Representing and Coding the Knowledge Embedded in Texts of Health Science Web Published Articles

Carlos Henrique Marcondes [1]; Marília Alvarenga Rocha Mendonça [1]; Luciana Reis Malheiros [2]; Leonardo Cruz da Costa [3]; Tatiana Christina Paredes Santos[4]; Luciana Guimarães Pereira[5]

[1] Department of Information Science
e-mail: marcon@vm.uff.br; e-mail: mariliaalvarenga@terra.com.br;
[2] *Department of Physiology and Pharmacology*
e-mail: malheiro@vm.uff.br
[3] *Department of Computer Science*
e-mail: leo@dcc.ic.uff.br
[4] *Biomedicine student*
e-mail: tatianacps.uff@gmail.com
[5] *Library and Information Science student*
e-mail: lucianaguipe@yahoo.com.br
Universidade Federal Fluminense
R. Miguel de Frias, 9 – Icaraí
24220-008 - Niterói – RJ Brazil

Abstract

Despite the fact that electronic publishing is a common activity to scholars, electronic journals are still based in the print model and do not take full advantage of the facilities offered by the Semantic Web environment. This is a report of the results of a research project with the aim of investigating the possibilities of electronic publishing of journal articles both as text for human reading and in machine readable format recording the new knowledge contained in the article. This knowledge is identified with the scientific methodology elements such as problem, methodology, hypothesis, results, and conclusions. A model integrating all those elements is proposed which makes explicit and records the knowledge embedded in the text of scientific articles as an ontology. Knowledge thus represented enables its processing by intelligent software agents The proposed model aims to take advantage of these facilities enabling semantic retrieval and validation of the knowledge contained in articles. To validate and enhance the model a set of electronic journal articles were analyzed.

Keywords: electronic publishing; scientific communication; semantic web; knowledge representation; ontologies

1 Introduction

Nowadays, electronic Web publishing is a common activity to scholars and researchers. However, scientific communication is still a slow social process which largely depends on discourse, text producing, reading/interpreting/inquiring and peer-reviewing by scholars until new knowledge is incorporated into the corpus of Science. The potential of new information technology (IT) has been applied to modern bibliographic information systems to improve scientific communication, providing fast notification and immediate access to full-text scientific documents. But IT is not yet used to directly process the knowledge embedded in the text of scientific articles.

Semantic Web Initiative is a future vision of the Internet which aims to structure today's vast Web content, adding semantic to this content [1]. The technologies and methodologies that have been developed in the context of Semantic Web will enable this content to be understandable not only by people but also by software agents, enabling them to *reason* on this content in achieving different intelligent and useful tasks. In the Semantic Web context, electronic publishing can be a cognitive tool with potential that is far from being explored. Today electronic journals are still based on the print mode. Electronic Web published articles are knowledge bases, but for human reading.

Before the rise of the Web, what constitutes the accented scientific knowledge of humanity was fuzzy, lacks formalization, and was scattered across journals collections throughout libraries. Today there are two main

barriers to a large scale use of this knowledge: the amount of information available throughout the Web and the fact that knowledge is embedded in the text of scientific articles in an unstructured way, not adequate for program processing.

Today, different scientific communities are developing Web ontologies which formally record the knowledge in a domain. W3C [2] defines ontology as "*a knowledge representation*". According to Jacob [3 p. 200] an ontology is "*a partial conceptualization of a given knowledge domain, shared by a community of users, that has been defined in a formal, machine-processable language for the explicit purpose of sharing semantic information across automated system*". In a near future, formal ontologies will be developed and recorded in program readable format, containing the accented knowledge in specific domains. Applying Semantic Web technologies to identify and record the knowledge embedded in the text of scientific articles in program-understandable format and compare it to the knowledge recorded in Web ontologies may be a key feature to the development of a future e-Science environment. Both these knowledge resources may be accessed by software agents on behalf of their owners, thus providing scientists with new tools to information and knowledge retrieval, to identify, evaluate and validate new contributions to Science.

The present research is looking for a new paradigm in scientific Web publishing: to publish not only text, for human reading, but also knowledge, formalized as ontologies, able to be processed by software agents. The objective of this research is to develop a Web publishing model which will be the basis for the future development of enhanced scientific authoring, publishing, retrieval and validating tools. These tools will enable the electronic publishing of scientific articles not only as texts for human reading, but also as a knowledge base in program-understandable format. The model aims to identify and record the semantic elements which constitute the knowledge embedded in the text of a scientific article.

What is the nature of scientific knowledge? This knowledge today, although recorded in digital format as Web published articles, are unstructured and not in adequate format for processing by software agents. According to Brookes [4 p. 131]: "*knowledge is a structure of concepts linked by their relations and information is a small part of such a structure*". Sheth [5 p. 1] states that "*Relationships are fundamental to semantics – to associate meaning to words, items and entities. They are a key to new insights. Knowledge discovery is about discovery of new relationships*". Miller [6 p. 306] answer these questions as: "*The above remarks imply-that science is a search after internal relations between phenomena".* Here scientific knowledge is considered as discovering relations between phenomena.

By the 16th century, a mark in the institutionalization of Science is the establishment of the scientific method as a procedure to achieve and communicate true statements in Science. A special element of scientific method is the hypothesis. As Scientific Methodologies handbooks emphasize, the role of hypotheses are central to Science in providing a provisory explanation to a phenomena and thus guiding the scientific inquiry. In the scientific method the hypothesis is the element which expresses a relation between phenomena.

Although a complex phenomena, scientific reasoning as expressed in the text of scientific articles must serve to an essential communicational role to Science as an institution: to validate the knowledge contained in the article, enabling any scientist to reproduce the steps taken by the author in his/her experiment. The need of this rigid protocol when communicating research results is stated by The International Committee of Medical Journals Editors, http://www.icmje.org:

> "*The text of observational and experimental articles is usually (but not necessarily) divided into sections with the headings Introduction, Methods, Results, and Discussion. This so-called "IMRAD" structure is not simply an arbitrary publication format, but rather a direct reflection of the process of scientific discovery*"

It is assumed here that knowledge in the text of articles – scientific methodology elements as the Problem, Hypothesis, Results and Conclusions – are all interrelated, constituting the content of the reasoning process developed by the author through which he/she communicates a new discovery. With the support of a Web authoring/publishing tool these semantic elements – the knowledge contained in the article -, can be identified, extracted and recorded in machine-understandable format, as an ontology. Knowledge thus recorded can be processed by software agents thus enabling semantic retrieval, consistence and validate checking. The ontology representing the knowledge extracted from the article can also be compared, matched and aligned to public Web ontologies which more and more represent the corpus of public knowledge in specific domains, thus enabling the establishment of formal relationship between both ontologies. Fails to establish these relationships may be evidences of new discoveries, since it can indicate that the knowledge in the article is not yet represented in the ontology which stores the accented knowledge in a specific domain.

2 Methodology

Building models is an important tool in Science. It enables Science to cope with complex phenomena such as scientific reasoning in communicating new discoveries through the text of scientific articles. An initial semantic model was developed, based on literature on Scientific Methodology, Philosophy and Epistemology of Science. Using the initial framework 53 articles on Health Science were analyzed with the aim of enhancing and validating the model. Articles were choose from two outstanding Brazilian research journals, 20 articles from the Memórias do Instituto Oswaldo Cruz, which scope is mainly Microbiology, http://www.scielo.br/revistas/mioc, 20 articles from the Brazilian Journal of Medical and Biological Research, http://www.scielo.br/revistas/bjmbr. Both are international journals using English as primary language. These journals were selected because initially we intended to interview authors personally. 14 additional articles about stem cells were analyzed too. Stem cells as an emerging research area in rapid development, was chosen expecting to find articles reporting important discoveries. Articles analyzed were selected from three recent reviews which present the stem cells research development in a historical perspective, promoting the advances in research, which was of special interest to this research. These reviews are "The Human Embryonic Stem Cell and the Human Embryonic Germ Cell", the official National Institute of Health (USA) resource for stem cells research, http://stemcells.nih.gov/, the article by Bongso et al. [7] and the article by Friel et al. [8].

The analysis simulates the tasks to be performed by an authoring/publishing tool when interacting with an author to identify and record the knowledge embedded in the text of an article. Scientific articles are highly conventional text types, with clear goal shared by authors and readers. Articles in Health Science are chosen for analysis due to their highly standardized structured, the so-called IMRAD – Introduction, Material and Methods, Results and Discussion - structure.

In order to explore the possibilities of using the model to identify new discoveries in Science, it is also verified if concepts found in the knowledge extracted from each article's text exist in a public knowledge base. DECS – Descritores em Ciência da Saúde - http://www.bireme.br/php/decsws.php, a Portuguese version of MeSH – Medical Subject Headings – http://www.nlm.nih.gov/MeSH/, and MeSH itself were both used in this experience in the role of a public knowledge base, with which subject headings found in the article's corresponding Lilacs (Latin America and Caribbean Literature on Health Science) or Medline database records are compared. MeSH is a component of UMLS - Unified Medical Language System -, http://www.nlm.nih.gov/pubs/factsheet/umls.html. It is a project of National Library of Medicine, USA, which aims to unify and encompass different medical specialized terminologies, thesaurus and classification schemas. UMLS evolves towards an ontology – the UMSL Semantic Network - in which concepts are organized in 134 classes or "semantic types" and 53 "types of relations".

The article analysis used the following form:

ARTICLE ANALYSIS FORM	
Journal: Memórias do Instituto Oswaldo Cruz	**URL**: http://www.scielo.br/revistas/mioc
Reference CAMARA, Geni NL, CERQUEIRA, Daniela M, OLIVEIRA, Ana PG *et al*. **Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil.** *Mem. Inst. Oswaldo Cruz*. [online]. Oct. 2003, vol.98, no.7 [cited 10 March 2005], p.879-883. Available from World Wide Web: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762003000700003&lng=en&nrm=iso>. ISSN 0074-0276	
METHOD OF REASONING	
Deductive: X Inductive: Abductive:	
PROBLEM (extracted from the text)	
As a contribution to the public health authorities in planning prophylactic and therapeutic vaccine strategies, we describe the prevalence of human papillomavirus (HPV) types in women presenting abnormal cytological results in Pap smear screening tests in the Federal District, Central Brazil.(Abstract) In contrast to what is observed in developed countries, cervical cancer mortality in Brazil is still high. (Introduction)	

HIPOTHESIS – previous (extracted form the text)
The chronic infection by certain types of human papillomavirus (HPV) is definitely related to the incidence of cervical cancer **(Lorincz et al. 1992, IARC 1995)** and the HPVs –16, -18, -31, -33, -35, -45, -51, -52, and -58 can now be considered as cervical carcinogenic agents **(Muñoz 2000)**. Squamous carcinomas and adenocarcinomas are the most frequent cervical neoplasias, and may develop from intraepithelial lesions, easily detected in preventive cytological exams **(Sherman et al. 1994)**.
Normalized Relation HPV infection is related to the incidence of cervical pre-neoplasic and neoplasic lesions
Antecedent: HPV, different types / Papillomavirus Humano,
Type fo relation: causes / T147 UMLS Semantic Network
Consequent: cervical pre-neoplasic and neoplasic lesions / Infecções Tumorais por Vírus, Neoplasias do Colo
Mapping to DECS: M (mapped)
DECS Subject Headings Papillomavirus Humano/classificação, Infecções Tumorais por vírus/epidemiologia, Neoplasias do Colo Uterino/virologia, Papillomavirus Humano/genética, Infecções Tumorais por Vírus/patologia Infecções Tumorais por Vírus/virologia, Neoplasias do Colo Uterino/diagnóstico Doenças do Colo Uterino/patologia, Doenças do Colo Uterino/virologia DNA Viral/genética, Esfregaço Vaginal, Reação em Cadeia da Polimerase Polimorfismo de Fragmento de Restrição, Genótipo, Fatores de Risco Prevalência
Citations: (Lorincz et al. 1992, IARC 1995), (Muñoz 2000), (Sherman et al. 1994).
EXPERIENCE
Results
Measure: prevalence
Context: Environment: Place: Distrito Federal, Brazil / Brasil/epidemiologia Time: Group: women / Humano, Feminino, Adulto, Meia-Idade
Methodology:
Conclusions
Observations:

Figure 1: Article Analysis Form

3 Results

We envisage an authoring/publishing software tool which will be available to the author during the process of Web publishing his/her article, and interactively will capture the articles knowledge, recording it in a standard program readable format. This knowledge can then be retrieved and processed by semantic retrieval tools. Validation tools or software agents could also compare the knowledge extracted from articles with that held in public ontologies like the UMLS and thus indicate inconsistencies, faults and even new discoveries. The overall authoring/publishing environment is discussed in Marcondes [9] and illustrated in Figure 1. The authoring/publishing software tool development and how to identify new discoveries using the model proposed are in our agenda and will be object of future research. The present research is conceived only with proposing, testing and validating a model to the knowledge extracted from the article's text by a future authoring/publishing tool to be developed.

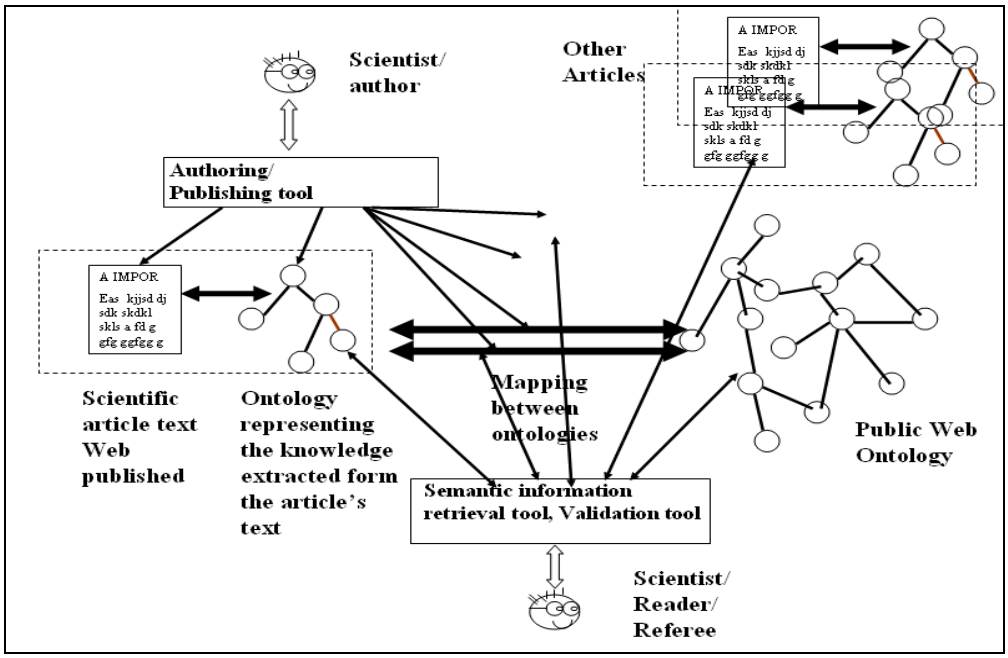

Figure 2: Author's editing/Web publishing environment

What are the methods to achieve the truth in Science? These questions date back to Greek Philosophy with Epistemology, Rhetoric, Dialectics and Sophistic. Aristotle proposed patterns of reasoning from which true statements could be achieved from previous statements. He invented the reasoning method called *deduction,* through which particular statements can be derived from general statements. These patterns were systematized by Medieval Scholastics.

A branch of this discussion with important contributions came at the Modern Age, with the establishment of the scientific method by Francis Bacon [10]. In opposition to Medieval Scholastics, Bacon emphasized the importance of observational experiments to achieve general laws in Science. His reasoning method of deriving general statements from a particular number of observational cases was called *induction*. Besides all criticisms to the bases of the scientific method induction reasoning is still a strong basis to experimental Science.

Pierce adds to deduction and induction the abduction method of reasoning. According to him abduction is essentially the creative process of generation new explanatory hypotheses from apparently unstructured observational data. Pierce also integrated abduction with deduction and induction, proposing a whole method to scientific inquiry: a new hypothesis is abductively generated; its consequences are deductively inferred and inductively tested.

Abduction is considered as the logic of discovery by many researchers as Hoffmann [11], Magnani [12] and Paavola [13]. Pierces' example of abductive reasoning is Kepler discovery that planet orbits are not circles, as believed Copernic, but ellipsis. Abduction has always been associated with new discoveries both by Pierce himself and by researchers working on his legacy. Induction and Deduction are always associated with hypotheses testing and their ratification or refusal, an incremental increase to knowledge stock.

An article's knowledge - or semantic elements - appears according to the reasoning procedure employed by the author. It is important to identify these semantic elements to the development of an ontology which will guide a future authoring/publishing software tool while interacting with the author during knowledge extracting from article's text as a by product of the writing/publishing activity.

The article analysis showed three patterns of reasoning procedures. According to the reasoning procedure employed scientific articles can be classified as *theoretical articles*, which employ abductive reasoning and *experimental articles* which employ inductive or deductive reasoning. The elements complaining the structure of knowledge contained in the text of the article differs depending on the type of reasoning procedure used by the author.

These elements are: the PROBLEM the article is trying to address, the HYPOTHESIS, where the author states a RELATION between phenomena, a possible empirical controlled EXPERIMENT with the aim of observing the phenomena described, specific of experimental articles, divided in RESULTS – tables, figures, numeric data, reporting the observations made -, MEASURE used, a specific CONTEXT where the empirical observations take place, subdivided in ENVIRONMENT – a hospital, a crèche, a high school -, a geographical PLACE where the empirical observations take place, TIME when the empirical observations occurs, a specific GROUP – pregnant women, early born babies, mice - in which the phenomena occurs, and CONCLUSION – a set of propositions made by the author as a result of his/her findings.

Although all these elements are important to reasoning procedure, the hypothesis is the element which has the potential to hold new knowledge. The hypothesis has the form of a RELATION formed by two or more ARGUMENTS linked by a TYPE_OF_RELATION. In every article analyzed concepts found in the ARGUMENTS were tentatively mapped to concepts taken from the UMLS verifying if these concepts correspond to DECS/MeSH subject heading extracted from the article's record in Medline or Lilacs databases.

Theoretical-abductive model of articles are based on synthesis of Gross [14] and Hutchins [15] proposals. *Theoretical-abductive* articles analysis different previous hypotheses, show their faults and limitations and propose a new hypothesis; the reasoning is as follows:

> *a PROBLEM is identified, with the following aspects and data;*
> *the previous authors/HYPOTHESES are not satisfactory to solve the PROBLEM due to the following criticism;*
> *so, we propose this new HYPOTHESIS which we consider as a new pathway to solve the PROBLEM.*

Experimental-inductive articles propose a hypothesis and develop experiments to test and validate it; 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 it comes at the following RESULTS.*

In experimental-inductive articles, a CONCLUSION is one of the following types: or it corroborates the hypothesis, or it refuses the hypothesis or it partially corroborates the hypothesis. However in some cases, the CONCLUSION is neither the former, it just reports intermediate, not conclusive results toward the hypotheses corroboration.

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

> *a PROBLEM is identified, with the following aspects and data;*
> *in literature the previous authors/HYPOTHESIS are proposed;*
> *we choose the following previous HYPOTHESIS;*
> *we enlarge and re-contextualize this HYPOTHESIS; we developed a EXPERIMENT to test it in this new context;*
> *the EXPERIMENT shows the following RESULTS in this new CONTEXT.*

Experimental articles also can compare various phenomena or hypotheses, as in a comparative study, a very usual type of article in Health Sciences. The different reasoning procedures can be formalized in an Ontology for Scientific Knowledge in Articles, as illustrated in Figure 2. This ontology has the following Classes and Properties:

Classes: THEORETICAL reasoning and
EXPERIMENTAL reasoning
Subclasses: INDUCTIVE reasoning and
DEDUCTIVE reasoning
Properties: PROBLEM
HYPOTHESIS (previous or new)
Sub-properties: ANTECEDENT
TYPE-OF-RELATION
CONSEQUENT

REFERENCES (just in previous HYPOTHESIS)
EXPERIMENT
Sub-properties: RESULTS (quantitative data, tables, etc.)
MEASURE
CONTEXT
Sub-properties: SPACE
TIME
GROUP

Two Classes of articles were identified: Theoretical and Experimental. Experimental articles in turn have two Subclasses, Inductives and Deductives. The Properties of articles are the following: Theoretical-abductive articles have a PROBLEM, one or more previous HYPOTHESIS, that are discussed, criticized and rejected as solutions to the PROBLEM posed. So, the author proposes a new HYPOTHESIS which may be a solution to the PROBLEM. Theoretical-abductive articles do not present experimental results.

Experimental articles in turn always present experimental results. Experimental-deductive articles have the following Properties: a PROBLEM, one or more previous HYPOTHESIS, by different authors, that are adopted to guide an experiment. Previous HYPOTHESIS are extended, restricted or inserted in a new CONTEXT. An experiment is developed bases in the previous HYPOTHESIS applied to the new CONTEXT and the results of the EXPERIMENT are reported.

Experimental-inductive proposes an original new HYPOTHESIS to address a PROBLEM, develop an experiment to test this HYPOTHESIS and the results of the EXPERIMENT are reported.

HYPOTHESES have an ANTECEDENT, a TYPE-OF-RELATION and a CONSEQUENT. HYPOTHESES hold the knowledge embedded in the article as it proposes a relation between phenomena.

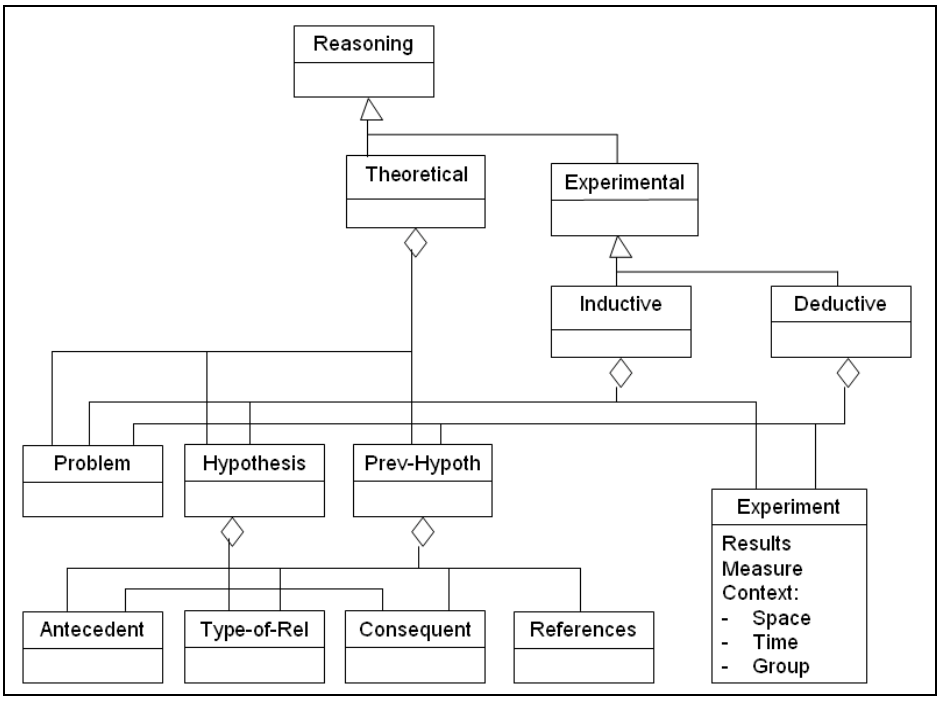

Figure 2: Class diagram of the Ontology for Scientific Knowledge in Articles

We plan to implement the Ontology for Scientific Knowledge in Articles in OWL [16]. The ontology will guide a future authoring/publishing tool in its interaction with an author to extract and record the knowledge embedded in the text of an article. Quantitative results of the analysis done on 53 articles are showed in Table 1. According to the classification proposed the majority of articles are experimental articles, 50 out of 53. Just 3 are theoretical-abductive articles.

Articles analyzed	Exp-inductives	Exp-deductives	Theor-abductives	TOTAL
MIOC	4	15	1	20
BJMBR	4	13	2	19
STEM CELLS	10	4	0	14
TOTAL	17	33	3	53

Table 1: Results of the articles analysis

In all articles the HYPOTHESIS was generally found in the Introduction section, in the Title or in the Abstract. Articles were considered Fully Mapped when concepts in both ARGUMENTs and the TYPE OF RELATION where fully mapped to one or more DECS/MeSH concepts that index the record in databases as Medline and Lilacs and there is a UMLS Semantic Network Relation corresponding to the TYPE OF RELATION. Articles were considered Partially Mapped when concepts in at least one of the ARGUMENTs or in the TYPE OF RELATION where fully mapped to one or more DECS/MeSH concepts and UMLS Semantic Network Relations. Articles were considered Not Mapped when any concept in neither the ARGUMENTs nor in the TYPE OF RELATION were fully mapped to DECS/MeSH concepts and UMLS Semantic Network Relations. The mapping of concepts to the DECS/MeSH is lower - which may be an indicative of new discoveries -, in a research area as stem cells in comparison to the two Brazilian journal. Table 2 shows these results.

Articles analyzed	MIOC	BJMBR	STEM CELLS
Total of articles	20	19	14
Fully mapped	11	4	0
Partially mapped	9	10	11
Not mapped	0 (0%)	5 (25%)	2 (7%)

Table 2: results of the mapping of concepts found in hypotheses to DECS/MeSH

4 Discussion

The majority of articles found are experimental, 50 out of 53. The experimental articles all fit in the IMRAD model, with definite textual parts while the theoretic-abductive articles not. This fact may indicate a pattern of research characterized as "normal Science" according to Kuhn's [17] theory.

Although foreseen in the literature only three theoretical-abductive articles were found among the articles analyzed. As this is the type of article which reports expressive paradigm changes in a scientific area it is expected that they are not very usual. But their existence is certain. For example, Watson and Crick article proposing a model to the DNA molecule is a typical theoretical-abductive article. All three articles found do not fit into the IMRAD structure. They do not have sections such as *Material and Method* and *Results*. Some review articles and letters to the editor have some traces of theoretical-abductive articles and must be object of future research.

Stem Cells potentialities constitute a new paradigm in cell biology. "*A new era in stem cell biology began in 1998 with the derivation of cells from human blastocysts and fetal tissue with the unique ability of differentiating into cells of all tissues in the body, i.e., the cells are pluripoten*." (http://stemcells.nih.gov/). Since then two problems face the researches in the area: how to maintain stem cells cultures indefinitely undifferentiated in specialized cell types as bone, skin, liver, etc., and how to start and control differentiation into specific cells types. In the Stem Cells articles group there is a predominance of experimental articles reporting culture or control methods, in all of which the TYPE OF RELATION was mapped to relation "method" (UMLS Semantic Network T183). All articles of this group seem to report incremental advances in knowledge. Nome theoretical-abductive article was found in this group.

Few articles are totally mapped to DECS/MeSH concepts and to UMLS Semantic Network Relations. The process of mapping the concepts found in the ARGUMENTs and in the TYPE OF RELATION of each HYPOTHESIS is just a by-product of the data generated by the analysis process, just an explorative pathway to generate data for future research. In the majority of cases concepts in the ARGUMENTs were too specific in comparison to DECS/MESH concepts used to index the record. On the other hand the majority of TYPE OF

RELATIONs identified was satisfactorily mapped to UMLS "relations". This fact may be due to the difference in numbers: there are more than 730.000 concepts in UMLS and just 53 "relations". Relations are more stable across the time and more generic in comparison to concepts in a scientific area. Another explanation to this fact is that there is always a delay to these concepts be incorporated in the UMLS, so it is in dead an indicative of new discoveries. Anyway, operational results enabling software agents to compare the knowledge extracted from the text of articles to the knowledge record in Web ontologies according to the model proposed deserves more research.
The analysis performed shows that the scientific reasoning elements, according to the type of reasoning employed, are structured, forming an ontology, in the sense used in knowledge engineering, as in Sowa [18]. This enables a software agent to perform *inferences* on this structure. Based on the example analysis presented in Figure 1 knowledge extracted from articles, marked up and recorded as described would enable the following queries by 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 to 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 other pathologies different from neoplasias?
- which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in different contexts? (not in women from Federal District, Brazil).

To publish scientific articles both as text and as machine readable knowledge bases seems to be a promising approach. It will enable the processing of this knowledge by software agents, thus improving critical inquiry, semantic querying and validation of scientific contributions to Science. Experimental Science, as Health Science, offer a solid basis to the development of the model, due to its formalism, derived from the use of the Scientific Method as an reasoning strategy in the text of scientific articles. The model outlined is a semantic model which aims to identify the semantic content of scientific reasoning. It is intended to be the basis to the development of a Web authoring/publishing tool. To reach this objective new research on computational techniques must be developed. We envisage an authoring/publishing tool that offers researchers/authors an interactive Web environment which, through a rich dialogue and using text extraction techniques, interactively identify and extract relevant contents of the article been written/published. This content will then be represented in machine-understandable format as an ontology, using OWL. Scientific articles so published throughout the Web can then be interlinked and linked to the increase number of Web ontologies, forming a rich knowledge network, thus enabling software agents to help scientist identify and validate new discoveries to Science. As the model proposed became more robust, there are plans to test it in other empirical science areas and even in areas as social sciences.

5 Conclusion

In all articles analyzed a relation expressing the mainly findings reported in the article was identified. This seems to indicate that scientific knowledge as expressed in the text of scientific articles can be represented as relations between phenomena. The amount of scientific knowledge now available throughout the Internet is so vast that it can only be processed with the aid of computer power. Here is proposed a standard representation to this knowledge feasible to be processed by software agents. This is essential if the intention is to use software agents to large scale processing of this knowledge in tasks as knowledge validation, semantic retrieval, identification and evaluation of discoveries.

Articles analyzed are very few and restricted to a single scientific area. If we are going to establish a new paradigm in electronic scientific publishing in which articles are published not only to human reading but also to be processed by software agents, this deserves more research. The model proposed is just a starting point to be discussed and enhanced by the scientific community.

Indexing language, as different Thesaurus largely used in information systems, select a set of concepts to describe a document. All knowledge organization effort is oriented toward the organization of systems of concepts. Generally all these concepts play an identical role when representing and retrieving a document. Although relations play a key role in scientific knowledge conventional indexing languages play no attention to them. Indexing language to no express the relations held between the subject headings indexing a document.

Indexing language must include relations between subject headings. There is also a need of the development of a taxonomy of relations used in Science to help indexing/retrieval scientific articles.

The model proposed also points to the standardization of a SkML - Scientific Knowledge Mark up Language - encompassing the semantic content of scientific articles Web published. This article highlights the benefits of a semantically richer format to represent the knowledge in scientific articles. With the aid of adequate software tools, this knowledge can be extracted as a by-product of authoring/publishing an article by the author. This opens an all new perspective in scientific knowledge acquisition, storage, processing and sharing.

Notes and References

[1] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, May, 2001. Available from Internet: <http://www.scian.com/2001/0501issue/0501berners-lee.html>.

[2] *OWL Web Ontology Language Guide.* Available form Internet: <http://www.w3c.org/TR/2004/REC-owl-guide-20040210/>.

[3] JACOB, E. K. Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, Abril/May, 2003.

[4] BROOKES, B. The foundations of Information Science. Part I. Philosophical aspects. *Journal of Information Science*, vol. l2, 1980, pp. 125-133.

[5] SHETH, A; ARPINAR, I. B.; KASHYAP, V. Relationships at the heart of semantic web: modeling, discovering and exploiting complex semantic relationships. In: NIKRAVESH, M. Et al. *Enhancing the power of the internet studies in fuzziness and soft computing*. Springer-Verlag, 2002. Available from Internet: <http://cgsb2.nlm.nih.gov/~kashyap/publications/relations.pdf>.

[6] MILLER, D. L. Explanation Versus Description. Philosophical Review, vol.. 56, no. 3, May, 1947. pp. 306-312. doi:10.2307/2181936.

[7] BONGSO, A; RICHARDS, M. History and perspective of stem cell research. Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 18, no. 6, 2004. pp. 827-842.

[8] FRIEL, R; SAR, S; MEE, P. Embryonic stem cells: Under standing their history, cell biology and signalling. Advanced Drug Delivery Reviews, vol.57, no.13, 2005. pp. 1894-1903.

[9] MARCONDES, C. H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: Proc. 9th ICCC ElPub - International Conference on Electronic Publishing, Leuven, Belgium, 2005, 9, p.119-27. Available from Internet: <http://elpub.scix.net> .

[10] BACON, F. *Novum Organum*. São Paulo : Abril Cultural, 1973.

[11] HOFFMANN, M. Is there a "Logic" of Abduction? In: Proc. 6th. Congress of the IASS– AIS International Association for Semiotics Studies, Guadalajara, Mexico, 1997. Available from Internet: <http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html>.

[12] MAGNANI, L. *Abduction, Reason, andScience: Processes of Discovery and Explanation*. New York : Kluwer Academic; Plenun Publishers, 2001.

[13] PAAVOLA, S. Abduction as a Logic and Methodology of Discovery: the Importance of Strategies. Foundations of Science, Vol.9, No. 3. November, 2004. p. 267-283. doi: 10.1023/B:FODA.0000042843.48932.25.

[14] GROSS, A. G. The Rhetoric of Science. Cambridge, Massachusetts; London: Harvard University Press, 1990.

[15] HUTCHINS, J. On the structure of scientific texts. In: Proc. 5th. UEA Papers in Linguistics, Norwich.. Norwich, UK: University of East Anglia, 1977. p.18-39.(Conference Proceedings). Available from Internet: <http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>.

[16] OWL- Ontology Web Language, a W3C standard language to represent ontologies. Available form Internet: <http://www.w3.org/2004/OWL/>

[17] KUHN, T. The structure of scientific revolutions. In: Foundations of the unity of Science, vol. 2. Chicago : the University of Chicago Press , 1970.

[18] SOWA, J. *Knowledge representation: logical, philosophical and computational foundations*. Pacific Grove, CA : Brooks/Cole, 2000.