A.A. Baptista et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-065-9-72

Knowledge Network of Scientific Claims Derived from a Semantic Publication System

Carlos Henrique MARCONDES[a,1]

[a] *University Federal Fluminense, Department of Information Science, R. Lara Vilela, 126, 24210-590, Niterói, Rio de Janeiro, Brazil, phone 55 21 26299758, marcon@vm.uff.br*

Abstract. Currently, the conventional communication channel for reporting scientific results is Web electronic publishing of scientific articles in paper print formats, such as PDFs. The emergence of the Semantic Web and Linked Data environment provides new opportunities for communicating, sharing, and integrating scientific knowledge in digital formats that could overcome the limitations of the current print format, which is only suitable for reading by people. The results of scientific research can be published electronically and shared in structured, interlinked formats. This integrated knowledge network could be crawled by software agents, thereby facilitating semantic retrieval, knowledge reuse, validation of scientific results, identification of traces of scientific discoveries, new scientific insights, and identification of knowledge contradictions or inconsistencies. This paper explores the possibilities of this new environment for scientific publishing and reports the implementation of a prototype semantic publishing system, which publishes scientific articles in a paper print format and publishes the claims made in the conclusions of each article as structured triples using the Resource Description Framework format.

Keywords: semantic publishing, semantic web, linked data, knowledge networks, knowledge representation, terminological knowledge bases, e-science

Introduction

The Web is fast becoming a universal platform for the disposal, exchange, and access of knowledge records. An increasing amount of human cultural records, derived from text, static or motion images, sound, and multimedia, are now being created directly in a digital format. Scientific activity has always involved the intensive processing of data and information. It is also being affected by the emergence of Web facilities for Grid technologies that can process vast amounts of scientific data, as well as the publication and maintenance on the Web of full text, datasets, and algorithms used for scientific data processing.

The current scholarly Web publishing environment is still an electronic metaphor of the paper print publishing environment used throughout the twentieth century. Despite numerous advances in information technology, Web electronic publishing is

[1] Corresponding Author: Carlos Henrique Marcondes, Department of Information Science, R. Lara Vilela, 126, 24210-590, Niterói, Rio de Janeiro, Brazil; phone 55 21 26299758; E-mail: marcon@vm.uff.br.

still based on the print text model. Scientific results are traditionally published in articles with a textual format, which limits the possibilities for the reuse and validation of scientific results that are published in the Web environment. Reuse requires the identification of similar patterns and the ability to compare similar parameters and differences in vast amounts of data. Validation requires reproducible methods, which can only be achieved if the same data processes, and algorithms are in the original experiments.

The reuse and validation of scientific results demands tools for information discovery, retrieval, and comparison in a very specific, precise, and meaningful manner. Current information retrieval systems do not contain explicit meaningful relations between elements, the content of documents, or resources they represent. Boolean operators are too general and they lack the semantic expressiveness necessary for content retrieval in specific scientific domains. Relations expressed using Boolean operators are processed as extensive set operations using the keywords included in the bibliographic records, rather than as intensive semantic relations among concepts.

Scientists now have to search through a vast amount of information resources that may be of potential interest, which is available on the Web. These resources range from electronic publications, digital libraries, repositories of full-text papers, algorithms, datasets of scientific data, terminological knowledge bases, and virtual machines that can process these data. Scientists have to devote greater efforts to discovering, examining, comparing, and integrating these resources [1]. The Web provides a potentially integrated and comprehensive environment for scientific activities, including scientific publication, but the textual format used for publishing results prevents computer programs from directly processing the content of scientific articles, which would facilitate scientific knowledge management tasks.

The Web environment now provides an opportunity to publish and integrate all these elements. The challenges posed by this scientific environment demands the development of methods and technologies for the direct processing of knowledge by computers. Scientific terminologies are one of the most essential components of scientific communication, and they are evolving towards more formal knowledge bases [2] that can be processed by computers. Comprehensive biomedical terminologies include the Unified Medical Language System (UMLS) and the Semantic Network (SN)[2], which is a classification schema of the UMLS Meta-thesaurus that organizes every concept into hierarchical trees, where each has as its root in a top level Semantic Type. The UMLS SN uses 54 Relation Types to express the semantic relations among concepts in the Semantic Type hierarchies used to index Biomedical scientific articles. The UMLS SN contains the permitted relations among Semantic Types. According to National Library of Medicine, USA, UMLS Fact Sheet [3]: **"***The purpose of NLM's Unified Medical Language System (UMLS®) is to facilitate the development of computer systems that behave as if they understand the meaning of the language of biomedicine and health***".**

In particular, the Semantic Web and Linked Data technologies [5] provide new opportunities for communicating, sharing, reuse, interlinking, and integrating scientific knowledge published in digital formats that may overcome the limits of the current print format used for the publication of scientific results, which is only suitable for reading and processing by people. We are now beginning to use these technologies for sophisticated tasks such as knowledge discovery, knowledge comparison, and

[2]UMLS Semantic Network, http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

integrating multiple sources, which facilitates inference capabilities among different and autonomous information resources.

Scientific publications were first recorded in bibliographic databases and citation analysis, citation models, and citation networks were developed as tools to understand and manage the development of science [5]. The Web offers the possibility of developing a richer and more multifaceted scientific knowledge environment where navigating throughout a citation network will be only one of the many possibilities.

This paper addresses the integration of a scholarly semantic publishing environment with the future e-science environment, which is being designed based on the Semantic Web and Linked Data technologies. Specifically, we address the question of how to identify, extract, and represent the knowledge embedded in the text of Web published scientific articles in a machine processable format in compliance with Semantic Web standards. The paper is structured as follows. The next section outlines the proposed semantic publication model and reports the development of a Web author submission interface to a journal system, which partially implements it. Section 2 presents the proposed semantic record model and discusses its implementation in the Resource Description Framework (RDF). Section 3 discusses how the model proposed may provide the core of a knowledge network, thereby facilitating a range of scientific knowledge management methods, presenting conclusions, and future research developments.

1. A Semantic Model for Scholarly Electronic Publishing

In a previous study, we proposed [6] a semantic model for electronic publishing. The model was developed through the analysis of scientific articles in the area of biomedical science, where we identified patterns of reasoning and semantic elements. The aim of the model was to achieve a semantically richer content for the representation of biomedical articles in a computer program “understandable” format. This proposal is based on the following hypotheses: (a) scientific knowledge that appears in the text of scientific articles consists of scientific conclusions made by authors and takes the form of relations between phenomena; (b) these relations are expressed linguistically using propositions and related concepts. Thus, it is feasible to use an authoring/Web publishing tool to ask authors to enter a conclusion and perform natural language processing (NLP) of the conclusion text to identify, extract, and represent this knowledge in a structured format.

This knowledge representation format allows programs to perform “inferences” based on the knowledge content of articles, thereby facilitating more semantically powerful content retrieval and knowledge management compared with current Bibliographic Information Retrieval Systems.

The proposed model has two components: an enhanced record model, i.e., a semantic record, and a Web interface allowing authors to self-publish and self-submit articles to a journal system. The record model *extends* conventional bibliographic record models, which comprise of conventional descriptive elements such as authors, title, abstract, bibliographic source, publication date, and content information, such as keywords, descriptors, and references to cited papers. In addition to these elements, the model includes claims made by authors in the conclusions of their articles. Miller [7] states that “science is a search after internal relations between phenomena”. Scientific knowledge found in the text of scientific articles consists of scientific claims made by authors throughout the article text, which is synthesized in the article's conclusion. It takes the form of relations between phenomena or between a phenomenon and its characteristics. These relations are expressed linguistically using propositions that relate concepts. Scientific claims are represented as *relations* between two different phenomena or between a phenomenon and its characteristics [8], e.g., (a) "telomere shortening (Phenomenon) causes (Type_of_relation) cellular senescence (Phenomenon)" [9], (b) "telomere replication (Phenomenon) involves (Type_of_relation) nontemplate addition of telomeric repeats onto the ends of chromosomes(Phenomenon)?" [10], or (c) "tetrahymena extracts (Phenomenon) show (Type_of_relation) a specific telomere tranferase activity (Characteristic)" [11]. Such relations could be modelled as triples of <Antecedent><Type_of_relation> <Consequent>.

Our research also includes the development of a prototype Web author submission interface for a journal system, which partially implements the model [12] where a general framework is proposed for identifying discoveries in scientific papers based on two aspects: their rhetoric elements and patterns and by comparing the content of the article conclusions with terminological knowledge bases [13].

In the proposed model, authors use a journal system Web submission interface to type the article conclusions and other standard metadata during the submission/upload of their article text. The system performs NLP of the text of the conclusion, before formatting it as a relation. Thus, we propose to engage authors in the development of a richer content representation of their own articles. The system interacts with authors and asks them to validate the extracted relations while the system of concepts found in the conclusion is mapped to concepts in a domain terminological knowledge base.

The result of this processing are recorded as a richer semantic content bibliographic record where scientific claims made by authors throughout articles are expressed as relations. In addition to being published in a textual format, each article's claims are also represented as structured relations and recorded in a machine-understandable format using Semantic Web standards such as RDF [14] and OWL (Web Ontology Language) [15], allowing the claims to be formally related by the author, i.e., mapped and annotated to concepts in a domain terminological knowledge base.

The terminological knowledge base used by our prototype system is UMLS. The author is asked to validate the automatic mapping made by the system, selecting other terms from a list displayed by the system and deciding whether satisfactory mapping options are offered. If satisfactory options are not available, the system assigns this specific element of the relation to "no mapping." The result is that the conclusion terms annotated by the author are related to terms in a terminological knowledge base *at the time of article publication*.

Once the relations are recorded in a database the machine-understandable records resulting from this publishing model can be processed and compared with public knowledge using software agents, e.g., published scientific articles, or with terminological knowledge bases throughout the Web. This provides scientists with new tools for knowledge retrieval, claims comparison, identification of contradictory claims, the use of these claims in different contexts, and the identification and validation of new contributions to science made in specific articles.

The proposed model is described in detail in [6].

A prototype of the submission system was developed to evaluate the dialogue with authors and the extraction routine. The prototype of the interface is in its initial development phase. It is not yet a final version or a production system. It has been developed with the specific aim of demonstrating the feasibility of the proposed system. It was developed as a Java application using MetaMap[3] program to perform NLP of article conclusions with LingPipe[4].

The prototype system processes selected parts of the article text uploaded by authors, i.e., the title, abstract, keywords, introduction, methods, and results. The introduction and abstract are used to extract the objectives of the article by the identification of phrases such as the "*objectives of our work…*" and *"the goal of the present work…"* The author is asked to type the *conclusions* of the article that is being submitted. The extraction routine uses a formula, which is based on the frequency of occurrence of a term in the title, abstract, keywords, method, results, and objective, and this method weights the terms in the conclusion in order to format them as a relation. The syntactic components with higher weights in the conclusion are candidates for the Antecedent and Consequent of the relation.

After the author validates the relation, the system records it as a record using the format shown in Figure 4, together with the conventional bibliographic metadata and the article full text. In the future, we plan to integrate this prototype with the PKP Open Journal System[5], which is an electronic journal system that is largely used in Brazil. In its current implementation, the prototype processes only the article conclusion.

Some of the steps involved in the processing of the conclusion, i.e., "*The results presented herein emphasize the importance to accomplish systematic serological screening during pregnancy in order to prevent the occurrence of elevated number of infants with congenital toxoplasmosis*" [16], are shown in the following Figures.

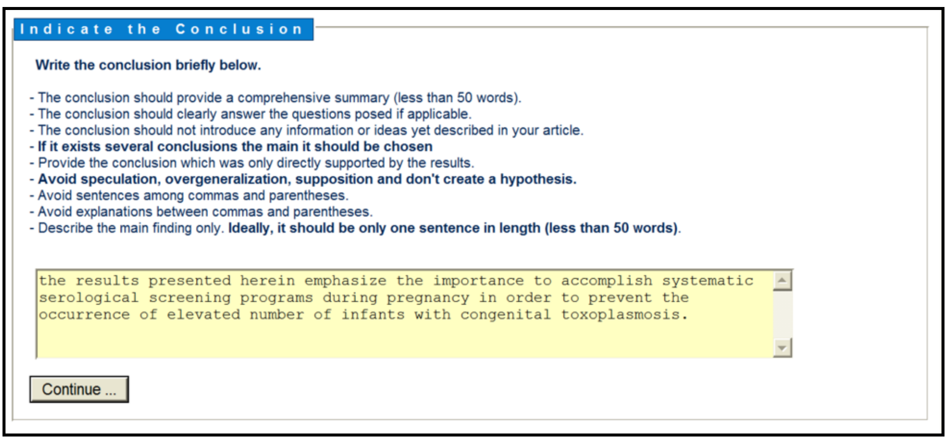

Figure 1. The author specifies the article conclusion.

[3]MetaMap, http://mmtx.nlm.nih.gov/.

[4] Lingpipe home, http://alias-i.com/lingpipe/.

[5] PKP Open Journal System, http://pkp.sfu.ca/

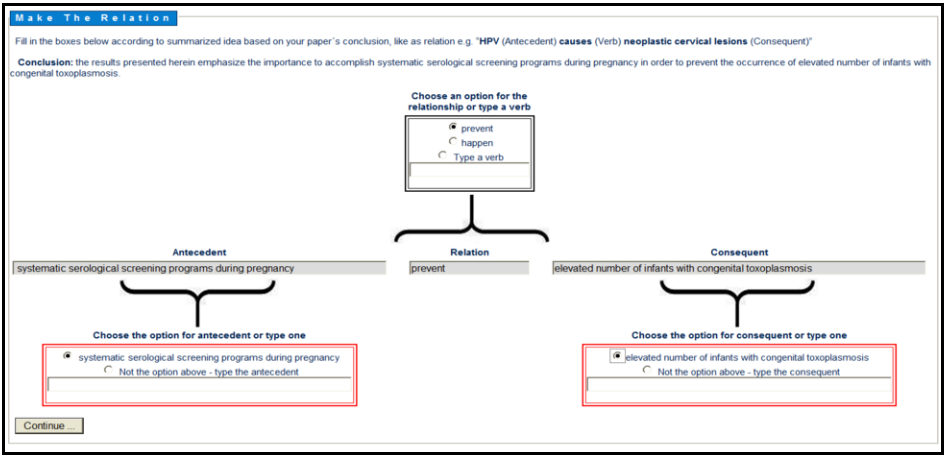

Figure 2. The article conclusion is formatted as a relation.

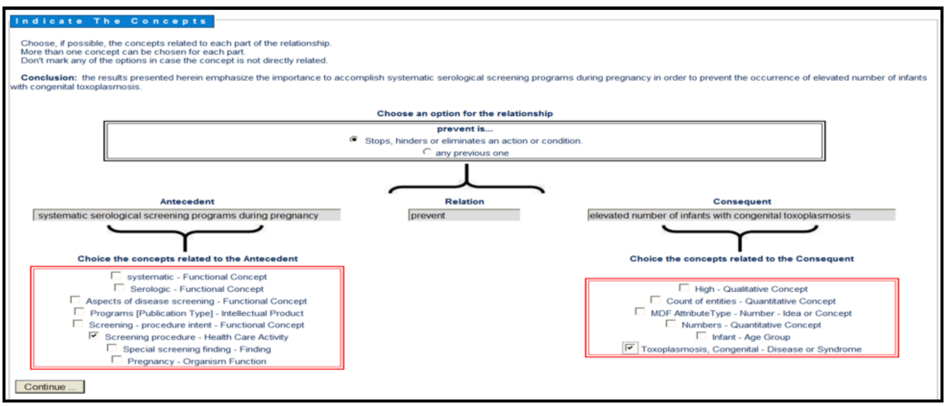

Figure 3. Authors are asked to map/annotate concepts in the article's conclusion to UMLS terms.

2. A Semantic Record Model in RDF for Scientific Articles

A key component in the scope of the proposed publication model is a *semantic record model* for scientific articles that focuses on the semantic elements of the scientific methodology in terms of scientific reasoning, e.g., questions, hypotheses, experiments, and conclusions. These elements are related to the knowledge content and reasoning patterns of an article, which this study aims to identify and record in a machine-processable format. The *Conclusion* is an essential semantic element that synthesizes the knowledge content of an article. In the scope of a recently published article, this is provisional knowledge, although it is verified by the experiment reported in the article. Semantic elements such as Questions and Hypothesis are also important because they enable researchers to trace the evolution of a research question and its resolution in a paper. Other elements have rhetoric functions, as extensively discussed in [17] and [18], or they may serve to describe methodological options more clearly, such as the experiment performed, its context, or the results obtained.

Thus, relations are the core of the proposed knowledge representation scheme. A relation has the form of an Antecedent (a concept referring to a phenomenon), a Semantic Relation, and a Consequent (a concept referring to another phenomenon or a characteristic of the phenomenon in the Antecedent). A Semantic Relation may be a specific Type_of_relation such as "causes," "affects," or "indicates," or a "has_as_characteristic" relation.

Relations may also appear in different semantic elements throughout the article text, such as within the Problem that the article addresses as a *Question* where either one of the two *relations* or the type of relation is unknown, in the *Hypothesis,* or in the *Conclusion*. Frequently, the Conclusion also poses new Questions.

The implementation of a model described using an electronic journal systemsubmission interface poses several challenges, e.g., how to represent the model, even partially, in a machine-understandable format, and how to extract and format a relation from the text of the article conclusion. We began to address these challenges during the development of the prototype.

The following figure considered the conclusion *"telomere replication (Antecedent) involves (Type_of_relation) a terminal transferase-like activity (Consequent),"* found in [19], which has been formatted in RDF.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:sa="http://example.org/semarticles/"
   xmlns:umls="http://www.nlm.nih.gov/research/umls/">
   <rdf:Description rdf:about="http://art_id/">
     <dc:title>title</dc:title>
     <dc:creator>creator</dc:creator>
     <dc:subject>subject</dc:subject>
      <dc:date.published>date</dc:date.published>
      <sa:conclusion>
         <rdf:Description rdf:about="http://art_id/conclusion">
           <sa:antecedent content="telomere
replication">http://www.nlm.nih.gov/research/umls/CUI01</sa:antecedent>
           <sa:type_rel
content="involves">http://www.nlm.nih.gov/research/umls/CUI02</sa:type_rel>
           <sa:consequent conten="a terminal transferase-like activity">
http://www.nlm.nih.gov/research/umls/CUI03</sa:consequent>
     </rdf:Description>
    </sa:conclusion>
   </rdf:Description>
</rdf:RDF>
```

Figure 4. Conclusion of article, represented in RDF. CUI means UMLS's concept unique identifier.

3. Knowledge Network Derivation from a Scholarly Semantic Publishing Environment

The *Conclusion* is an essential semantic element of the article text that synthesizes its knowledge content. Within the scientific community, the agreed mechanism for validating scientific results is the number of citations an article receives. Thus, the Conclusion represents provisional knowledge in recently submitted/published articles without any citations. However, these results are verified by the experiments they report and by the peer review process performed by the journal editorial board that approved the article's publication.

The enhanced bibliographic records found in the proposed model are nodes in a knowledge network with links to full-text articles, conventional bibliographic metadata, the bibliographic references cited, and the article conclusions formatted as RDF triples. Authors are also asked to annotate terms in the conclusion text using terms in a domain terminological knowledge base. Bermes [23] stressed the important role of terminological knowledge bases in a Linked Data environment, because they function as "hubs" that connect resources from different domains. All of these elements may also be identified using URIs, thereby facilitating semantic navigation by scholars, as suggested in the Linked Data proposal [4].

Figure 5 illustrates this process.

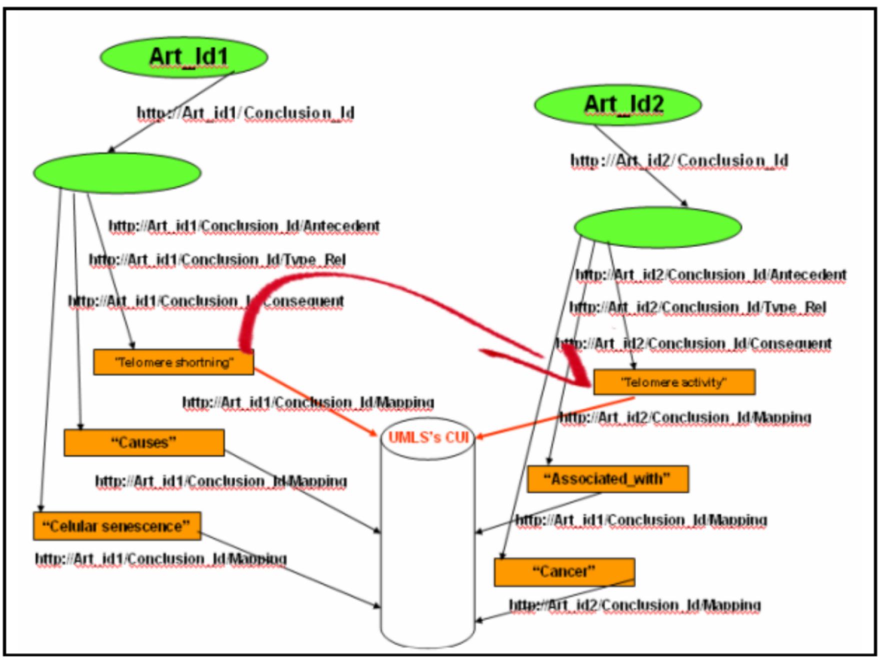

Figure 5. Two articles with related claims connected by a UMLS Semantic Type.

The example shown in Figure 5 is based on a free interpretation of the two claims found in following statement: "*It has been proposed that the finite cell division capacity of human somatic cells is limited by telomere length. This is consistent with reports that telomerase activity is often high in cancer and immortalized tissue culture cells*" [20].

This example illustrates the situation where two articles have related claims. The first article's claim states that "telomere shortening causes cellular senescence," whereas the second article states that "telomerase activity is associated with cancer." The concepts "telomere shortening" and "telomerase activity" are both mapped, i.e., linked, to the same concept in UMLS, which is identified by its Concept Unique Identifier (CUI) as "telomerase activity," which can function as a URI.

Even a partial implementation of the record model proposed in RDF, where the only semantic element captured is the *conclusion,* will facilitate more expressive semantic retrieval from a knowledge network using SPARQL, as shown in Figure 5 [21].

In this example, a software agent might infer a new claim, i.e., that (maybe) "telomere shortening" "is associated with" "cancer". The claim can only be trusted based on the evidence presented in the experiments described in both articles and by the judgement of journal referees, who certified that both articles had sufficient scientific quality to merit publication. The result is a knowledge network of *claims* extracted from refereed scientific articles, which are coded in machine-readable format, and linked to *terminological knowledge bases* [23] and to *citing/cited articles.* Such a network overcomes the problems of conventional citation networks [5] when analyzing scientific developments and managing scientific knowledge. Semantic Web technologies are increasingly used in electronic publishing environments to enhance citation analysis [22], so it is reasonable to expect that a citation network would merge from the proposed generalized knowledge network.

Software agents can navigate through this richer network and perform "inferences" that facilitate sophisticated tasks such as hypothesis formulation, hypothesis comparison, knowledge discovery, and the identification of traces of scientific discoveries and knowledge misunderstandings.

For example, if CUI01 is the UMLS CUI for the relation "Functionally_related_to" and CUI02 is the UMLS CUI for the concept "Telomerase activity." Then a knowledge network, such as that shown in Figure 5, enables SPARQL queries, as follows:

```
@PREFIX sa: <http://example.org/semarticles/>
SELECT ?Consequent
WHERE {type_rel/mapping=”CUI01”}
AND        antecedent/mapping=”CUI02”}
```

We have also shown elsewhere [13] that our proposed semantic publication model facilitates the identification of traces of scientific discoveries. We hypothesize that there is a correlation between the article content and the fact that these articles report scientific discoveries. Based on this approach, we propose a new scientific discovery indicator that is different from bibliographic/citation indicators. Since there is a structural delay between the publication date of an article and the date of citation, this new indicator stresses the importance of such metadata elements for the proposed schema. The proposed approach is based on a content comparison of the conclusions of scientific articles and biomedical terminological knowledge bases, to verify that article conclusion terms can be mapped to terms in a terminological knowledge base. Evidence was found that articles, where just a few terms in the conclusions or even none were mapped or just mapped at a generic level to terms in biomedical terminologies such as UMLS, indicated new discoveries.

We envisage other applications that could enhance current literature-based discovery methods [24], [25] by comparing indirect indicators such as citations or use the same content descriptors with the claims made in articles. The proposed semantic publication model may be part of a future wholly integrated e-science environment.

References

[1] A.H. Renear, C.L. Palmer, Strategic reading, ontologies and the future of scientific publishing, Science 325, (2009), 828-832.

[2] G.C. Marion, An Expanded Review of Information-System Terminology, Proceedings of AFCEA 's Sixteenth Annual Federal Database Colloquium and Exposition, "Information Dominance and Assurance" San Diego, CA, September (1999), 21-23.

[3] Unified Medical Language System Fact Sheet, http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

[4] C. Bizer, T. Heath, T. Berners-Lee, Linked data – the story so far, In: T. Heath, M. Hepp, C. Bizer (eds.), Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS).

[5] E. Garfield, I.H. Sher, R.J. Torpie, The Use of Citation Data in Writing the History of Science, The Institute for Scientific Information, Philadelphia, (1964).

[6] C.H. Marcondes. A semantic model for scholarly electronic publishing. In: International Workshop on Semantic Publication - SePublica2011-, 1st, at the Extended Semantic Web Conference (ESWC), 8th, in Hersonissos, Crete, Greece, Proceedings... CEUR Workshop Proceedings, 721, (2011). ISSN: 1613-0073.

[7] D.L. Miller, Explanation Versus Description, *Philosophical Review* 56 (3) (1947) 306-312.

[8] Ingetraut Dahlberg. Conceptual structures and systematization. International Forum on Information and Documentation 20, (3), (1995) 9-24.

[9] L.Y. Hao et al. Short telomeres, even in the presence of telomerase, limit tissue renewal capacity. Cell 123 (2005) 1121–1131.

[10] J. Shampay, J.W. Szostak, E.H. Blackburn. DNA sequences of telomeres maintained in yeast. *Nature* 310, (1984) 154-157.

[11] C. W. Greider, E. H. Blackburn. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell*, 43 (1985) 405-413.

[12] L.C. da Costa, *Um proposta de processo de submissão de artigos científicos à publicações eletrônicas semânticas em Ciências Biomédicas*, Tese (doutorado), Programa de Pós-graduação em Ciência da Informação UFF-IBICT. Niterói, (2010).

[13] C.H. Marcondes, L.R. Malheiros. Identifying traces scientific discoveries by comparing the content of articles in biomedical sciences with web ontologies. In: 12 ISSI - International Conference on Informetrics and Scientometrics, 2009, UFRJ, Rio de Janeiro, v. 1, *Proceedings…* São Paulo, BIREME/PAHO/WHO, (2009), 173-177.

[14] RDF, http://www.w3.org/TR/rdf-primer/

[15] OWL Ontology Web Language Overview, http://www.w3.org/TR/owl-features/.

[16] G.R.S. Segundo *et al*. A comparative study of congenital toxoplasmosis between public and private hospitals from Uberlândia, MG, Brazil. *Mem. Inst. Oswaldo Cruz*. [online], 99 (1) (2004) 13-17.

[17] J. Skelton. Analysis of the structure of original research papers: an aid to writing original papers for publication. *British Journal of General Practice*, 44, (1994), 455-459.

[18] Kevin Ngozi Nwogu. The Medical Research Paper: Structure and Functions. *English for Specific Purposes,* 16 (2) (1997) 119-138.

[19] C.W. Greider, E.H. Blackburn. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell* 51, (1987) 887--898.

[20] M.J. McEachern, E.H. Blackburn, Runaway telomere elongation cause by telomerase RNA mutations, *Nature* 376 (1995), 403-409.

[21] SPARQL Query Language for RDF (2008), http://www.w3.org/TR/rdf-sparql-query/

[22] The citation ontology, CITO, http://speroni.web.cs.unibo.it/cgi-bin/lode/req.py?req=http:/purl.org/spar/cito

[23] E. Bermes. Convergence and Interoperability: a Linked Data perspective. In: IFLA World Library and Information Congress, 77th. Puerto Rico, 2011, *Proceedings*. (2011)

[24] D.R. Swason, N. R Smalheiser, V. I. Torvik, Ranking indirect connections in literature based discovery. The role of Medical Subject Headings, *Journal of the American Society for Information Science and Technology* 57 (11) (2006), 1427–1439.

[25] R.N. Kostoff, M.B. Briggs, J.L. Solka, R.L. Rushenberg, Literature-related discovery (LRD): Methodology, *Technological Forecasting and Social Change*, 75 (2), (2008), 165–185.