Towards a Semantic-Based Content Management System for Journalistic Writing

Invited paper

Vitor Silva de Deus
University of Brasília
Brasília, Brazil
vitor.sddf@gmail.com

Edison Ishikawa
University of Brasília
Brasília, Brazil
ishikawa@unb.br

Edgard Costa Oliveira
University of Brasília
Brasília, Brazil
ecosta@unb.br

Marcio Victorino
University of Brasília
Brasília, Brazil
mcvictorino@unb.br

Benedito Medeiros Neto
University of Brasília
Brasília, Brazil
medeirosneto@unb.br

Tor-Morten Groenli
Kristiania University College
Oslo, Norway
tmg@kristiania.no

Gheorghita Ghinea
Brunel University London
London, UK
Kristiania University College
Oslo, Norway
george.ghinea@brunel.ac.uk

ABSTRACT

Semantics is still a challenge to automation and to the improvement of the relationship between humans and machines. Content management systems have a lot of improvement possibilities using semantics. A news writing process incorporating a content management system that provides semantic search, semantic relationships between articles and communication with external semantic systems can improve the productivity of the writers and make the reader's task easier. This work presents a functional prototype of a content management system which focuses on the construction of semantic annotations based on domain ontology, reuse annotations in search and on relationship construction between stored texts, providing a semantic interface for external systems. We present in this paper the annotation algorithm, the use cases of annotation, article creation and editing, as well as an approach for doing a semantic search and creating semantic relationships between texts. The system enables users to create semantic annotations quickly and allows them to remove and add annotations, including those suggested by the annotation algorithm. The two approaches for semantic relationships between texts are accurate and useful, while the search tool is versatile because it allows users to search in semantic and non semantic fields at the same time and it also uses logic operators in all fields.

CCS CONCEPTS

• **Information systems** → *Information systems applications*; • **Software and its engineering** → *Software creation and management*;

KEYWORDS

Semantic Web, Semantic Search, Ontology, Semantic Computing, Web 3.0, Semantic Annotation, Content Management System

ACM Reference Format:
Vitor Silva de Deus, Edison Ishikawa, Edgard Costa Oliveira, Marcio Victorino, Benedito Medeiros Neto, Tor-Morten Groenli, and Gheorghita Ghinea. 2018. Towards a Semantic-Based Content Management System for Journalistic Writing: Invited paper. In *The 10th International Conference on Management of Digital EcoSystems (MEDES '18), September 25–28, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3281375.3281407

1 INTRODUCTION

A huge amount of information is produced daily on the Internet, and most of the data can be processed for general information retrieval purposes, as well as for data analytics, metrics and statistics. However, at what level does this processing happen? Lexical, semantic, context-based, or none of them? The reason is that whilst considerable information is being retrieved, this happens with high recall and low precision, due to the fact that web content is not context-free and machines cannot understand context nor meaning in each specific content. There is thus abundant semantic ambiguity and considerable concept variation due to regional and knowledge area restricted use of terms or concepts that cannot be automatically extracted from texts. [1, 2]

Concepts have semantic relationships with other concepts that cannot be extracted by simply making textual references to them.

MEDES '18, September 25–28, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-5622-0/18/09…$15.00
https://doi.org/10.1145/3281375.3281407

For example, the term Zika virus is associated with the concept of Zika disease, however extracting information from texts that mention both concepts separately is not possible. This is what happens in the present web (2.0) which is based on syntactic recognition of characters, text mining and statistical analysis of word counting or word occurrence in a text or set of texts.

These limitations have led to the development of technologies, tools and structures that allow machines to process semantic content for the Web: the Semantic Web. Semantic Web content provide structures that make semantic processing and inferencing in order to extract information which could only be processed or undestood by humans.The concept of Zika virus can be automatically related to the concept of Zika disease, to a certain local epidemy or to a specific case of Zika during pregnancy. This will happen only if these concepts are linked and shared on the web to be processed by machines. This new approach is only possible with the use of ontologies and Semantic Web tools.

In Brazil, the Zika epidemy is now under control but it still represents a public health problem, and has been highly documented, and commented upon on the web by experts and journalists in various news agencies. However, retrieving this content is difficult, due to the many shortcomings in the way information is being published. The main issue we target in this paper is the need to enhance semantic representation of content about a certain topic, Zika in this case, by content management systems of newsrooms. We are searching for a way to help journalists, general writers and readers to make use of semantic enhanced content, with the use of ontologies, in order to produce and access Zika-related content.

Let's assume that a journalist wants to write a text about *Zika disease*, and he/she needs references about the subject. If the content on the Web is not semantically linked, he/she will spend a reasonable timespan searching and making relationships with the texts and contents. It is noteworthy that if the journalist searches a keyword *Zika disease*, there can be retrived results such as cities that have the Zika epidemy in Brazil, the total amount of infected people, the transmitting mosquito, etc. Indeed, a text concerning information on the mosquito *Aedes aegypti* could also relate to the Zika disease and other transmitting diseases.

Semantic content machine processing aims to augment the knowledge aquisition capacity of machines but also of humans, since only humans understand semantics in a natural way, and humans are not able to process terabytes of contents fast and to make deductions or inferences of web content. The Semantic Web is of great help to content production and sharing in a faster and powerful manner, other than the present web which is based on links between content.

The objective of this paper is to present the initial implementation of the content dimension of the newsroom workflow [6]. We have built the prototype of a newsroom content management system - CMS - that generates semantic annotations to be used to help produce semantic-related content and share this content on the Web. This system provides the journalist a tool to write, annotate, search and publish papers with the support of semantic information processing. The semantic annotations are based on a domain ontology [8] presented in previous work [6], which also represents the environment where the current solution was tested.

2 THE NEWSROOM CMS

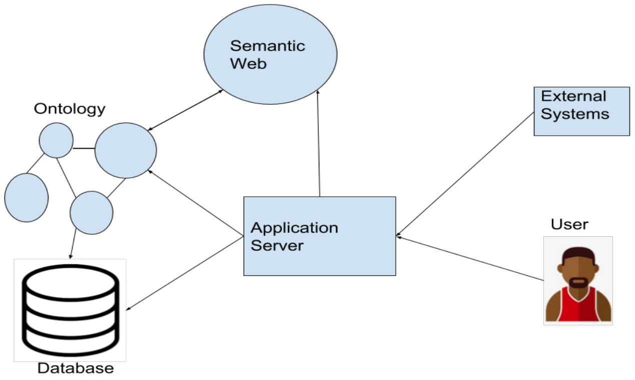


Figure 1: System's general architecture

The artifact is a prototype of a content management system developed with the purpose of extending the capabilities of journalistic writing in the production of journalistic articles as well as of increasing the capacity of its readers in the consumption of the information produced.

Semantic semi-automatic annotations of texts based on ontology produce information that can later contribute to the construction of semantic relationships between the texts. This helps both in the construction of new articles with reference to related articles already written and in the suggestion of articles related to an article that is being read. It also helps in the search of articles based on the semantics of their texts.

The ontology used as the basis for the semantic annotations in this first version of the artifact is described in [8], and was chosen because the artifact developed here is a possible practical alternative to what is also proposed in [8]. The artifact can also be presented as an authoring environment or knowledge production based on ontology and its requirements are inspired by what is presented in [7]. The general architecture of the artifact can be seen in Figure 1. The application server is accessed by external systems and users, and in turn accesses the database and the ontology that is stored locally. It also references Semantic Web concepts to build annotations.

2.1 Use Cases

The author's *persona* has the function of creating and reviewing articles. This function, in addition to creating the task itself, also includes those of annotating and editing a created article. The information produced by journalists can thus be consumed by the reader through the published articles via a search interface to them, and can also be consumed by other semantic systems through RDF files with the semantic annotations of each article.

The article creation and editing screen is shown in Figure 3. The title, text, and subtitle fields (or "soutien" [1]) are textual insertion fields whereas the editors and authors are multiple-selection. The "Annotate" button executes the annotation algorithm on the text

[1] French word which is a journalistic jargon in Brazil

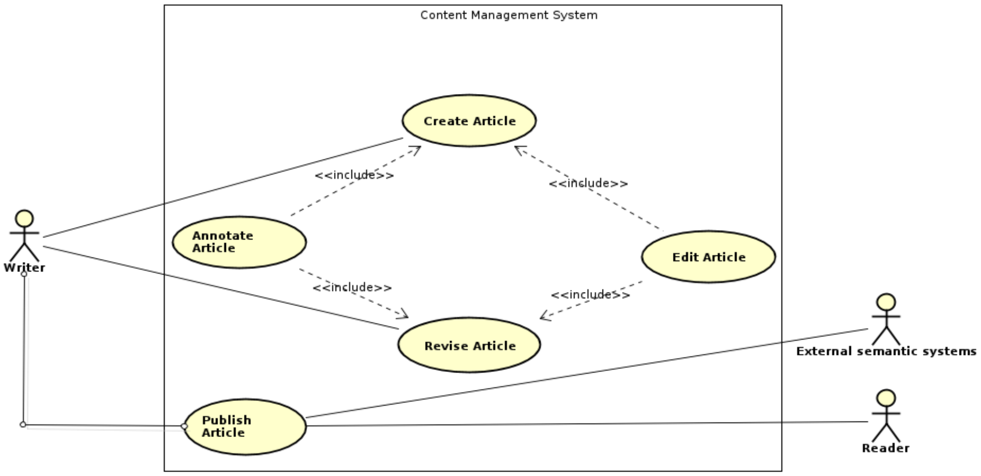


Figure 2: Use case diagram

and saves the article generating a list of concepts found according to Figure 4.

Concepts can be unmarked from the annotated list if they do not actually have a semantic match to the text and can also be added through the *Add concept* label selector. After the annotation, on the same page a list is generated with the five articles supposedly most related to the article being edited.

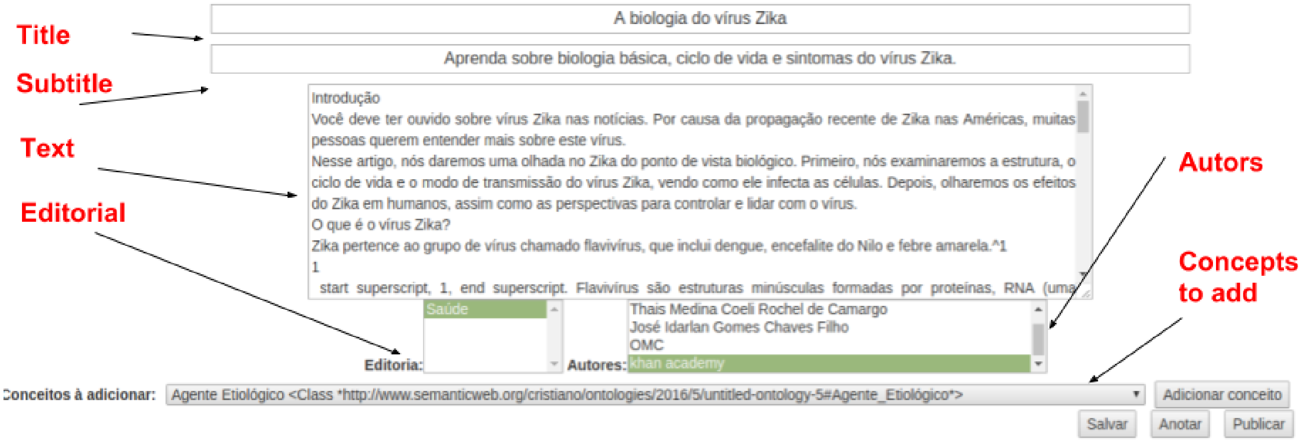
The annotation actor/user can unselect concepts from the list in case these concepts do not have semantic correspondence with the text. New concetps can be addded by a selector with a label *Add concept*. After the annotation is created, the system shows a list with the five more relevant existing articles semantically-related to the current article.

2.2 Data persistence architecture

It is important to show the database model of this work because it contains all the data of the CMS and the RDF triples of the semantic annotations in the same relational database. RDF triplesets form a graph structure and therefore should be stored in graph-based databases or unstructured ones. This is an option similar to those seen in the examples present in [11]. The existing options would be to use only a graph-based database or unstructured database in which it would be possible to represent all the semantic and non-semantic data of the prototype, to represent everything in a relational database, or to construct a hybrid solution. Given the time constraints, the tools used and the focus of the work, we chose to build a relational database for all data that is persisted and used internally in the system, including semantic annotations. However, for the data offered for external semantic systems, it was possible to provide the semantic annotations of each article in an RDF file with graph structure.

From Figure 5 it can be observed that an article consists of title, subtitle, text, editorials and authors. Each article can be related to several editorials and to several authors as well as a given editor, and a given author may be related to several articles. Each article has zero or more related published articles. The published articles have the content of the article that appears in the web (HTML), the article’s RDF file with semantic annotations and a date of publication. Each article is related to zero or more triples that are constructed from the semantic annotation of the text of the article. The triple is represented by a table in the relational database and contains a reference to its article (subject), a reference to a resource (predicate) and a resource (object). Each resource has a URI and a value such as http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Agente_Etiologico, the URI and Etiologic Agent being the value. Therefore the URI field is the replication of a Semantic Web identifier in the internal relational database of the prototype. Each resource can be related to its namespace (context) that can be for example a Semantic Web ontology, that is, this context can also be stored in the local bank as is the case of the ontology [8]. This database model was inspired by the solutions present in [11], which present solutions for storing RDF triples in relational databases.

Figure 6 shows an excerpt from the RDF file that is stored in the *published* relational database table and is offered to external semantic systems. In the beginning, within the rdf: RDF tag we have the *namespaces* or contexts that are referenced in the file, that is, it contains URIs that are referenced in the file, which in this case are those proposed by [3],[2] and [4]. The rest of the file is made up of rdf: Description tags each related to a concept found in the text of the article, such as Fase Viremia and Prurido. Thus, for each of these annotated concepts, five pieces of information are recorded inside the rdf: Description tag. The nodeID is a unique identifier for this item annotated inside the document. In the following line we have <aof: annotatesDocument rdf: resource = ” www.article2example.com ” /> where aof: annotatesDocument means that the annotatesDocument predicate of *namespace* aof is being referenced and rdf: resource = "www.article2example.com" is the predicate, we see that the nodeID "N976617b5b49e4bfcb0065a1274d8829d" annotates the article ” www.article2example.com.br ”.



A biologia do vírus Zika

Aprenda sobre biologia básica, ciclo de vida e sintomas do vírus Zika.

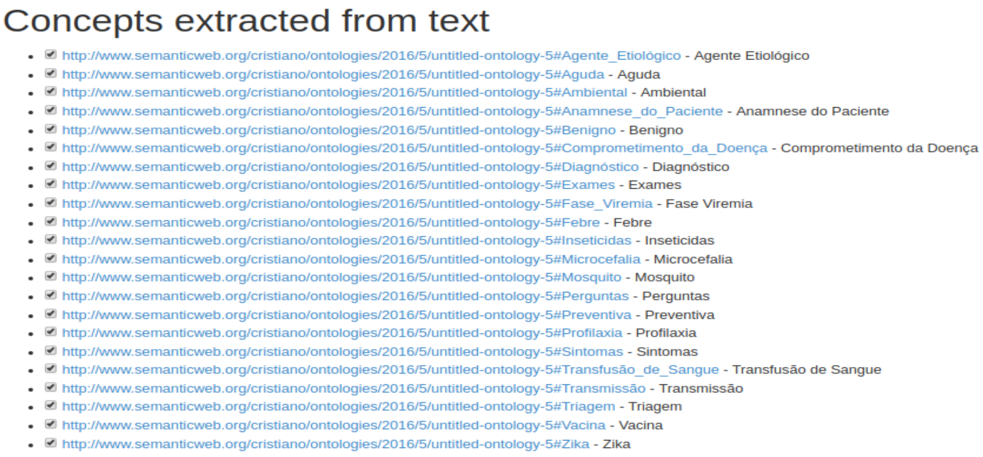
Introdução
Você deve ter ouvido sobre vírus Zika nas notícias. Por causa da propagação recente de Zika nas Américas, muitas pessoas querem entender mais sobre este vírus.
Nesse artigo, nós daremos uma olhada no Zika do ponto de vista biológico. Primeiro, nós examinaremos a estrutura, o ciclo de vida e o modo de transmissão do vírus Zika, vendo como ele infecta as células. Depois, olharemos os efeitos do Zika em humanos, assim como as perspectivas para controlar e lidar com o vírus.
O que é o vírus Zika?
Zika pertence ao grupo de vírus chamado flavivírus, que inclui dengue, encefalite do Nilo e febre amarela.^1
1
start superscript, 1, end superscript. Flavivírus são estruturas minúsculas formadas por proteínas, RNA (uma

Editoria: Saúde

Autores: Thais Medina Coeli Rochel de Camargo
José Idarlan Gomes Chaves Filho
OMC
khan academy

Conceitos à adicionar: Agente Etiológico <Class *http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Agente_Etiológico*>

Figure 3: Writer interface for creating and editing articles.



Concepts extracted from text

- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Agente_Etiológico - Agente Etiológico
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Aguda - Aguda
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Ambiental - Ambiental
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Anamnese_do_Paciente - Anamnese do Paciente
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Benigno - Benigno
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Comprometimento_da_Doença - Comprometimento da Doença
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Diagnóstico - Diagnóstico
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Exames - Exames
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Fase_Viremia - Fase Viremia
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Febre - Febre
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Inseticidas - Inseticidas
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Microcefalia - Microcefalia
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Mosquito - Mosquito
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Perguntas - Perguntas
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Preventiva - Preventiva
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Profilaxia - Profilaxia
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Sintomas - Sintomas
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Transfusão_de_Sangue - Transfusão de Sangue
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Transmissão - Transmissão
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Triagem - Triagem
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Vacina - Vacina
- ☑ http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Zika - Zika

Figure 4: Example of concepts annotated in the article.

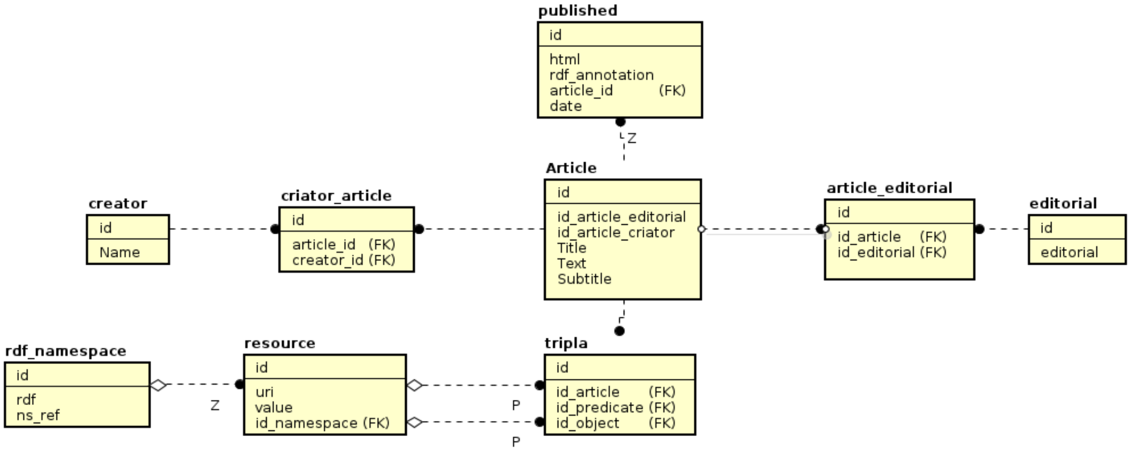
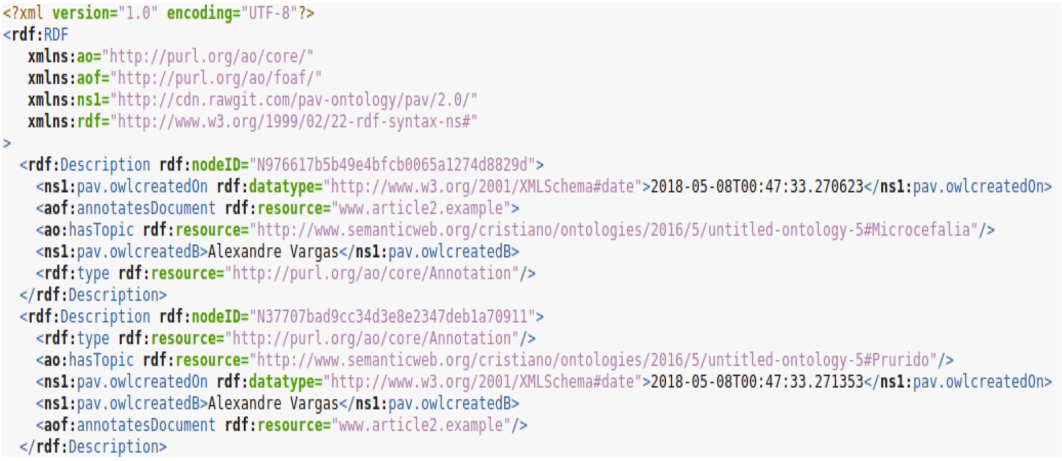


Figure 5: Relational Database Model



```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:ao="http://purl.org/ao/core/"
   xmlns:aof="http://purl.org/ao/foaf/"
   xmlns:ns1="http://cdn.rawgit.com/pav-ontology/pav/2.0/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:nodeID="N976617b5b49e4bfcb0065a1274d8829d">
    <ns1:pav.owlcreatedOn rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-05-08T00:47:33.270623</ns1:pav.owlcreatedOn>
    <aof:annotatesDocument rdf:resource="www.article2.example">
    <ao:hasTopic rdf:resource="http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Microcefalia"/>
    <ns1:pav.owlcreatedB>Alexandre Vargas</ns1:pav.owlcreatedB>
    <rdf:type rdf:resource="http://purl.org/ao/core/Annotation"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N37707bad9cc34d3e8e2347deb1a70911">
    <rdf:type rdf:resource="http://purl.org/ao/core/Annotation"/>
    <ao:hasTopic rdf:resource="http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Prurido"/>
    <ns1:pav.owlcreatedOn rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-05-08T00:47:33.271353</ns1:pav.owlcreatedOn>
    <ns1:pav.owlcreatedB>Alexandre Vargas</ns1:pav.owlcreatedB>
    <aof:annotatesDocument rdf:resource="www.article2.example"/>
  </rdf:Description>
```

Figure 6: Example of a file with an article’s semantic annotation

The <ns1:pav.owlcreatedOn rdf:datatype=http://www.w3.org/2001/XMLSchema#date> 2018-05-08T00:47:33.270623 </ns1:pav.owlcreatedOn> also makes a reference to a *namespace*, a data pattern, and stores a data value.

The line <rdf:type rdf:resource=”http://purl.org/ao/core/Annotation”/> shows that the register with a certain NodeID is the one referenced in http://purl.org/ao/core/Annotation, which is an annotation.

The line <ns1:pav.owlcreatedB> Alexandre Vargas </ns1:pav.owlcreatedB> also shows that the register with a certain NoteId was created by Alexandre Vargas, i.e. he annotated the term ”Microcefalia” in the article article2example.com.br. An association of the article to the concept ”Microcefalia” in the Semantic Web is done in the line <ao:hasTopic rdf:resource=”semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Fase_Viremia”,which shows that the article has a concept which is also referenced at : http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Microcefalia.

2.3 The Annotation Algorithm

The semantic annotation of the text, performed by the system, is an ontology-based semi-automatic annotation, that is, an annotation made by a person with the support of a computer and an ontology.

Figure 7 presents a diagram with the annotation algorithm receiving the text, and by using the ontology, it produces a list of suggested annotations [12]. This list is filtered and incremented by the author of the text according to his/her perception of the semantics of the concepts present in the text. After the author’s confirmation, the annotations are stored in the database.

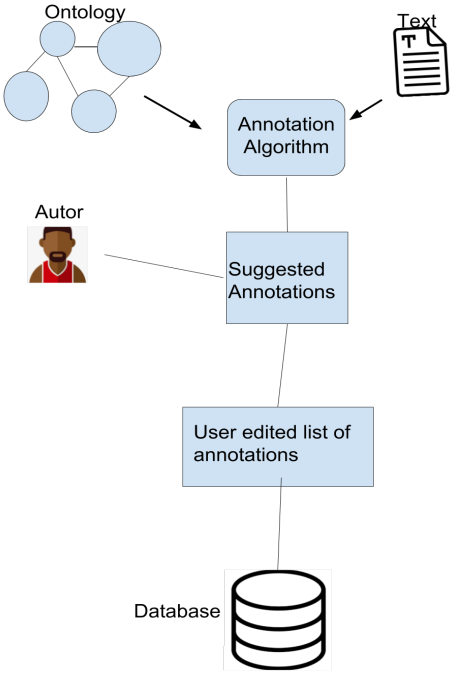


Figure 7: Semantic annotation process

The first task the algorithm performs is to chop the article and transform a list with a phrase into a list of sentences divided per each work organized in the same sequence as they appear in the text. Then we build a list with all the terms representing the searched concepts to be searched in the Zika ontology [8]. The list of terms/concepts is ordered by the number of terms of each concept.

For each concept size (number of words in the textual representation of the concept) there is a check to establish whether there are periods of size greater than or equal to that. If not, all concepts with that particular size are removed from the list. If yes, this particular concept is searched in all periods that have a size greater than or equal to it. If it is found in some, the concept is inserted into a list of concepts for annotation and is removed from the previous list. If it is not found, the concept is also removed from the list of

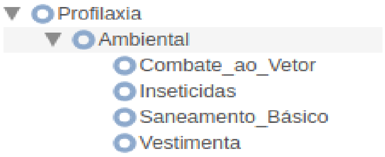


Figure 8: Example of reification.

concepts being sought. The algorithm continues until the concept list is empty.

For each annotated concept, all parent nodes in the ontology are also annotated. For example, in Figure 8, Clothing, Basic Sanitation, Insecticides, Combat Vector are instantiations of Environmental Prophylaxis and therefore a text that speaks of Clothing or Combat to Vector of Zika also speaks of environmental prophylaxis of Zika, for example.

The algorithm, in its pseudocode, is shown as follows.

Algorithm 1 Find concepts in text

```
procedure FIND(list concepts_list,texto)
    $found\_concepts \leftarrow empty\_list()$
    $sentences \leftarrow divide\_into\_sentences(text)$
    $remaining\_concepts \leftarrow concepts\_list$
    while $not\_empty(remaining\_concepts)$ do
        $t \leftarrow len(longest(i \in remaining\_concepts)$
        $long\_concepts \leftarrow list(i \in concepts\_list | len(i) = t)$
        $remaining\_concepts \leftarrow list(x | x \notin long\_concepts$
        for $sentence \in sentences$ do
            if $len(sentence) \geq t)$ then
                for $concept \in processing\_concepts$ do
                    if $concept \in sentence$ then
                        $found\_concepts.insert(concept)$
    $return(found\_concepts)$
```

2.4 Searching in the Relational Database

Search results are articles published in the system. They comprise six fields, two with semantic data and four with common data of articles which are text, title, " soutien " (subtitle), editorials and authors. The two semantic fields are the fields labeled *Concepts* and *URIs* according to Figure 9. The field Concepts refers to the field that stores a textual representation of a concept of the ontology in the relational database and is more specifically the field *value* of the *resource* table shown in Figure 5 which is: search for Semantic Web resources that are related to that article through a simple textual representation. The *URIs* field also performs a search in the *resource* table but in the *uri* field, which is: it searches the real representation of that resource or concept in the semantic Web. The *Editor* and *Authors* fields allow users to filter articles by their authors and publishers. The other fields enable the filtering of articles by the very own fields of the table *Article*.

For each field a different query is performed on the database, and then the union of those queries is generated as a result. Each field accepts a search expression that can contain parentheses, quotation marks, the OR operator (|) or the AND operator (&) in infix notation. Some examples of queries are presented in 9.

In Figure 9 there is an example of a query in which the articles that mention fever or treatment or that contain the word dengue in the title are searched. Another example would be to search for articles that have the terms Zika and disease or the term fever in the title.

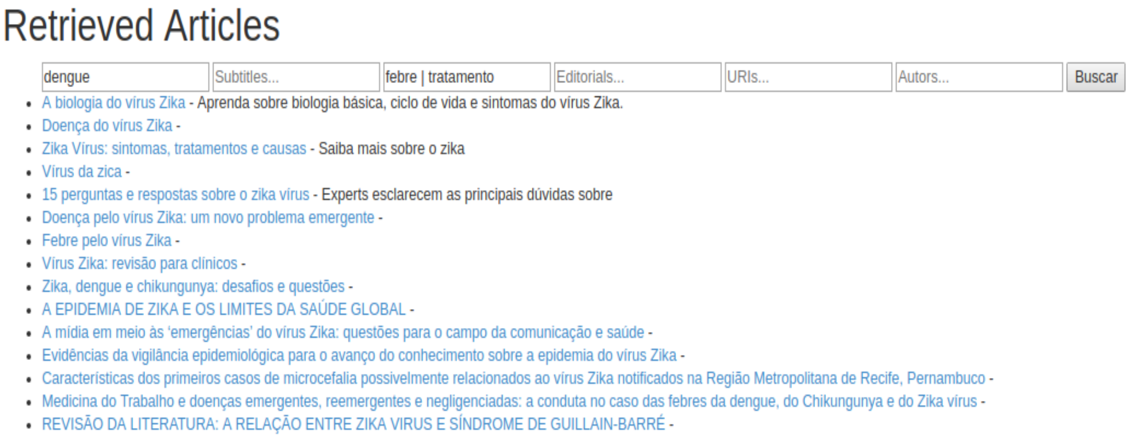
In a nutshell, the layer that performs each query from a search expression also performs the pre-processing of the expression. It does so by assembling a list of operators, parentheses, and operands in which each position in the list is an operator or parentheses or word, or words with inside quotation marks, all in the same infix position in which they were informed in the search field. Then the expression is passed to postfix notation and each operand is replaced by a set of articles to which that word or words are referenced. For example, searching for the word "Vitor" in the author's field will return all the articles containing the word 'Vitor' in the author's name, such as "Vitor Silva" and "Vitor Laerte". Then the expression in postfix notation in which the operands are sets of articles is processed by intersecting when the operand is "&" and by union when it is "|". Another example would be to search for "Vitor Silva" in the title and then return all articles that contain the word "Vitor" or the word "Silva" in the title. Indeed, one could also search for "Vitor Silva" and have as a return all articles containing "Vitor Silva" in the title.

2.5 Two Approaches for Semantic Relationship Inference between Texts

Two possible approaches to the semantic relationship between texts are proposed here. The first one essentially causes a text B of a set C of texts to be more related to another text A if B is the text whose intersection between the annotated concepts of B and A is the largest possible of all the others texts of C. Therefore, if we want to know, from this approach, the five texts of a set C more related to a text A, we obtain the five texts with the largest size of intersection of their annotated concepts with those annotated by the size of this intersection.

The second approach, which is slightly more complex, uses the structure of the ontologies, that contain the concepts related to the texts, to extract a metric for the relationship between two texts. In a nutshell, this metric is the amount of sibling concepts that exist between the annotated concepts of two articles A and B including the concept itself. In other words, the quantity of concepts of article A that has the same parent in the ontology of concept B. For example, by observing Figure 10 we notice that the intersection between the concepts of two articles shows that one speaks of the Zika Virus and the Zika Disease is empty. However the two articles are speaking of Etiologic Agent, which is the father of the two previous concepts in the ontology and therefore the two articles are talking about related subjects in a certain level of abstraction, and this second approach takes that into account.

The results obtained in both approaches are highly dependent on the data mass of the system. It is not trivial to extract a well-grounded metric on efficiency and characteristic of the two approaches without detailed study with a large mass of data, and this was not the focus of this work. However, for the purpose of exemplifying the operation of the system, some qualitative results obtained from the two algorithms are presented here from a small



Retrieved Articles

dengue | Subtitles... | febre | tratamento | Editorials... | URIs... | Autors... | Buscar

- A biologia do vírus Zika - Aprenda sobre biologia básica, ciclo de vida e sintomas do vírus Zika.
- Doença do vírus Zika -
- Zika Vírus: sintomas, tratamentos e causas - Saiba mais sobre o zika
- Vírus da zica -
- 15 perguntas e respostas sobre o zika vírus - Experts esclarecem as principais dúvidas sobre
- Doença pelo vírus Zika: um novo problema emergente -
- Febre pelo vírus Zika -
- Vírus Zika: revisão para clínicos -
- Zika, dengue e chikungunya: desafios e questões -
- A EPIDEMIA DE ZIKA E OS LIMITES DA SAÚDE GLOBAL -
- A mídia em meio às 'emergências' do vírus Zika: questões para o campo da comunicação e saúde -
- Evidências da vigilância epidemiológica para o avanço do conhecimento sobre a epidemia do vírus Zika -
- Características dos primeiros casos de microcefalia possivelmente relacionados ao vírus Zika notificados na Região Metropolitana de Recife, Pernambuco -
- Medicina do Trabalho e doenças emergentes, reemergentes e negligenciadas: a conduta no caso das febres da dengue, do Chikungunya e do Zika vírus -
- REVISÃO DA LITERATURA: A RELAÇÃO ENTRE ZIKA VIRUS E SÍNDROME DE GUILLAIN-BARRÉ -

Figure 9: Example of query

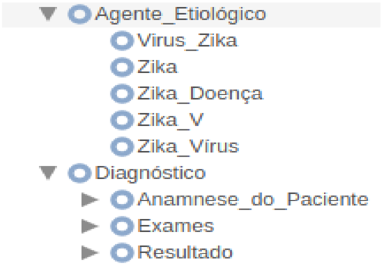


Figure 10: Example of sibling concepts in the ontology

mass and restricted to the Zika Virus domain, presented in Figure 10. Accordingly, journalistic and scientific articles dealing with Zika were inserted into the database as detailed in Table 1.

Table 1: Types of articles inserted in the system

Type of article	Quantity
Jornalístic	14
Scientific	16
Total	30

For the first approach, by using texts restricted to the same domain (which is the case of the test mass to which we are submitting the two algorithms) a feature is clear from simple observation: texts that refer to many different concepts of the domain end up being related to many other texts, due to the fact that there are greater chances that the intersection of their concepts occur with other concepts. An example is the article Labor Medicine and Emerging, Reemerging and Neglected Diseases: Dengue Fever, Chikungunya and Zika Fever [9] is among the five most related to eighteen of the other twenty articles (already registered at the time of this evaluation) and contains thirty-nine of the forty-five concepts present in the ontology.

In the second approach, the article [9] still appeared in seventeen out of twenty articles, which is logical since an article that has 39 of the 45 mapped concepts of the domain will also have more (or at the very least, the same number of) sibling concepts as other articles in the domain. In general, the first five related texts did not change much. As a rule, they changed their order and changed into two or three.

One way to generate more accurate results on the two approaches would be to use a domain expert to populate the system with a larger mass of domain texts so that groups of texts belong to common subdomains within the larger domain, and then extract metrics from how efficiently algorithms relate texts within subdomains and the domain that contains them.

Another important observation is that the scientific articles, on average, have more triplets generated by the automatic annotation, namely, suggestions of annotations made by the system, as shown in Table 2. This can be interpreted as a superficial content use in journalistic articles, that is, scientific articles on average cite more terms from the domain of Zika's ontology than journalistic articles. The standard deviation of the average of both types of articles shows that the samples are quite varied with respect to the number of triplets associated with each article.

Table 2: Average number of RDF triples associated with each type of article

Type of item	Average number of triples	Standard deviation
Scientific	16,93	9,3
Journalistic	11,07	8,06

3 CONCLUSIONS

In this work, we presented an RDF interface of a semantic-based authoring environment, operated by two algorithms, a system that performs semantic search in the texts, automatic relationship between texts and communication with external semantic systems. Of course these are some major bottlenecks of a non-semantic CMS. The semantic search proposed in this work solves the ambiguity problem by allowing queries with URIs as well as the relationship algorithms of the texts from their URIs. The semantic interface proposed for external systems containing the HTML and an RDF with the semantic annotations of each article offers a well defined structure that reuses Semantic Web resources and that can be perfectly consumed by any external agent. The reliability of semi-automatically-produced semantic annotations is guaranteed by the fact that in addition to the automatic generation of annotations, the person who is annotating the text can manually remove or add annotations according to his/her understanding of the text [5, 10] . The annotation of concepts that were not directly quoted in the text but are present due to their relation to some annotated concept is also made. For example, if the text is about Zika fever then it is also about a symptom of Zika and this information is also annotated, i.e. the automatic support made to the semantic annotations is providential and successfully supports the requirement to enable semi-automatic semantic annotations[8]. For the relationship between texts, two viable approaches were presented here. The first approach only takes into account the annotated concepts of an article in relation to the annotated ones of another article. The second approach also takes into account the generalization of the concepts, that is, if different concepts are instantiations of the same concept, then they are related.

ACKNOWLEDGMENTS

We would like to thank Kristiania University College for proving the means to publish this paper. Special thanks to Computer Science Departments of Brunel University London and University of Brasília, CIC/UnB, via the MDM Project of the Brazilian Ministry of Education - CAPES, under the Grant Number MEC/MCTI/CAPES/CNPq/FAPs 88881.068354/2014-01. The project also thanks the infrastructure of the Experimental Laboratory for the Study of Digital Languages for Mobile Devices (Labdim) of the Faculty of Communication, University of Brasilia, registered under number 485 707 in CNPq / 2013-6.

REFERENCES

[1] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* (5 2001), 1.

[2] Paolo Ciccarese, Marco Ocana, Leyla Jael Garcia Castro, Sudeshna Das, and Tim Clark. 2011. An open annotation ontology for science on web 3.0. *Journal of Biomedical Semantics* 2, 2 (2011), 1.

[3] Paolo Ciccarese, Marco Ocana, Leyla Jael Garcia Castro, Sudeshna Das, and Tim Clark. 2018. Anottation Ontology. http://annotation-ontology.googlecode.com/svn/trunk/. (2018). Acessed on 07/06/2018.

[4] Dan Brickley and Libby Miller. 2018. FOAF Vocabulary Specification. (2018). Retrieved June 7, 2018 from http://xmlns.com/foaf/spec/

[5] Maryam Hazman, Samhaa El-Beltagy, and Ahmed Rafea. 2012. An Ontology Based Approach for Automatically Annotating Document Segments. *International Journal of Computer Science Issues,* 9, 2 (2012), 3–9.

[6] International Center for Journalists. 2018. A Study of technology in newsrooms. (2018). Retrieved June 20, 2018 from https://medium.icfj.org/a-study-of-technology-in-newsrooms-cea3252ce5df

[7] Edgard Costa Oliveira. 2006. *Autoria de documentos para a Web Semântica: um ambiente de produção de conhecimento baseado em ontologias.* Ph.D. Dissertation. pages 107-127 ,http://repositorio.unb.br/handle/10482/4794 .Acessed on 07/06/2018.

[8] Edgard Costa Oliveira, Edison Ishikawa, George Ghinea, Thabata Hellen Granja, Marcos Nunes, Lucas Hiroshi Hironouchi, Rafael Batista Menegassi, Luciano Gois, and Daniel Rodriguez. 2016. Designing an Ontology-based Zika Virus news authoring environment for the Semantic Web. *roceedings of the 8th International Conference on Management of Digital EcoSystems* (11 2016), 1–7.

[9] Marcelo Pustiglione. 2018. Medicina do Trabalho e doenças emergentes, reemergentes e negligenciadas: a conduta no caso das febres da dengue, do Chikungunya e do Zika vírus | Biblioteca Virtual em Saúde. http://pesquisa.bvsalud.org/cvsp/resource/pt/lil-779356?lang=pt. (2018). Acessed on 07/06/2018.

[10] Quratulain Rajput and Sajjad Haider. 2011. BNOSA: A Bayesian network and ontology based semantic annotation framework. *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 2 (2011), 4–11.

[11] Stanford University. 2018. Storing RDF in a relational database. (2018). Retrieved June 7, 2018 from http://infolab.stanford.edu/~melnik/rdf/db.html

[12] Peng Wang, Bao wen Xu, Jian jiang Lu, Da zhou Kang, and Yan hui Li. 2004. A novel approach to semantic annotation based on multi-ontologies. *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference* (2004), 4–8.