A semantic model for scholarly electronic publishing

Carlos H. Marcondes

University Federal Fluminense, Department of Information Science, R. Lara Vilela, 126, 24210-590, Niterói, Rio de Janeiro, Brazil, marcon@vm.uff.br

Abstract. Despite numerous advancements in information technology, electronic publishing is still based on the print text model. The natural language textual format prevents programs from semantically processing article content. A semantic model for scholarly electronic publishing is proposed, in which the article conclusion is specified by the author and recorded in a machine-understandable format, enabling semantic retrieval and identification of traces of scientific discoveries and knowledge misunderstandings. 89 biomedical articles were analyzed for this purpose. A prototype system that partially implements the proposed model was developed. Four patterns of reasoning and sequencing of semantic elements were identified in the analyzed articles. A content model comprising semantic elements and their sequences in articles is proposed. The development and testing of a prototype of a Web submission interface to an electronic journal system that implements the proposed model are reported.

Key words: electronic publishing, scientific methodology, scientific communication, knowledge representation, ontologies, semantic content processing, e-Science

1 Introduction

Before the advent of the World Wide Web (hereafter referred to as "Web"), man's body of scientific knowledge was fuzzy and distributed across publications in libraries worldwide. The Web is fast becoming a universal platform for the disposal, exchange, and access of knowledge records. An increasing amount of records of human culture—from text, static and motion images, and sound, to multimedia—are now being created directly in a digital format.

With regard to scientific knowledge, one problem is the fact that although a large amount of knowledge can potentially be made available through the Web in digital formats, this knowledge is embedded in the text of scientific articles in natural language that is only comprehensible to humans. Scholarly electronic publishing is based on the print text model. These texts are also distributed across various information resources such as digital libraries, electronic journal systems, and repositories. Their textual format hinders the comparison of their semantic content by computers in order to identify gaps and contradictions and agreements in knowledge.

Metadata is essential for managing knowledge records in an increasingly complex digital environment. Since the MARC (machine-readable cataloging) record was established in the 1960s, bibliographic record models have hardly changed. A typical bibliographic record comprises sets of database fields, including a flat space of a list of unconnected fields for content description, where keywords or descriptors are assigned, each having an equal weight for retrieval purposes. Content access to documents in modern bibliographic retrieval systems is still achieved by matching user queries formed by keywords connected by Boolean operators to keywords comprising the bibliographic records, in a manner similar to early bibliographic retrieval and library automation systems.

A subtle distinction, rarely made by the Library and Information Science Community, must be made between the *aboutness* of a document, a concept that has been exhaustively discussed in this community, and the *claims* made by authors throughout the text of the documents. Indexing activities address the former but not the latter. The extraction and representation in machine-understandable format of claims in scientific article texts should constitute a step toward conventional information retrieval (IR) systems. It should enable direct knowledge management, its use in automatic reasoning and inference tasks applied to different and unpredicted contexts, and increased possibilities of the automatic processing of the rich digital content now available throughout the Web.

Relations between concepts are the core of meaning. Dictionary entries with definitions of terms, thesauri, and classification schemas are examples of this claim. Typical bibliographic records do not hold explicit semantic relations between elements comprising the content of documents they represent. Boolean operators are too general and lack the semantic expressiveness necessary for content retrieval in specific scientific domains. Relations expressed by Boolean operators are processed as extensive set operations on the keywords included in the bibliographic records, and not as intensive semantic relations.

In comparison with the poor expressiveness of the three Boolean operators, the UMLS (unified medical language system) Semantic Network (hereafter abbreviated as "SN") [1], which is the classification schema of the UMLS NIH (National Institutes of Health) Metathesaurus, organizes every concept in hierarchy trees, each having as its root a top level Semantic Type. The UMLS SN uses 54 Relation Types to express the semantic relations used between concepts in Semantic Type hierarchies used to index Biomedical Science scientific articles. The UMLS SN holds the permitted relations between Semantic Types. Although this semantically richer schema is supported by the UMLS, the bibliographic record models in databases such as Medline are incapable of exploiting this potential.

Semantic Web (SW) technologies [2] constitute a step toward semantic retrieval and processing in computational environments. The proposal content of a Web document is no longer a matter of keyword match as in conventional computational environments since the 1960s, but instead comprises structured sets of concepts connected by precise meaning relations as in RDF (Resource Description Framework) [3] and RDF Schema [4] statements. Such a rich knowledge representation schema enables software agents to perform "inferences" and more sophisticated tasks based on the document content.

Since the Actas of the Royal Society in the seventeenth century, scientific articles have become privileged channels of scientific communication. Through scientific articles, authors bring discoveries into the public knowledge. Nowadays, scholars and researchers commonly engage in electronic Web publishing. Most scientific journals are now available on the Web. Modern bibliographic information systems exploit the potential of information technology (IT). However, IT is not yet used to directly process the knowledge embedded in the text of scientific articles. Electronic-Web-published articles can serve as *knowledge bases,* as stressed by Gardin [5]. However, in the digital format, these knowledge bases are useful only to humans, who can read them. The content of scientific articles deserves critical reading, inquiry, and citation through a long social process until it becomes part of man's body of knowledge.

In the present proposal, a richer semantic content bibliographic record model is proposed, in which scientific claims made by authors throughout articles are expressed by relations between phenomena. In the proposed model, each article, in addition to being published in textual format, has its claims also represented as structured relations and recorded in a machine-understandable format using SW standards such as RDF [3] and OWL (Web Ontology Language) [6]. In the proposed model, article records comprise full-text, conventional bibliographic metadata, and semantic metadata conveying the claims made by the author. The machine-understandable records resulting from this publishing model can be compared by software agents either with public knowledge—e.g., published scientific articles—or with terminological knowledge bases throughout the Web, thus providing scientists with new tools for knowledge retrieval, claim comparison, identification of contradictory claims, use of these claims in different contexts, and identification and validation of new contributions to science made by specific articles.

We propose to engage authors in developing a richer content representation of their own articles; bibliographic record instances in compliance with the proposed model will be generated by a Web author's submission interface to a journal system, as a byproduct of submitting his/her articles to the system. Such a system, during the upload process of scientific article files, will perform an interactive dialog with authors in order to extract the semantic content of the claims made in the scientific articles and record them in a machine-readable format. We also report the initial steps toward the development of such a system.

Several alternatives have already been proposed as new types of publications that address the previously discussed issues; to try and exploit SW technologies to enhance scientific communication, management, sharing, and reuse of knowledge; and to provide direct access to semantic content of scientific articles. Thus, there is an increasing trend in electronic publishing experiences toward formalizing the text of articles or structuring them, marking them, and identifying significant parts to facilitate more direct reading by humans, potentially by relating the text to formal ontologies [7] as a means to overcome the ambiguity of the texts and allow their "semantic" processing by programs.

The remainder of this article is organized as follows. The next section presents a review of the theoretical concepts the proposed publication model is based on along with similar experiences and projects. Section 3 describes the materials and methods used. Section 4 describes the model, its elements, and the development of a prototype system of a Web author's submission interface to a journal system, which partially implements the model. Finally, section 5 presents the results obtained thus far and discusses the conclusions. It also outlines the future research steps.

2 Related studies

From an ontological point of view, scientific articles are (a) documents embedded in definite social relations concerning the scientific communication protocols exhaustively studied in Information Science [8], [9], and with regard to their textual structure, (b) a text-embedded rhetoric/logical theory [10], [11], [12]. The focus of the proposed model is the second aspect, i.e., the reasoning/rhetorical, and the semantic structure of the scientific articles in Biomedical Sciences.

In this field in particular, new research methods challenge the conventional Scientific Method and Popperian hypothesis-driven research. The so-called high-throughput methods like DNA microarrays and proteomics [13] allow scientists to process a great amount of data rapidly and in parallel, thus "conducting experiments about which no predictions can be made because no hypotheses have been constructed," as stressed by Westein [14]. This author also stresses the following:

> "Given the layered, evolutionary complexity of biological systems, it will not be possible to understand them comprehensively on the basis of hypothesis-driven research alone. Likewise, it will not be possible to do so solely through "omic" studies of genes, proteins, and other molecules in aggregate. The two modes of research are complementary and synergistic".

Several alternatives have been considered as new types of publications to address the previously discussed issues and to exploit SW technologies to enhance scientific communication, management, sharing, and reuse of knowledge, and to provide direct access to the semantic content of scientific articles. The following text comments on these experiences and their conceptual bases.

The Prospect project is a publishing initiative of the Royal Society of Chemistry, in which terms in the texts of articles that refer to chemical or biological entities have links to dictionaries or ontologies that define them. The Elsevier publishing group is developing a project called Article of the Future associated with the biomedical journal Cell in order to add functionality to several articles, including change in presentation (hierarchical presentations), summary charts, and a section on "Highlights" that briefly outline the conclusions of the article. These facilities are only possible in a Web environment for digitally published articles. Sample articles are available on the project Web site to demonstrate these facilities. A previous study [15] has described the experience of using different semantic technologies in the journal PloS, including biomedical ontologies, comments on the articles, and an ontology of types or reasons for citation.

HyBrow [16] is a system aimed at helping scientists with hypothesis formulation and evaluation against previous knowledge. The work by Hunter and co-authors [17] aimed to identify concepts for extracting protein interaction relations from biomedical text. The approach of [18] to semantic annotations in medical articles considers assertions to be the fundamental units of knowledge. The HypER approach [19] also considers claims to be the basic unit of scientific knowledge. Groth and colleagues [20] present a publication model called nanopublications, consisting of core scientific statements associated with their annotations which specify their context; scientific statements are coded as RDF triples. The Utopia project [21] proposed the assignment of semantic comments to articles.

A growing number of scientific publications, especially in the biomedical area, such as the BMJ (British Medical Journal) and the JAMA (Journal of American Medical Association), have been using structured abstracts [22] as a way to optimally extract the contents of articles.

3 Materials and methods

- The domain of biomedical sciences was chosen because scientific articles in this area follow a strict formal pattern in their texts, with sections defined according to a standard called IMRAD (Introduction, Method, Results, and Discussion).
- 89 articles in biomedical sciences were analyzed to develop the model with the aim of identifying the semantic elements of scientific methodology, reasoning patterns, and sequencing that combine these elements.

Articles analyzed comprise 3 groups.

- articles from two outstanding Brazilian research journals, 20 articles from the Memórias do Instituto Oswaldo Cruz, which has its scope mainly in Microbiology, (published during the period 1999-2004), 20 articles from the Brazilian Journal of Medical and Biological Research (published during the period 1998-2004).

- 20 articles about stem cells were also analyzed (published during the period 1994-2004). Stem cells, as an emerging research area in rapid development, were chosen expecting to find articles reporting important discoveries. The articles analyzed were selected from three reviews which present stem cell research development in a historical perspective, pointing out the advances in research, thus of special interest for our work.

- 29 articles from the Albert Lasker Basic Medical Research Award 2006 key publications were analyzed. This last group is of special interest to the objectives of this research because the articles report, step by step, the rise of new scientific discovery, the discovery of telomerase enzyme since 1978 - the first article - to 2001 - the last article of this group. The analysis of this group of articles was guided by an article [23] by the three winners of Lasker Award 2006 which comments the steps toward the discovery of telomerase enzyme.

- Each article was analyzed in 4 steps: (1) identify patterns of reasoning developed throughout the article; (2) identify the main conclusion posited by the author in the text; (3) format the claim made in the conclusion as a relation according to the proposed knowledge representation format; and (4) tentatively map each element of the relation to concepts in the UMLS/UMLS SN. Mapping is achieved by comparing terms in the relation extracted in step 3 to MeSH/UMLS terms indexing the article in PubMed records.

- A prototype of a submission interface to an electronic journal system was developed, which formats the natural language text of conclusions of articles submitted by authors as semantic relations; this was developed using MetaMap [24], a program that processes biomedical texts to identify terms from the UMLS Thesaurus.

4 Results and discussion

We have been working for years [25] on the development of a semantic model of electronic publishing. The aim of this model is to achieve a semantically richer content surrogate of biomedical articles in a program "understandable" format. Such a knowledge representation format allows programs to extract "inferences" about the knowledge content of articles, enabling semantically powerful content retrieval and management relative to current bibliographic IR Systems. The proposed model comprises two components: a semantic content model and a Web interface for authors self-publishing and self-submitting articles to a journal system. The semantic content model *extends* conventional bibliographic record models, which comprise conventional descriptive elements such as authors, title, bibliographic source, and publication date together with content information such as keywords or descriptors. Scientific claims made by authors in their papers are represented as *relations* between two different phenomena or between a phenomenon and its characteristics [26]. Our study also includes the development of a prototype system of a Web author's submission interface to a journal system, which implements the model [27] and the use of the general framework proposed to identify discoveries in scientific papers based on two aspects: their rhetoric elements and formats and by comparing the content of the conclusion of articles with terminological data banks [28]. This last aspect corresponds to step 4 of the analysis process described in section 2 and to the task performed by authors as illustrated in Figure 5.

The following figure shows an overview of the semantic model of electronic publishing, which includes the following components: the Web interface to a system for the submission of articles to electronic publications, the Database, the public Web knowledge base, and the Discoveries identification tool.

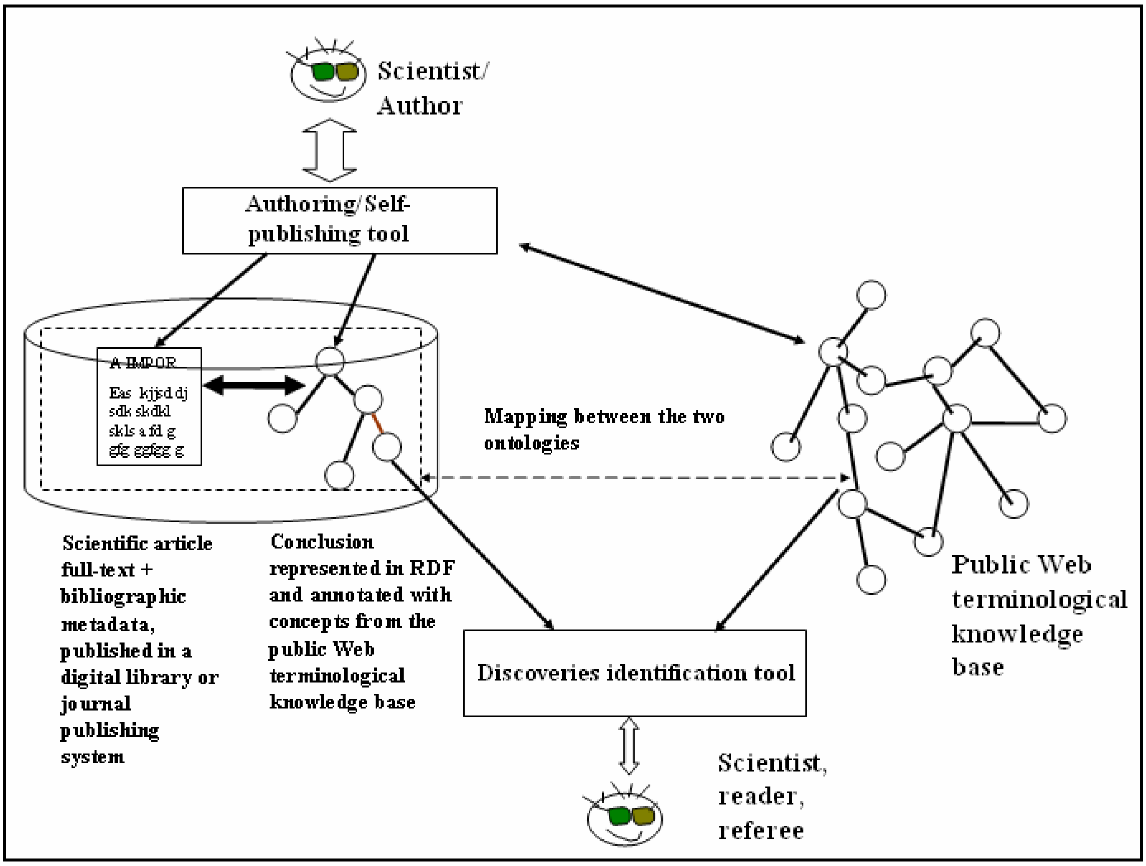

Scientist/ Author
Authoring/Self-publishing tool
A IMPOR
Eas kjjsd dj sdk skdkl skls a fd g gfg ggfgg g
Mapping between the two ontologies
Scientific article full-text + bibliographic metadata, published in a digital library or journal publishing system
Conclusion represented in RDF and annotated with concepts from the public Web terminological knowledge base
Public Web terminological knowledge base
Discoveries identification tool
Scientist, reader, referee

Fig. 1. Overview of the components of the semantic publication model

4.1. A semantic content model for electronic publishing

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 a 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/have) characteristic relation. Examples of knowledge representation according to this schema are the following:

- Tetrahymena extracts (Antecedent) have (Characteristic) a specific telomere terminal transferase activity (Consequent).
- Telomere shortening (Antecedent) causes (Type_of_relation) cellular senescence (Consequent).

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

Questions, *Hypothesis*, and *Conclusion* are the semantic elements comprising the proposed model. They are the elements related to the knowledge content of an article, which we aim 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, it is provisional knowledge; however, it is at least guaranteed by the experiment reported in the article. Semantic elements such as *Questions* and *Hypothesis* are important because they enable the evolution of a claim to be determined. Other elements have rhetoric functions, as extensively discussed in [29] and [30], or serve to describe methodological options, the experiment performed, its context, or the obtained results more clearly.

In Biomedical Sciences, there are some standardized methodological procedures, such as PCR (polymerase chain reaction), and some standardized contexts where experiments can take place, for example, in humans (e.g., children, women, embryo), rats, etc.

The semantic elements that comprise the proposed record model are as follows:

- the problem the article is addressing and the **question** derived from it,
- an **antecedent**,

- a **type_of_relation** (holding the semantic of the relation in a domain, for example, in Biomedical Sciences),

- and the **consequent**.

The **antecedent** and **consequent** may be two different phenomena or a phenomenon and its characteristics.

A possible empirically controlled **experiment** with the aim of observing the phenomenon described and specifics of experimental articles are divided into

- **results** – tables, figures, and numeric data reporting the observations made;
- **measure** used;
- a specific **context** where the empirical observations take place, subdivided into:
    - **environment** – a hospital, a daycare center, a high school,
    - a geographical **place** where the empirical observations take place,
    - **time** when the empirical observations occur,
    - a specific **population** – pregnant women, early born babies, mice – in which the phenomenon occurs,
    - **conclusion** – a set of propositions made by the author as a result of his/her findings.

A **conclusion** corroborates totally or partially the **hypothesis** of an article or negates it. A **conclusion** may also be conclusive or not yet conclusive.

In every analyzed article, concepts found in the antecedent, type_of_relation, and consequent were tentatively mapped (and will be annotated in the future web authoring/publishing tool) to concepts taken from the UMLS. Not all elements are present in all articles.

Articles differ in the way they are built around previously stated hypotheses—those stated by authors other than the author of the current article, or new, original hypotheses, i.e., those stated by the author of the current article. Articles may also differ by the existence of a documented experiment or simply theoretical considerations comparing previously stated hypotheses. We found four patterns of reasoning in the analyzed articles: *theoretical articles*, which employ abductive reasoning and *experimental articles,* which may simply be *exploratory* or employ *inductive* or *deductive* reasoning.

Theoretical-abductive (TA) articles analyze different, previous hypotheses, showing their faults and limitations and proposing a new hypothesis; the reasoning is as follows:

*A **problem** is identified, with the following aspects and data…;*

*The **previous hypotheses** (from other authors) are not satisfactory to solve the problem due to the following criticism…;*

*Therefore, we propose this **new hypothesis** (original), which we consider a new pathway to solve the problem.*

Experimental-inductive (EI) articles propose a hypothesis and develop experiments to test and validate it; the reasoning is as follows:

*A **problem** is identified, with the following aspects and data…;*

*A possible solution to this **problem** can be based on the following new **hypothesis…**;*

*We developed an **experiment** to test this **hypothesis** and obtained the following **results**.*

In experimental-inductive articles, a **conclusion** may be mainly one of these alternatives: it corroborates the hypothesis, refutes it, or partially corroborates the hypothesis. However, in some cases, the Conclusion is not one of the former; it simply reports intermediate, and not conclusive, results toward the hypothesis corroboration.

Experimental-deductive (ED) articles use a hypothesis proposed by other researchers cited by the articles' author and apply it to a slightly different context; the reasoning is as follows:

*A **problem** is identified, with the following aspects and data…;*

*In the literature, the **previous hypotheses** (by other authors) have been proposed…;*

*We choose the following **previous hypothesis…**;*

*We enlarge and recontextualize this **hypothesis**; we develop an **experiment** to test it in this new context…;*

*The **experiment** shows the following **results** in this new **context**.*

Experimental-exploratory (EE) articles are not usually hypothesis driven; their objective is to acquire knowledge about a poorly understood scientific phenomenon by performing an **experiment**; the reasoning is as follows:

There is a phenomenon that is poorly understood in a scientific domain.

*We developed an **experiment** that permits the identification of the following characteristics of this phenomenon.*

Within the group of 89 articles that were analyzed, we classified 27 as experimental-inductives (EI), 44 as experimental-deductives (ED), 15 as experimental-exploratories (EE), and 3 as theoretical-abductives (TA).

These basic semantic elements of scientific articles are interrelated and structured. Together with the corresponding bibliographic metadata and article full-text, they form richer article surrogates in machine-understandable formats and constitute single digital objects stored in a digital library or electronic journal publishing system.

The different reasoning semantic elements and reasoning procedures discussed previously can be formalized in the Model of Knowledge in Articles (MKA), as illustrated in Figure 2 with the hierarchy of classes and properties.

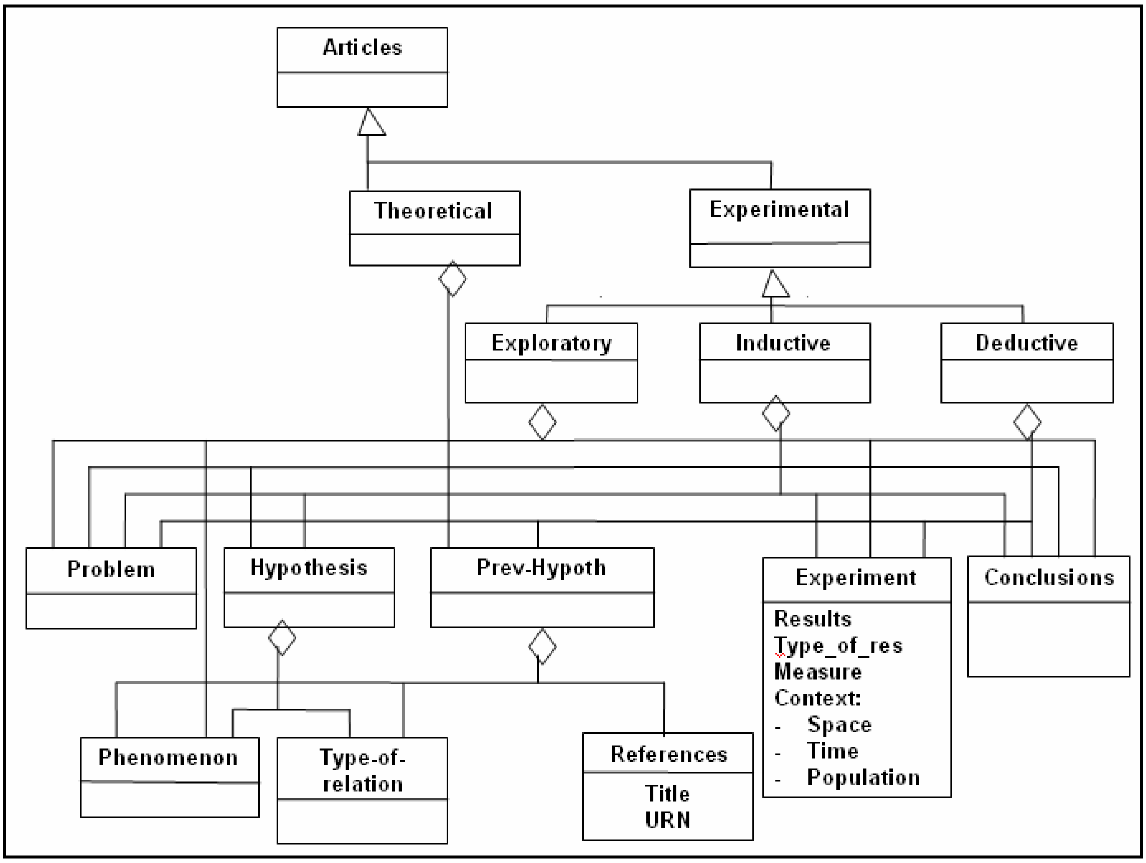


Fig. 2. MKA: model of knowledge representation in articles

The proposed knowledge representation framework enables the following types of queries to a semantic information retrieval system:

- Which other articles have hypotheses suggesting HPV as the cause of cervical neoplasias in women?

- Which articles have hypotheses suggesting other causes of cervical neoplasias different from HPV in women?

- Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in groups different from women?

- Which articles have hypotheses suggesting HPV as the cause of pathologies different from neoplasias?

- Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in different contexts (not in women from the Federal District, Brazil)?

The model also enables queries that may indicate new discoveries, for example, new causes for cellular senescence:

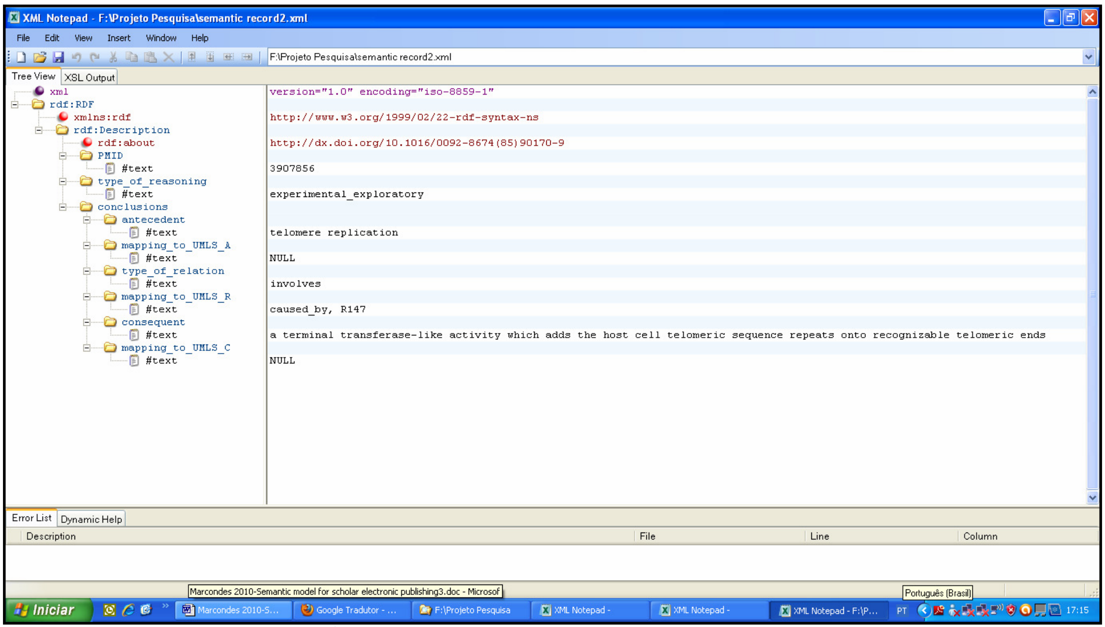
- Which experimental-inductive articles propose (Antecedent?) causes (Type_of_relation) for cellular senescence (Consequent) that are not mapped to UMLS concepts?
- Is there any confirmation of the hypothesis that “Several aspects of both the structural and dynamic properties of telomeres (Antecedent) led to the proposal that telomere replication

involves (Type_of_relation) nontemplate addition of telomeric repeats onto the ends of chromosomes (Consequent)?" [31]?

- Who and when first maintained that "the RNA component of telomerase (Antecedent) may be directly involved in (Type_of_relation) recognizing the unique three-dimensional structure of the G-rich telomeric oligonucleotide primers (Consequent*)*" [32]?

Previous examples show how the proposed knowledge representation schema may improve semantic retrieval and the use of knowledge in different and unpredicted contexts.

The implementation of the model described in a Web submission interface to an electronic journal system poses the following different challenges: representing the model, even partially, in a machine-understandable format, and extracting and formatting a relation from the article conclusion. We address these challenges as follows. We opt for an initial and partial implementation of the model of content in articles in RDF as it enables semantic retrieval using SPARQL. The following figure shows as the conclusion "telomere replication (Antecedent) involves (Type_of_relation) a terminal transferase-like activity which adds the host cell telomeric sequence repeats onto recognizable telomeric ends (Consequent)," found in [32], which is implemented in RDF format.



```
xml                            version="1.0" encoding="iso-8859-1"
rdf:RDF
  xmlns:rdf                    http://www.w3.org/1999/02/22-rdf-syntax-ns
  rdf:Description
    rdf:about                  http://dx.doi.org/10.1016/0092-8674(85)90170-9
    PMID
      #text                    3907856
    type_of_reasoning
      #text                    experimental_exploratory
    conclusions
      antecedent
        #text                  telomere replication
      mapping_to_UMLS_A
        #text                  NULL
      type_of_relation
        #text                  involves
      mapping_to_UMLS_R
        #text                  caused_by, R147
      consequent
        #text                  a terminal transferase-like activity which adds the host cell telomeric sequence repeats onto recognizable telomeric ends
      mapping_to_UMLS_C
        #text                  NULL
```

Fig. 3. Conclusion of article, represented in RDF

4.2. Web submission interface to an electronic journal system

We developed a prototype of the submission system to evaluate the dialog with authors and the extraction routine. In the future, we plan to integrate this prototype with the PKP Open Journal System [33], an electronic journal system largely used in Brazil. In its present implementation, among the semantic elements that comprise the content model, the prototype processes only the conclusion.

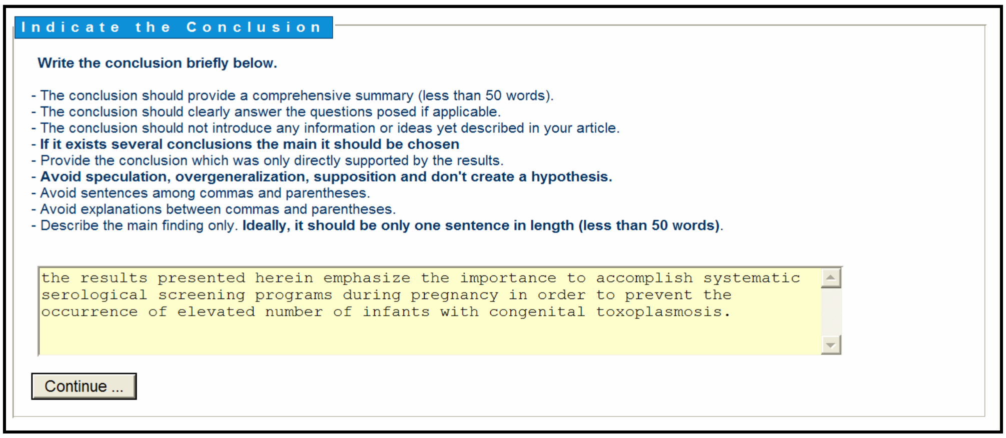
This prototype processes selected parts of the text, namely, the title, abstract, keywords, introduction, methods, and results; the introduction and abstract are used to extract the objective of the article through the identification of phrases such as *objectives of our work…* and *The goal of the present work…* The author is asked by the system to enter the conclusion of the article 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, to weigh terms in the conclusion in order to format it from a textual format to a relation. The syntactic components found in the conclusion with higher weights are candidates for the Antecedent and Consequent of the relation. The Antecedent and Consequent must not be consecutive. The identification of a Relation requires the use of a dictionary that relates the 54 UMLS relations to a set of verbs with the same meaning, obtained from Wordnet (2010) [34].

The systems interacts with authors as follows: (1) authors are asked to enter conventional bibliographic metadata; (2) authors are asked to upload a file with article full-text; (3) authors are asked to choose the type of reasoning used in the article, either theoretical or experimental; (4) authors are asked to validate the article objective extracted by the system; (5) authors are asked to specify the conclusion of the article; (6) after identifying its elements, the article conclusion is formatted as a relation and authors are asked to validate the Antecedent, Relation, and Consequent prompted by the system; (7) authors are asked to map concepts in the article's conclusion to UMLS terms.

After the author validates the Relation, the system records it as an instance of the MKA according to the format illustrated in Fig. 3, together with the conventional bibliographic metadata and the article full-text.

Some of the steps described above when processing the conclusion "*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*" are shown in the following Figures.



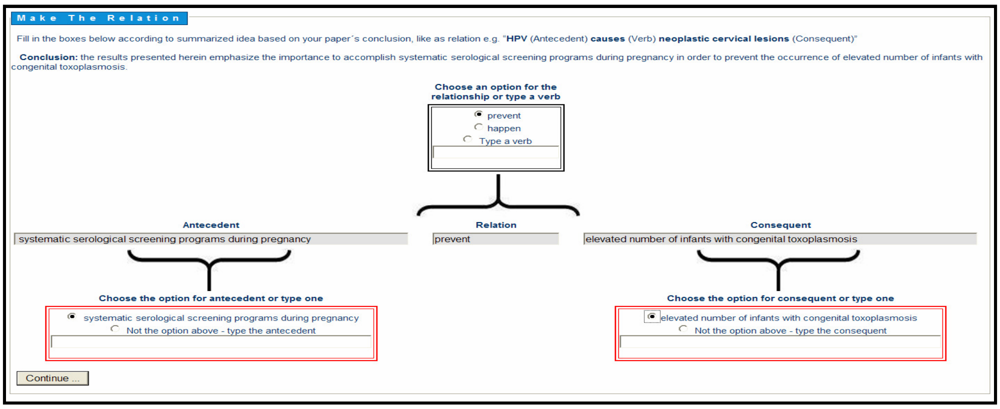
Indicate the Conclusion

Write the conclusion briefly below.

- The conclusion should provide a comprehensive summary (less than 50 words).
- The conclusion should clearly answer the questions posed if applicable.
- The conclusion should not introduce any information or ideas yet described in your article.
- **If it exists several conclusions the main it should be chosen**
- Provide the conclusion which was only directly supported by the results.
- **Avoid speculation, overgeneralization, supposition and don't create a hypothesis.**
- Avoid sentences among commas and parentheses.
- Avoid explanations between commas and parentheses.
- Describe the main finding only. **Ideally, it should be only one sentence in length (less than 50 words).**

```
the results presented herein emphasize the importance to accomplish systematic
serological screening programs during pregnancy in order to prevent the
occurrence of elevated number of infants with congenital toxoplasmosis.
```

Fig. 4. Author specifies the article conclusion



Make The Relation

Fill in the boxes below according to summarized idea based on your paper´s conclusion, like as relation e.g. "**HPV** (Antecedent) **causes** (Verb) **neoplastic cervical lesions** (Consequent)"

Conclusion: the results presented herein emphasize the importance to accomplish systematic serological screening programs during pregnancy in order to prevent the occurrence of elevated number of infants with congenital toxoplasmosis.

Choose an option for the relationship or type a verb

● prevent
○ happen
○ Type a verb

Antecedent: systematic serological screening programs during pregnancy

Relation: prevent

Consequent: elevated number of infants with congenital toxoplasmosis

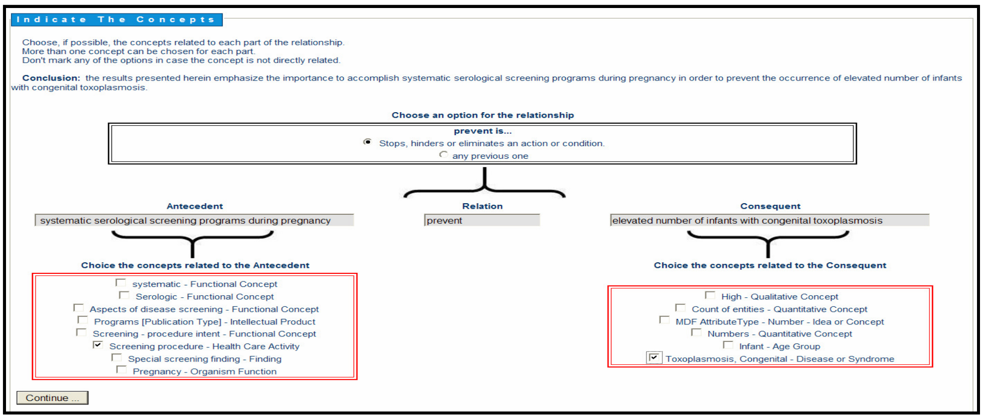
Choose the option for antecedent or type one

● systematic serological screening programs during pregnancy
○ Not the option above - type the antecedent

Choose the option for consequent or type one

● elevated number of infants with congenital toxoplasmosis
○ Not the option above - type the consequent

Fig. 5. The article conclusion is formatted as a relation



Indicate The Concepts

Choose, if possible, the concepts related to each part of the relationship.
More than one concept can be chosen for each part.
Don't mark any of the options in case the concept is not directly related.

Conclusion: the results presented herein emphasize the importance to accomplish systematic serological screening programs during pregnancy in order to prevent the occurrence of elevated number of infants with congenital toxoplasmosis.

Choose an option for the relationship

prevent is...
● Stops, hinders or eliminates an action or condition.
○ any previous one

Antecedent: systematic serological screening programs during pregnancy

Relation: prevent

Consequent: elevated number of infants with congenital toxoplasmosis

Choice the concepts related to the Antecedent

☐ systematic - Functional Concept
☐ Serologic - Functional Concept
☐ Aspects of disease screening - Functional Concept
☐ Programs [Publication Type] - Intellectual Product
☐ Screening - procedure intent - Functional Concept
☑ Screening procedure - Health Care Activity
☐ Special screening finding - Finding
☐ Pregnancy - Organism Function

Choice the concepts related to the Consequent

☐ High - Qualitative Concept
☐ Count of entities - Quantitative Concept
☐ MDF AttributeType - Number - Idea or Concept
☐ Numbers - Quantitative Concept
☐ Infant - Age Group
☑ Toxoplasmosis, Congenital - Disease or Syndrome

Fig. 6. Authors are asked to map concepts in the article's conclusion to UMLS terms

The prototype of the interface is in its initial phase of development. In addition to the 10 interviews, the prototype was tested with 5 of the 10 authors and in all cases, it was able to format a second relationship from the conclusion of the article.

5. Conclusions

Nowadays, researchers are accustomed to publishing and describing their papers themselves when submitting them to a digital library, conference management system, digital repository, or journal system. We consider the submission of an article to a journal system to be a privileged process during which authors are particularly motivated to clarify and disambiguate questions about their articles. The pathway that seems more feasible to reach this objective is to provide authors with an interactive interface that enables them to validate the automatic natural language processing carried out by the system. Some elements of the proposed model can be directly obtained by asking questions of the authors, such as whether the article is theoretical or experimental, whether the conclusion confirms or denies the hypotheses, and whether the article is based on the hypothesis of other authors or is original.

After the claims made by an author from anywhere in the article text, for example, the conclusion, are extracted, they will be represented in a structured form as relations. All these semantic elements can be added to conventional bibliographic elements such as the title, author, abstract, publication data, abstract, and key words, forming richer article surrogates. This knowledge content will then be represented in a standard machine-understandable format such as RDF. Articles published according to the model proposed can be interlinked and have their content annotated with an increasing number of Web public ontologies, forming a rich knowledge network. This will enable software agents to help scientists to identify and validate new discoveries in Science by comparing the knowledge content of articles with the knowledge content held in public knowledge bases such as the UMLS.

Although relations play a key role in scientific knowledge, conventional indexing languages do not take them into consideration. The inclusion of relations in knowledge representation makes an expressive difference [35] by enhancing meaning and making more precise the role of subject headings used to represent the document content.

The inclusion of articles conclusions formatted as relations to enhance article metadata is just a proposal. The prototype developed aims at testing it feasibility. The complete article record lay-out is under development.

The body of scientific literature published on the Web is becoming increasingly vast and complex. It will be necessary for scientists to have enhanced software tools in order to make inferences based on this content. Library and Information Science can go beyond conventional indexing techniques to provide fast access to full-text scientific articles. This would help scientists to directly process the knowledge content of scientific articles and to recover the reasoning that leads to a scientific discovery. The proposed model also recommends the standardization of an SkML (Scientific Knowledge Markup Language) encompassing the knowledge content of scientific articles published on the Web, as also proposed by other studies [36], [37], [38]. This opens a new perspective in scientific electronic publishing, knowledge acquisition, storage, processing, and sharing. The proposed model depends on the development of software tools that are not available yet. Our research group has not been able to fully develop the model to the potentialities outlined here. The proposed model should, however, serve as a starting point that can be discussed and built upon by the scientific community.

Acknowledgments

This research was supported, at different times, by CNPq, CAPES, FAPERJ, and PROPPi/UFF. We would also like to thank Marília Alvarenga Rocha Mendonça, Luciana Reis Malheiros and Leonardo Cruz da Costa.

References

1. UMLS Semantic Network, http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html
2. Berners-Lee, T., Hendler, J., Lassila, O. The semantic web, Scientific American. (2001)
3. RDF Resource Description Framework, http://www.w3.org/RDF/ (accessed 10 Jan. 2007)
4. RDF Schema Specification, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/
5. Gardin, J-C. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. In: Filtrage et Résumé Automatique de l'Information sur les Reseaux - Actes du 3ème Colloque du Chapitre Français de l'ISKO. Paris, Université de Nanterre-Paris X (2001)
6. OWL Ontology Web Language Overview, http://www.w3.org/TR/owl-features/
7. Renear, A. H., Palmer, C. L. Strategic reading, ontologies and the future of scientific publishing. Science 325, pp. 828--832 (2009)
8. Frohmann, B. Documentation redux: Prolegomenon to (another) philosophy of information. Library Trends 52, (3) pp. 387--407 (2004).
9. Cronin, B. Scholarly communication and epistemic cultures. Journal New Review of Academic Librarianship 9, (1) pp. 1--24 (2003)
10. Bezerman, C. Shaping written knowledge: Rhetoric of the human sciences. Madison, The University of Wisconsin Press (1988)
11. Gross, A. G. The Rhetoric of Science. Cambridge, Massachusetts; London: Harvard University Press (1990)
12. Hutchins, J. On the structure of scientific texts. In: Proceedings of the 5th. UEA Papers in Linguistics, Norwich pp. 18--39. Norwich, University of East Anglia (1977)
13. Franklin, L. R. Exploratory Experiments. In: Philosophy of Science Assoc. 19th Biennial Meeting - PSA2004: Contributed Papers, Austin, TX; 2004. Austin, Texas (2004)
14. Weinstein, J. N. 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. Current Opinion in Pharmacology 2, (4) pp. 61--65 (2002)
15. Shotton, D., Portwin, K., Klyne, G., Miles, A. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. PLoS Comput. Biol. 5, (4) (2009)
16. Racunas, S. A., Shah, N. H., Albert I., Fedoroff, N. V. HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics 20, (1) pp. 257--264 (2004)
17. Hunter, L., Baumgartner, W. A., Lu, Z., Johnson, H. L., Caporaso, J. G., Paquette, J., Lindemann, A., White, E. K., Medvedeva, O., Cohen, K. B. Concept recognition for extracting protein interaction relations from biomedical text. Genome Biol. 9 (Suppl 2), (2008)
18. Dinakarpadian, D., Lee, Y., Vishwanath, K., Lingambhotla, R. MachineProse: An ontological framework for scientific assertions. Journal of the American Medical Informatics Association 13, (2) pp. 220--232 (2006)
19. De Waard, A., Buckingham Shum, S., Carusi, A., Park, J., Samwald, M., Sandor, Á. Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. In: Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science. Springer Verlag Berlin, Washington DC (2009)
20. Groth, P., Gibson, A., Velterop, J. The anatomy of a nanopublication. Information Services & Use 30, pp.51--56 (2010)
21. Attwood, T. K., Kell, D. B., Mcdermott, P., Marsh, J., Pettifer, S. R., Thorne, D. Calling international rescue: knowledge lost in literature and data landslide! Biochemical Journal, Dec (2009)
22. Guimarães, C. A. Structured abstracts: Narrative review. Acta Cirúrgica Brasileira, 21, (4) (2006)
23. Blackburn, E. H, Greider, C. W., Szostak, J. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. Nature 12 (10), pp.1133--1138 (2006)
24. MetaMap, http://mmtx.nlm.nih.gov/
25. Marcondes, C. H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: ICCC ElPub - International Conference on Electronic Publishing, Leuven, Bélgica, 2005, 9, Leuven, Bélgiun pp. 119--127. Peeters Publishing, Leuven (2005)
26. Dahlberg, I. Conceptual structures and systematization. International Forum on Information and Documentation 20, (3) pp. 9--24 (1995)
27. Costa, L. C. 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)
28. Marcondes, C. H., Malheiros, L. R. 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, Rio de Janeiro, v. 1. pp. 173--177. São Paulo, BIREME/PAHO/WHO, UFRJ (2009)

29. Skelton, J. Analysis of the structure of original research papers: an aid to writing original papers for publication. British Journal of General Practice, 44, pp. 455--459 (1994)
30. Nwogu, K. N. The Medical Research Paper: Structure and Functions. English for Specific Purposes 16, (2) pp. 119--138 (1997)
31. Shampay, J., Szostak, J. W., Blackburn, E. H. DNA sequences of telomeres maintained in yeast. Nature 310, pp. 154-157 (1984)
32. Greider, C. W., Blackburn, E. H. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51, pp. 887--898, (1987)
33. PKP Open Journal System, http://pkp.sfu.ca/
34. WordNet. A lexical database for English, http://wordnet.princeton.edu/
35. Kajikawa, Y, Abe, K., Noda, S. Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords. Journal of Information Science 32, pp. 511--524 (2006)
36. Murray-Rust, P., Rzepa, H. S. Chemical Markup, XML and the World Wide Web. I: Basic principles, Journal of Chemical Information and Computer Science 39, pp. 928--942 (1999)
37. Hucka, M., Finney, A., Suro, H., Bolouri, H. System Biology Markup Language (SBML) Level 1: Structures and facilities for basic model definitions. (2003)
38. Murray-Rust, P., Rzepa, H.S. STMML. A markup language for scientific, technical and medical publishing, Data Science Journal 1, (2), pp. 128--193 (2002)