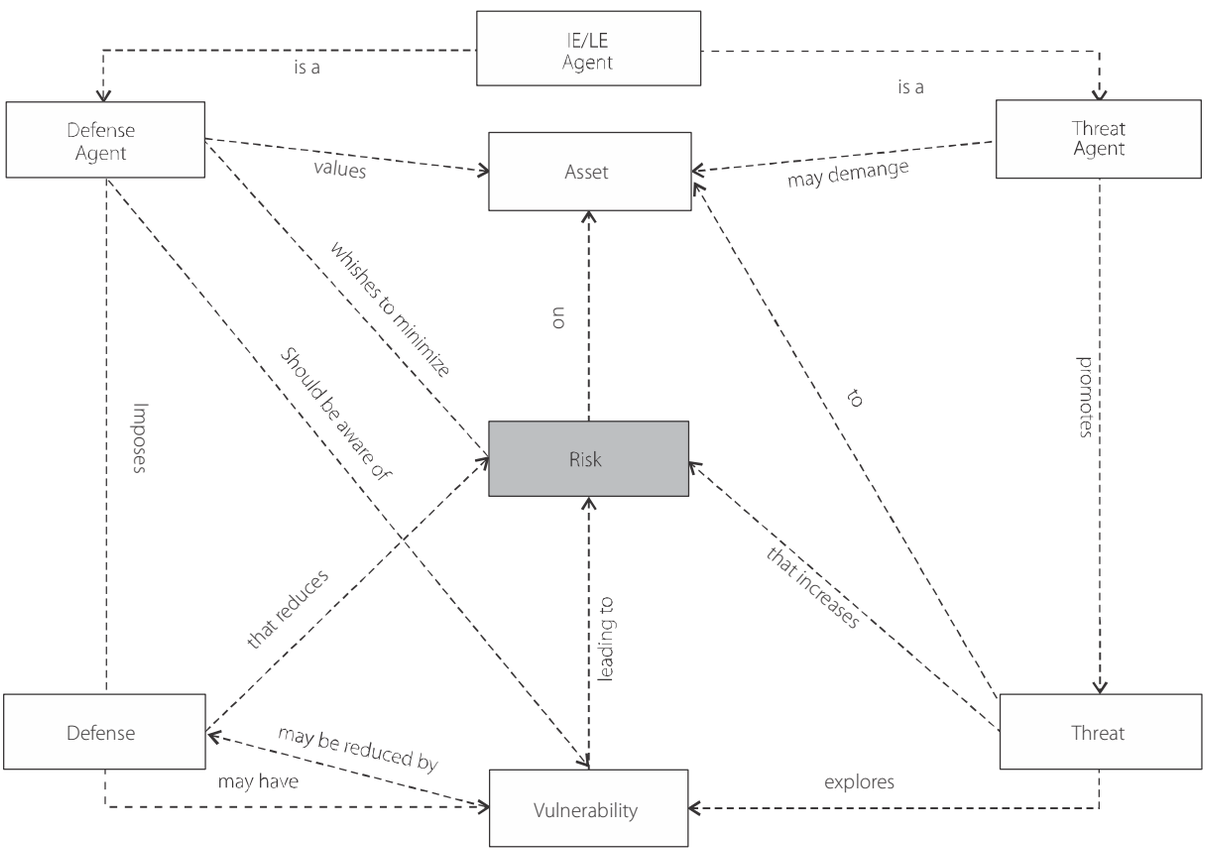


**Figure 2**. Top level view of the Financial Risk Domain.
Source: Adapted from Gresser *et al.* (2010, p. 12).

eXtensible Markup Language (XML) format. The reason was the ability to find different types of lexicalizations or ontology properties that enable better generalization of standards. For this search, we used the corpus *Floresta* that is incorporated into the Natural Language tool Toolkit (NLTK). There were 9,266 phrases corresponding to "*Floresta Sintá(c)tica Corpus*", version 7.4, *Bosque* part. The following computational resources were used for processing: Protégé <http://protege.stanford.edu/>, version 4.3, Python 2.7 programming language with the library (NLTK), by Bird, Klein, and Loper (2009); SciKit-Learn <http://scikit-learn.org/stable/>, by Pedregosa *et al.* (2011), RDFlib <https://rdflib.readthedocs.org/en/latest/>, and the applications Apache Jena Fuseki <http://jena.apache.org/index.html>, version 1.0.0, and Solr <http://lucene.apache.org/solr/>, version 4.6.0.

**Lexicalization Approach**

In order to represent the linguistic information, the principles defined by McCrae *et al*. (2011) for the proposal of the Lexicon Model for Ontologies (lemon) were applied. This model was designed to develop a standard RDF format of linguistic information, which includes declarative specifications of a machine readable lexicon that captures morphological, syntactic, and semantic aspects of the lexical items related to an ontology.

Semantic similarity was determined using the following lexical resources that are structured in groups of semantically related lexical items and that can be used freely because they are in the public domain: Priceton WordNet, proposed by Fellbaum (1998); Open Multilingual Wordnet (OMW) proposed by Paiva *et al*. (2012), which resulted in the OpenWN-PT; PWN proposed by Bond and Foster (2013); Onto.PT proposed by Oliveira (2013); and DBnary, proposed by Seìrasset (2014). These resources combined are the key sources for the selection of semantically related lexicons in the domain of interest, financial risk, in the present study.

The proposal for the construction of RiscoLex is to extract the labels of classes and properties of the ontology, identify and retrieve their respective synonyms and the morphosyntactic features of each term, convert them into RDF format, and provide the lexical database with the Lemon model. Figure 3 shows the steps of the generation RiscoLex process.

The approach includes the proposal of one or more lexical entries for each class and property of the ontology. The first step involves the extraction of the labels of the ontology and additional information such as synonyms and syntactical features, from external resources. The task steps were configured to do the following: (1) All s and p labels are extracted from the ontology triple (s, p, o) to create a list of terms in natural language; (2) Labels in CamelCase (*nascimentoLocal*), hyphenated words (*presidentes-do-Brazil*) or separated by underscore (*instituições_financeiras*) must be represented in NL found in texts. This step aims to transform formats such as *paísDeOrigem* into *país de origem* or *gerenciamento_de_risco em gerenciamento de risco*; (3) These terms are searched in the corpora for validation. This step aims to characterize frequent terms which are, therefore, preferred in the domain and in the Portuguese language; (4) in natural language, it is common to use more than one word to convey the same meaning. Thus, the aim is to find the greatest possible number of synonyms for the terms of the list. Linguistic ontologies for the Portuguese language were used in this task; and (5) The Lesk (1986) approach was used to treat polysemous terms and collect those that are more relevant to the domain.

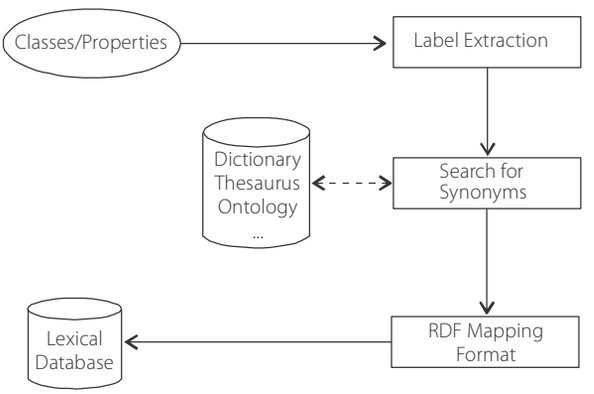


**Figure 3**. Riscolex construction flowchart.
Source: Created by the author.


https://doi.org/10.1590/2318-08892017000100006

### Riscolex and information retrieval

Traditional IRS rely on keywords or descriptors to index documents, but this is not enough. The problem is that if the query term does not match with the keywords, the document will not be retrieved. For example, the query term is *perigo* (danger), in the representation of the document the synonym is *risco* (risk); no mechanism based on measuring similarity between terms will retrieve that document.

Therefore, a corpus and an ontology represent the same domain to different users: machines and people. In general, there is no correspondence between the labels available in ontological entities and the document descriptors. For example, the label of the class *Pessoa Física* would be expressed in the descriptors as *Pessoa Física* (Individual Entity). In this case, the RiscoLex, linked to the ontology, provides the lemma and the synonyms to the descriptors. If the descriptor is not inserted in the RiscoLex or in the ontology, it can be semi-automatically inserted in both of them to emphasize the dynamic nature of knowledge.

Moreover, the extent to the comprehensiveness of the indexation system can be increased through inferences that explicitly provide semantic meanings. The inclusion of ontology to support the IRS provides more meanings through inferential engines. In this case, it can be seen as a dynamic extension of the document descriptors. For example, from the class *Especialista* (Specialist) it can be inferred that the members also belong to the Stakeholder and *pessoaFísica* classes. Therefore, they inherit all of their attributes and restrictions through axioms, without being explicitly expressed. The automatic hypernym resolution, such as *banco* (bank) and *instituiçãofinanceira* (financial institution) and other forms of dependence between words, increase precision in the representation of a given document.

### SIRM: An overview

In the Semantic Information Retrieval Model (SIRM), it is assumed that the ontology was constructed and associated with the textual information sources that include the concepts to be represented. In addition, it is also assumed that although this search is restricted to the financial risk domain, the model can be applied to any other domain since there is structured and unstructured information that could represent the concepts understood by the domain.

The domain is represented by ontologies and corpora. On the one hand, ontological entities represent concepts, and inference engines automatically infer non-explicit information. On the other hand, descriptors describe document contents, and people interact using natural language to infer unexpressed meanings. They complement each other for the task of providing information but in different formats or even incompatible formats. Influenced by Fernández *et al*. (2011) and Kara *et al*. (2012), the model has the information retrieval structure based on the descriptors including a semantic module. Documents and ontological entities are indexed together. This modeling option facilitates the interaction with the end user as it keeps searching in the same way it does in traditional search engines. Additionally, the final result is at least as good as that of the traditional approach; *i.e*., if the query does not find relevant correspondents in the knowledge base, the system retrieves the information related to the document descriptors.

Figure 4 illustrates the information retrieval process with the addition of the semantic module. The user interacts in a traditional way to submit the query. The query processing standardizes the terms for the search. The lexicon-ontological knowledge includes the ontology and the RiscoLex. The corpus characterizes the database containing the documents to be retrieved. The joint indexation of the databases involved provides the lexical-semantic index, which is used in the retrieval and ranking of retrieved documents to be presented to the user.

The semantic annotation process is therefore essential to link documents to the semantic space created by the domain ontology. NLP is the main tool for document identification, comparison, and annotation. However, seeking to minimize possible ambiguity effects, it is complemented by human validation.

## Results

The first result to be highlighted is the creation of RiscoLex, the first lexical database in Brazilian Portuguese

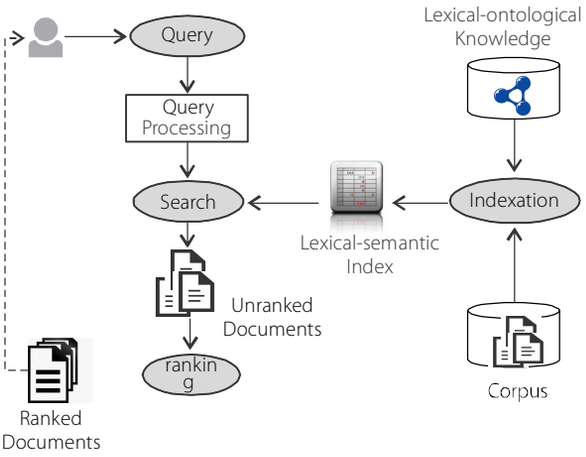


**Figura 4**. SIRM overview.
Source: Adapted from Fernández *et al.* (2011, p.438).

built with the Lemon model, which differs from the others by the interpretation of language restricted to the well-defined domain. Additionally, the ontology, as a resource for natural language interpretation, puts the lexical database at the center of the interpretation process. In this line, the level of representational granularity, at which the meaning of natural language is captured, is not driven by language but by the semantic distinctions made in an ontology. Thus, these distinctions are relevant only in the context of a specific domain.

Another result is the construction of the first ontology for risk management in Portuguese. Difficulties in building this type of resource have often been reported in the academic literature. In our study, it was not different. Although different resources were used as the starting point, the specificity of the topic demanded the construction of this ontology as if it were new. The diversity between the international and national financial markets has led us to rethink the concepts and their relationships according to the Brazilian market, especially for public companies. This adaptation required great effort to represent such knowledge.

Thus, the final ontology in the domain of *Risco Financeiro e Corporativo* (Finance and Corporate Risk) – *OntoRisco* – in Portuguese was developed based on the combination of existing ontologies in English adapted to the Brazilian needs. This resulted in 2,178 triples comprising the subject, predicate, and object; 65 classes – or concepts – and 47 properties – or relationships between concepts. In addition, 476 axioms were created for the inference engine that enables logical deductions from existing information and knowledge discovery about the domain.

In the present study, the results of the validation of the ontology labels with the corpus were below expectations. Only 50.7% of subjects or classes were found in the corpus and only 20,0% in the properties. These low results reflect the poor choice of terms to be included in the labels of the classes and properties. This process indicates that the selection of synonymous should improve the representation of knowledge in relation to the written material available. The participation of experts or specialists in this field is essential for choosing the most appropriate terms.

The clustering procedure generated three groups. The group chosen by the specialists was the one with the largest number of words related to *Risco Financeiro* (financial risk). Then, the most representative term was used as the source for the creation of the bag of words (BoW) of the risk. The group went through several processing steps until finding the appropriate BoW for the comparison with the synonyms. In each processing step, it was observed a reduction of the number of terms that would be part of the BoW, resulting in a 53% reduction, *i.e.*, from 90,533 terms to 42,394 terms in the group.

Therefore, the labels of 65 classes and 47 properties, *i.e.*, 112 entries, were searched to find lexical variations or synonyms in the dictionaries and lexical ontologies to compose the RiscoLex. A total of 122 new terms were found and validated. Thus, the final version was increased by 109%, totaling 234 terms to compose the RiscoLex.

On the one hand, the ontology provides the labels to start the search for synonyms in the support databases. On the other hand, the corpus is segmented and the group that contains the terms that best represent the domain is transformed into a BoW. Therefore, the synonyms and terms in the BoW are validated using the similarity measure. The terms with similarity equal to 1, *i.e.*, identical terms, were automatically added to the RiscoLex. For the terms that were not found in the BoW, a manual analysis was carried out to verify whether they are in fact related to the domain; if so, they were inserted into the RiscoLex.

## Discussion

According to the objectives of our investigation, it was considered that there was no need for a vast number of documents, but rather a set of data that enabled examining the advantages and disadvantages of the methodology used and the feasibility of investigating the entire corpus manually to verify and validate the results of the automatic procedures. Consequently, a total of 785 documents containing the main definitions of risk domain were selected. These documents were indexed with and without semantic annotation to enable comparison between the approaches.

The terms identified in the corpus that corresponded to the ontology labels were assigned weight to increase the relevance of the document and make it more visible to the search engine. In the platform Solr, term weighting is based on the tf-idf algorithm, which presents an ordered list with detailed information of the scores assigned to each retrieved document to rank the relevance.

### Retrieving information

In the search for the same term, for example the term *'ameaça'*, there is the semantic space provided by the RiscoLex, which also means searching for the terms *risco, perigo,* and *ameaça*. As previously explained in section 3.3.1, the semantic similarity between these terms was obtained from the lexical resources used in search and that are related to the financial risk domain. Thus, the syntactic search looks for explicit terms only, whereas the semantic search looks for any type of term. For example, in the syntactic search, the term *perigo* retrieves only one document, but in the semantic search, it retrieves 159 documents.

The term *risco* (risk), the most frequent term, was present in most documents, and its variants were used in only three documents. In the case of a syntactic search for the term *ameaça* (threat), there would be absence of 99% of semantically related texts. Therefore, the semantic result is the set of texts containing any semantically related term present in the RiscoLex database. The first benefit of this technique is the recall increase.

As highlighted in the literature, recall and precision are inversely correlated, and therefore a balance should be sought to achieve maximum recall and precision. For instance, an ambiguous common behavior is observed for terms that have different syntactic functions, according to their use. For example, let's take the noun '*bem*', which in the RiscoLex has the same concept of *propriedade* (property), *posse* (ownership), *ativo* (asset), and *recurso* (resource), that is, something that is owned or possessed. This shows that the syntactic differentiation of terms helps removing the ambiguity caused by policategorization and improving precision.

A third procedure to deal with ambiguity by homography was also used. It refers to the semantic identification of terms that have the same syntactic category but different meanings. For example, the search for the term '*produto*' (product) which is synonym for '*artigo*' (article). This term is also common in the risk domain, but it usually refers to a part of law or legal agreement that deals with a particular point.

In addition, in terms of semantic similarity, the procedure refers to the identification of terms with related meanings aiming at measuring their semantic similarity. Given a particular term, the entire collection can be

**Table 1**. Results' evaluation.

| Query | ts | RRD | P% | R% | F% | ss | RRD | P% | R% | F% | RDD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P-1 | 2 | 2 | 100.00 | 1.26 | 2.48 | 159 | 159 | 100.00 | 100.00 | 100.00 | 159 |
| P-2 | 175 | 14 | 8.00 | 93.33 | 14.74 | 15 | 15 | 100.00 | 100.00 | 100.00 | 15 |
| P-3 | 720 | 18 | 2.50 | 100.00 | 4.88 | 19 | 18 | 94.74 | 100.00 | 97.30 | 18 |
| P-4 | 30 | 10 | 33.33 | 6.54 | 10.93 | 154 | 153 | 99.35 | 100.00 | 99.67 | 153 |
| P-5 | 2 | 2 | 100.00 | 25.00 | 40.00 | 7 | 7 | 100.00 | 87.50 | 93.33 | 8 |

Notes: Relevant Retrieved Documents (RRD); Relevant Documents in the Database(RDD).
Source: Created by the author.

scrolled trough to identify others that have the same semantic category. Identifying semantically related terms is very useful in indexing a corpus so that a search for a broad meaning such as '*mercadoria*' (goods) also retrieves documents with specific terms such as '*artigo*' (article).

Finally, it is known that human supervision increases the annotation accuracy. However, the task is not feasible for several million annotations that can be obtained in textual databases. The automatic annotation processing described can present a list of terms to be investigated by domain experts and about which there is some uncertainty regarding annotation. This list is a debugging tool to identify polysemy cases, semantic annotations which do not correspond to the concept or to the syntactic category, and the absence of important terms for the domain in the knowledge base.

## Evaluation

To evaluate the performance of our proposal, two databases with the same documents were indexed. The first one, represents the traditional search, *i.e.*, an index built based on unprocessed texts, and it was used as a starting point for comparison. The second contained the index built based on texts with the semantic annotations, and it was used to represent the semantic search.

The evaluation was based on recall and precision. Therefore, the relevance of each document should be determined according to the user's interest, in this case, the query made into the system. Since these databases have not been previously classified, it is necessary to evaluate the references relevant to the topic of the query. Therefore, in order to determine the relevance, the databases were evaluated by 5 experts, who also assessed the documents that were not retrieved, according to the following queries:

– P-1 Documents related to *ameaça* (threat) (query: *ameaça*).

– P-2 Documents related to *risco operacional* (operational risk) (query: *risco operacional*).

– P-3 documents related to *risco de crédito* (credit risk) (query: *risco de crédito*).

– P-4 documents related to *bens* (goods) (query: *bens*).

– P-5 documents related to *crime* (crime) (query: *crime*).

All queries have characteristics that may influence the results obtained by search engines. From the language processing point of view, the complexity increases demonstrating an improvement with the use of a semantic information retrieval system. The results, the linguistic complexities that affect the performance of traditional search engines, and the way the proposed approach dealt with them are discussed below.

For each query, Table 1, shows: the number of documents retrieved by the traditional search for documents considered relevant; the number of documents retrieved by the semantic search; the values of precision, recall, and measurement; and the number of documents in the database that were considered relevant by the experts (last column).

There was a considerable difference between the traditional search and semantic search in the P-1 query due to the preference for the term *risco* (risk) in the documents that compose the database. Which was therefore reflected in the low recall, only 1.26 documents. The traditional index is not able to retrieve the term '*risco*' (risk) because it is not explicit in the query. However, the semantic search can recognize other terms that are present in the RiscoLex, and thus they were added to the index. '*Ameaça*' (threat), '*risco*'(risk), and '*perigo*' (danger)were therefore indexed and considered as access points.

The P-2 query showed good recall but low precision in traditional search, *i.e.*, many documents were retrieved but most were irrelevant. This is due to the lack of identification of compound terms, resulting in the search for the isolated terms *risco* (risk) or *operacional* (operational). Furthermore, there is no processing for plural terms. Thus, a document containing '*riscos operacionais*' (operational risks) was also not retrieved in the semantic search either. On the other hand, the ss finds a single term '*risco operacional*' (operational risk) that leads to the accurate retrieval of all documents containing the compound term, including the one with its plural form.

The P-3 query indicated the importance of processing stop words and compound terms, as well as the proper identification of diacritical marks characteristic

of the Portuguese language. Traditional searches usually eliminate the accent marks and therefore, the words '*crédito*' (credit) and '*credito*' (to credit) (verb '*creditar*', conjugated in the first person of present indicative), without the acute accent, will have the same form. Both words are found in the domain investigated, and thus the recognition and change to the canonical form before annotation is convenient. Therefore, high recall and low precision in traditional search indicate the lack of processing of the mentioned topics. In the semantic search, high recall and precision values were found, as expected.

In the P-3 query, there was one extra document in the ts, making automatic processing more difficult. This reference is included in the document as follows:

*... além dos e de mercado, introduziu-se o risco operacional...*

With the normalization of the plural form, the document was retrieved, but it was considered irrelevant by the experts because it is a text in which the term appears only in one list of several risks in the context of *risco operacional* (operational risk). This is a typical case in which only human judgment can determine the relevance of the term, and there is no way to treat it automatically.

The P-4 query showed an improvement in the results considering the syntactic categories as a way of defining the meaning of the terms. The traditional search retrieved documents that have the term '*bem*' (goods), but it did not include its plural form '*bens*' (goods), resulting in 16 references in the database. It is important to mention that out of the 30 documents retrieved by the ts, 20 had the comparative phrase '*bem como*' (as well as), and therefore, with the exception of 3 references that besides this comparative phrase also contained other relevant terms and terms with the same meaning, they should not be retrieved. In the semantic search, 154 documents that also included the synonyms *ativo* (asset), *propriedade* (property), *posse* (ownership), and *recurso* (resource) were retrieved. Thus, there was one extra document retrieved in the ss, which indicated a mistaken annotation, which was observed due to this result. Moreover, of the 20 references to the term '*bem como*' (as well as), 17 were excluded because they did not have the synonyms in the text.

The last query, P-5, showed the need for more attention to details since it also involves the inference process that includes the search for hyponyms. "*Violação*" (violation) is the synonym for 'crime' (crime). The meaning adopted refers to an action that breaks a law, principle, or agreement from perspective of the ordinary citizen; *i.e.*, it does not include all technical aspects that the word expresses in terms of '*Direito*' (Law). Thus, it includes the following hyponyms for the domain: *ataque* (assault), *assalto* (robbery), *falsificação* (counterfeit), *roubo* (theft), *rapto* (kidnapping), and *infração* (infringement).

It is noteworthy that, in general, a search includes from the most general to the most specific concepts, thus in the existential path towards the concept, an ordinary user does not take into account the complexities and existing linguistic relations between his/her query and the expected result. When a general term is searched, the user is satisfied by the retrieved documents that meet his/her information needs without realizing that the results include more specific concepts, such as '*roubo*' (theft), which is an extension of the concept '*crime*'. Duly noted hyponyms play this "specifying" role in a semantic query.

Conversely, in the intentional path towards the concept, there is an increase in the scope and vagueness in the search, leading to an increase in recall but a reduction precision, which is not the objective of most search engines. The example below illustrates the taxonomy of the concepts of the term '*crime*', according to WordNet; from left to right, the most general to the most specific meaning:

*entidade→abstração→característica psicológica→evento→ato→atividade→transgressão→crime*

*(entity→abstraction→psychological characteristic→event→action→activity→transgression→crime)*

It was observed that the more general the term, the more comprehensive the search would be. For example, the search for the term '*atividade*' (activity) would certainly retrieve documents that are not related to '*crime*' (crime) at all; which is not desirable in this example. Thus, in this semantic search approach, only the hyponymic relationships that aim to improve the specificity of the search, and consequently its precision, are considered.

It can be seen from Table 1 that the traditional search retrieved 2 documents with 100,0% precision but with low recall, only 25,0%. The semantic search, however, also had great precision but with much higher recall, 87.5%, due to the fact that this search also identified hyponyms annotated at the database.

An important fact that deserves attention is that the failure to identify all relevant documents in the database, since there was one more document in addition to the 7 documents retrieved. In the manual analysis, one document that had the term '*atentado*' (attack), which is also a hyponym of '*crime*' was found, but it was not annotated or identified to be inserted into to the RiscoLex database.

This case could be easily solved with the introduction of the annotation and the insertion of the term into the RiscoLex. However, this fact is important to emphasize that human verification is a very important part of the process for debugging the search and the lexicon-ontological databases, OntoRisco and RiscoLex, resulting in a substantial information retrieval improvement. Furthermore, it shows the dynamic nature of the maintenance of knowledge bases, which require periodic maintenance.

Finally, it can be said that the queries submitted to the model showed that the semantic search outperforms the traditional search and validates the methodology. The process involving the preparation of ontolexical databases and the annotation of the corpus of the domain to be indexed are complex, but they promote considerable improvement in the task of providing the most appropriate information to the user. Although more complex, the procedure proposed can be used in all kinds of domain optimizing the results obtained.

## Conclusion

The present study addressed the use of Semantic Web technologies and textual information processing for the construction of a lexical-semantic database that is in accordance with the standard adopted by the W3C. The aim of this is to support the proposed semantic information retrieval model that uses linguistic and semantic information to build a lexical-semantic index that improves precision in the information retrieval process.

From the perspective of Information Science, research in this field is still scanty because the literature is mostly produced in the Computer Science field. There is "plenty of room" for scientific research that would foster the development of the information science field and promote the popularization of the Semantic Web also, including the view of information scientists. Therefore, one of the contributions of the present study is to bridge the gap between the development of computational resources and the management and organization of information, from the perspective of information science.

Another contribution of this study is to provide resources in Portuguese for the financial segment. The first resource is the construction of the RiscoLex, a novel lexical database containing morphological, syntactic and semantic information about terms related to *risco finaceiro* (financial risk). The second resource is the development of the ontology, OntoRisco, for the same domain.

For all of these reasons, it can be said that the objective of this study was achieved since the proposal was to create a lexical database, RiscoLex, in Brazilian Portuguese containing morphological, syntactic, and semantic information that can be read by machines in the RDF format allowing the link between structured *OntoRisco* data and unstructured textual corpus and integrate it into a semantic retrieval information model in order to improve precision.

As a suggestion for future research, we recommended a study on users in order to collect ideas to improve the vocabulary and, at the same time, consider the idiosyncrasies and jargons of the domains to be explored. This could contribute to improve the search results of lexical databases with semantic IRS. Moreover, the adoption of different weighting factors, other than the tf-idf, to address the lexical-semantic indexing would be highly useful. Moreover, the databases created by ontology lexicalization could be used as tools to improve automatic summarization or automatic text writing. Finally, the participation of lexicographers,

terminologists, and linguists in the building of lexical databases could greatly contribute to the interpretation and adequacy of linguistic phenomena to the ontology environment.

## Contributors

All authors contributed to the conception and design of the study, data analysis and final editing.

## References

ALLEMANG, D.; HENDLER, J. *Semantic web for the working ontologist*: Efective modeling in RDFS and OWL. San Francisco: Morgan Kaufmann, 2008.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*. Boston: Addison-Wesley Longman, 1999.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, v. 284, n. 5, p. 34-43, 2001. Available from: <https://www.scientificamerican.com/article/the-semantic-web/>. Cited: Sept. 17, 2011.

BERNERS-LEE, T. *et al*. *Semantic web road map*. [S.l.]: W3C, 1998. Available from: <https://www.w3.org/DesignIssues/Semantic.html>. Cited: Aug. 22, 2011.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with python*. Boston: O'Reilly Media, 2009. Available from: <http://www.nltk.org/book/>. Cited: Feb. 21, 2012.

BOND, F.; FOSTER, R. Linking and extending an open multilingual wordnet. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 51., 2013, Sofia. *Proceedings*... Sofia: Association for Computational Linguistics, 2013. p. 1352-1362. Available from: <http://aclweb.org/anthology/P13-1133>. Cited: May 17, 2014.

BRÄSCHER, M. *Tratamento automático de ambiguidades na recuperação da informação*.1999. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 1999.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDNS Systems*, v. 30, n. 1, p. 107-117, 1998. Available from: <http://www.sciencedirect.com/science/article/pii/S016975529800110X>. Cited: Jun. 13, 2014.

BUITELAAR, P. *et al*. Ontology lexicalisation: The lemon perspective. In: INTERNATIONAL CONFERENCE ON TERMINOLOGY AND ARTIFICIAL INTELLIGENCE, 9th, 2011, Paris. *Proceedings*... Paris: [s.n.], Institut National des Langues et Civilisations Orientales, 2011. p. 33-36.

CASTELLS, P.; FERNANDEZ, M.; VALLET, D. An adaptation of the vector-space model for ontology-based information retrieval. *Knowledge and Data Engineering*, *IEEE Transactions on*, v. 19, n. 2, p. 261-272, 2007.

CASTELLS, P. *et al*. Semantic web technologies for economic and financial information management. In: THE SEMANTIC WEB: Research and applications: European Semantic Web Symposium, 1st, 2004, Heraklion, Crete, Greece. *Proceedings*... Washington (DC): Springer, 2004. p. 473-487. Available from: <https://repositorio.uam.es/handle/10486/664140>. Cited: Jul. 15, 2014.

CASTELLS, P. *et al*. Neptuno: Semantic web technologies for a digital newspaper archive. In: *THE SEMANTIC WEB*: *Research and applications*. Washington (DC): Springer, 2004. p. 445-458. Available from: <http://nets.ii.uam.es/neptuno/publications/neptuno-esws04.pdf>. Cited May 15, 2013.

CIMIANO, P.; UNGER, C.; McCRAE, J. Ontology-Based Interpretation of natural language. *Synthesis Lectures on Human Language Technologies*, v. 7, n. 2, p. 1-178, 2014. Available from: <http://www.morganclaypool.com/doi/abs/10.2200/S00561ED1 V01Y201401HLT024>. Cited: Jun. 25, 2014.

CONTRERAS, J. *et al*. A semantic portal for the international affairs sector. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT, 14th, 2004, Whittlebury Hall, Northamptonshire, UK. *Proceedings*... [S.l.]: Springer, 2008. p. 203-215. Available from: <http://oa.upm.es/2632/>. Cited: May 13, 2014.

DAHLBERG, I. Teoria do conceito. *Ciência da Informação*, v. 7, n. 2, p. 101-107, 1978. Discponível em: <http://revista.ibict.br/ciinf/article/view/115/115>. Acesso em: 22 out. 2011.

FELLBAUM, C. *Wordnet*: An electronic lexical database. Cambridge (MA): The MIT Press, 1998.

FERNÁNDEZ, M. *et al*. Semantically enhanced information retrieval: An ontology-based approach. *Web Semantics*: Science, Services and Agents on the World Wide Web, v. 9, n. 4, p. 434-452, 2011. Available from: <http://dx.doi.org/10.1016/j.websem.2010.11.003>. Cited: Feb. 20, 2011.

FERNÁNDEZ, M. *et al*. Semantic search meets the web. In: IEEE INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, 2nd, 2008, Santa Clara, California, USA. *Proceedings*… Santa Clara (CA): USA: ICSC, 2008. p.253-260. Available from: <http://dx.doi.org/10.1109/ICSC.2008.52>. Cited: Jun. 12, 2014.

GRESSER, J. Y. *et al*. *Parsifal*: Project Protection and trust in financial infrastrcutures: D2.1 Draft ontology of financial risks & dependencies within & outside the financial sector. [S.l]: EDGE, 2010. Available from: <http://www.tssg.org/files/archives/PARSIFAL D2.1 Draft Ontology of Financial Risks.pdf>. Cited: Apr. 13, 2011.

GUARINO, N. Formal ontology in information systems. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS, 1st, 1998, Trento, Italy. *Proceedings*… Amsterdam: IOS Press, 1998. p. 3-15.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: STAAB, S.; STUDER, R. (Ed.). *Handbook of ontologies*. 2nd ed. Berlin: Springer, 2009. p. 1-17. Available from: <http://www.springerlink.com/index/10.1007/978-3-540-92673-3>. Cited: Sept. 27, 2011.

GUHA, R.; MCCOOL, R. Tap: A semantic web platform. *Computer Networks*, v. 42, n. 5, p. 557-577, 2003. Available from: <http://

www.sciencedirect.com/science/article/pii/S1389128603002251>. Cited: Jan. 18, 2011.

GUHA, R.; MCCOOL, R.; MILLER, E. Semantic search. In: ACM INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 12., 2003, Budapest, Hungary. *Proceedings*… New York: ACM, 2003. p. 700-709.

HEATH, T.; BIZER, C. Linked Data: Evolving the web into a global data space. Williston (VT): Morgan & Claypool Publishers, 2011. (Synthesis Lectures on the Semantic Web: Theory and Technology).

HOGAN, A. *et al*. Searching and browsing linked data with SWSE: The semantic web search engine. *Web Semantics*: Science, Services and Agents on the World Wide Web, v. 9, n. 4, p. 365-401, 2011. Available from: <http://www.sciencedirect.com/science/article/pii/S1570826811000473>. Cited: Apr. 30, 2012.

KARA, S. *et al*. An ontology-based retrieval system using semantic indexing. *Information Systems*, v. 37, n. 4, p. 294 305, 2012. Available from: <http://www.sciencedirect.com/science/article/pii/S030643791100113X>. Cited: Oct. 30, 2012.

LESK, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: ANNUAL INTERNATIONAL CONFERENCE ON SYSTEMS DOCUMENTATION, 5th, 1986, Toronto. *Proceedings*… New York: ACM, 1986. p. 24-26. Available from: <http://dl.acm.org/citation.cfm?doid=318723.318728>. Cited: Jun. 30, 2011.

MAEDCHE, A. *et al*. Seala framework for developing semantic web portals. In: ADVANCES IN DATABASES: British National Conference on Databases, 18th 2001, Chilton, UK. *Proceedings*... Washington (DC): Springer, 2001. p. 122. Available from: <http://link. springer.com/chapter/10.1007%2F3-540-45754-2_1>. Cited: Jun. 30, 2012.

McCRAE, J. *et al*. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, v. 46, n. 4, p. 701-719, 2012.

McCRAE, J.; SPOHR, D.; CIMIANO, P. Linking lexical resources and ontologies on the semantic web with lemon. In: THE SEMANTIC WEB: Research and applications. Extended Semantic Web Conference, 8th, 2011, Heraklion, Crete, Greece. *Proceedings*... Washington (DC): Springer, 2011. p. 245-259. Available from: <http://dl.acm. org/citation.cfm?id=2008914>. Cited: May 15, 2012.

NARDI, D.; BRACHMAN, R. J. An introduction to description logics. In: Baader, F. *et al*. (Ed.). *The description logic handbook*: Theory, implementation, and applications. Cambridge (MA): Cambridge University Press, 2003. p. 544. Available from: <https://www.inf.unibz.it/~franconi/dl/course/dlhb/dlhb-01.pdf>. Cited: Feb. 13, 2013.

NAVIGLI, R.; VELARDI, P.; GANGEMI, A. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, v. 18, n. 1, p. 22-31, 2003. Available from: <http://ieeexplore.ieee.org/document/1179190/?reload=true>. Cited: May 3, 2012

OLIVEIRA, H. G. Onto. PT: Towards the automatic construction of a lexical ontology for portuguese. Thesis (Doctoral dissertation) – University of Coimbra, Portugal, 2013.

PAIVA, V.; RADEMAKER, A.; MELO, G. *Openwordnet-pt*: An open brazilian wordnet for reasoning. Rio de Janeiro: FGV, 2012. (EMAp Technical Reports). Available from: <http://biblioteca digital.fgv.br/dspace/bitstream/handle/10438/10274/emap-techreport.pdf?sequence=3&isAllowed=y>. Cited: Aug. 6, 2012.

PEDREGOSA, F. *et al*. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, v. 12, p. 2825-2830, 2011.

POPOV, B. *et al*. Kim a semantic platform for information extraction and retrieval. *Natural Language Engineering*, v. 10, n. 9, p. 375-392, 2004.

REYMONET, A.; THOMAS, J.; Aussenac-Gilles, N. Modelling ontological and terminological resources in OWL DL. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 6th, 2007, Busan, South Korea. *Proceedings*… Toulouse: Institut de Recherche in Informatigue, 2007. v. 7.

ROCHA, C.; SCHWABE, D.; ARAGÃO, M. P. A hybrid approach for searching in the semantic web. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 13rd, 2004, New York. *Proceedings*… New York: ACM, 2004. p. 374-383. Available from: <http://doi.acm. org/10.1145/988672.988723>. Cited: Apr. 28, 2011.

SÉRASSET, G. Dbnary: Wiktionary as a lemon based rdf multilingual lexical resource in RDF. *Semantic Web Journal*, n. 1, p. 1-7, 2014.

SILVA, F.; GIRARDI, R.; DRUMOND, L. An information retrieval model for the semantic web. In: IEEE INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: NEW GENERATIONS, 6th, 2009, Las Vegas. *Proceedings*… Las Vegas: ICSC, 2009. p. 143-148. Available from: <http://ieeexplore.ieee.org/document/507 0607/>. Cited: Aug. 24, 2014.

UNGER, C. *et al*. A lemon lexicon for DBpedia. In: INTERNATIONAL WORKSHOP ON NLP AND DBPEDIA, 1st, 2013, Sydney. *Proceedings*… Aachen: CEUR-WS, 2013. Available from: <http://dl.acm.org/citation.cfm?id=2874479.2874491>. Cited: Sept. 18, 2014.

VALLET, D.; FERNÁNDEZ, M.; CASTELLS, P. The quest for information retrieval on the semantic web. *Upgrade*: The Europhan Journal for the Informatics Professional, n. 6, p. 19-23, 2015. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.287&rep=rep1&type=pdf>. Cited: Dec. 6, 2014.

WORLD WIDE WEB CONSORTIUM. OWL 2 Web Ontology Language. Cambridge (MA): W3C OWL Working Group, 2014. Available from: <http://www.w3.org/TR/2012/REC-owl2-new-features-20121211/>. Cited: May 17, 2014.

WALTER, S.; UNGER, C.; CIMIANO, P. A corpus-based approach for the induction of ontology lexica. In: *NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS*. Heidelberg: Springer, 2013. p. 102-113.

WALTER, S.; UNGER, C.; CIMIANO, P. Atoll: A framework for the automatic induction of ontology lexica. *Data & Knowledge Engineering*, v. 94, p. 148-162, 2014.

WILKS, Y.; BREWSTER, C. Natural language processing as a foundation of the semantic web. *Foundations and Trends in Web Science*, v. 1, n. 34, p. 199-327, 2007. Available from: <http:// www.nowpublishers.com/article/Details/WEB-002>. Cited: Oct. 18, 2014.