A proposed model of knowledge representation and the coding of knowledge embedded in texts of Web published scientific articles

Carlos Henrique Marcondes[a1], Marília Alvarenga Rocha Mendonça[a], Luciana Reis Malheiros[b]
[a] *Department of Information Science*
[b] *Department of Physiology and Pharmacology*
Federal Fluminense University, R. Lara Vilela, 126, 24210-590, Niterói, RJ, Brazil

This article reports results of a research project with the aim of investigating the possibilities of electronic publishing 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 in XML the article contribution, new knowledge and scientific novelty. The use of XML language to represent this knowledge enables its processing by intelligent software agents 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 Web environment. 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 methodology, scientific communication, knowledge representation, ontologies.

1 INTRODUCTION

Nowadays, electronic Web publishing is a common activity to scholars and researchers. Despite this fact, electronic journals are still based in the print model and do not take full advantage of the facilities offered by the Web environment. Since the Philosophical Transactions of Royal Society in the 17th century the scientific article is the container of new scientific knowledge. Before the raise of the Web, paper journals collections in libraries constituted the humanity scientific knowledge bases. 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.

Scientific communication is a slow social process that largely depends on discourse, text producing and reading/interpreting/inquiring these texts by scholars until new knowledge is incorporated to the corpus of Science. The potential of new information technology 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. In the Semantic Web context [1], electronic publishing could be a *cognitive tool* which its potential is far from being explored [2]. A related project which points toward the same objective of enlarging this potential is W3C Scientific Publishing Task Force Ontology for Experiment Self-Publishing[1].

The objective of this research is to propose a Web-publishing model which enables the electronic publishing of scientific articles not only as texts for human reading, but also as a knowledge base in machine-readable format in XML[2]. As Scientific Methodologies handbooks emphasize [3], [4], [5], [6], scientific knowledge has the form of relations between phenomena. In special, the hypothesis is the element which contains a relation. We envisage an authoring/self-publishing environment in which knowledge in the text of articles – the elements of scientific reasoning - are identified and recorded in machine readable format.

A framework to analyze text is proposed by Kintsh and Van Dijk [7]. Gardin [8] proposes that scientific articles have embedded in their texts the scientist reasoning. In Brazilian Information Science literature, Smit [9] and Kobashi [10], applied both the proposals of Kintsh/Van Dijk and Gardin to the analysis of document for indexing and abstraction. This proposal intends to go further than indexing for providing access, it intends to enable the processing of the knowledge embedded in articles texts.

New discoveries in Science are validated comparing them with the assented knowledge in a specific domain. Before the raise of the Web, what constitutes this assented knowledge was fuzzy, lacks formalization, was scattered across journals collections in libraries. The main mechanism of Science validation was and still is reading, interpreting, inquiring, criticizing and, in brief, citing journal articles by scholars, until new knowledge was finally incorporated in the fuzzy corpus of Science. This knowledge, through to the scientific communication process, is turned into what Ziman [11] calls the public knowledge.

[1] http://esw.w3.org/topic/HCLS/ScientificPublishingTaskForce
[2] XML- Extensible Markup Language, a standard from W3C, http://www.w3c.org/xml

Today this scenario is rapidly changing. De Roure [12] describes a Web environment called The Semantic Grid/eScience, in which scholars will have tools that "*getting hold of the information that is around, and turning it into knowledge by making it usable. This might involve for instance, making tacit knowledge explicit, identifying gaps in the knowledge already held, acquiring and integrating knowledge from multiple sources*". Different scientific communities are at present developing Web ontologies which formally record the knowledge in a domain, like the UMLS – Unified Medical Language System[3]. In a near future, formal ontologies will be developed and recorded in program readable format, containing the knowledge in specific domains. This knowledge may be accessed by software agents, thus helping scientist use it to order to validate new contributions to Science.

Within this scenario, the model proposed, as a by-product of articles writing and self-publishing process in a Web environment, will extract knowledge from articles text, record it in machine readable format and also link it to public Web ontologies. This process will enable the establishment of formal relationship between the article content and ontologies which represent the public knowledge in a domain. The model is proposed as the basis to the future development of enhanced authoring, publishing and retrieval tools. While Information Science cognitive paradigm considers knowledge as process occurring in user's mind [13], we adopted the view from Artificial Intelligence [14], which represents knowledge as production rules.

The working hypothesis of this research are thus the following: a- scientific knowledge, as it appear in the text of scientific articles, has the form of relations between phenomena; b- it is feasible, through the aid of an authoring/Web publishing tool, to identify, extract it, record this knowledge in machine readable format and, in addition, formally link it to Web public ontologies; c- in such framework, a fail to formally link the knowledge thus extracted and recorded could be an indication of a new scientific discovery.

2 MATERIAL AND METHODS

A key point to the development of an authoring/publishing environment as described is the development of a solid model to this process. We are engaged in the development of such model. Since Bacon [15] and Descartes [16] the Scientific Method is a solid basis to Science. Its actual version, the Hypothetic-Deductive Method, was systematized by Popper [17]. All scientific methodology handbooks stress the role of hypothesis in guiding scientific inquiry. Hypothesis have the form of relations between phenomena.

An initial model was developed [18], having the literature about Scientific Methodology, Philosophy and Epistemology of Science as basis. Accordingly, Gross [19] emphasizes the existence of two types of scientific articles, theoretical and experimental. Hutchins [20], based also in Van Dijk model, proposes that "*hypothesis testing*" articles use induction reasoning and that "exploratory" articles employ "*abduction*" reasoning. Hoffman [21], using Pircean framework, discusses the potentialities of abduction reasoning to arise to new scientific discoveries.

The proposed model can be divided in a- an Web authoring/publishing environment model b- a scientific reasoning model identifying, the elements that constitutes the scientific reasoning and c- a record of the scientific reasoning model in program readable format using XML. Item a- is presented in [18]; the present paper mainly describes items b and c.

A crucial point in the scientific reasoning model is to identify relations between phenomena expressed in the text of scientific articles. Our assumption is that those relations contain the knowledge in the articles text. Using the initial model framework 40 articles were analyzed simulating the tasks that an authoring tool would perform, in order to validate and enhance the scientific reasoning model. Articles were choose from two outstanding Brazilian research journals, the Memórias do Instituto Oswaldo Cruz, which scope is mainly Microbiology, http://www.scielo.br/revistas/mioc and the Brazilian Journal of Medical and Biological Research, http://www.scielo.br/revistas/bjmbr. Articles in Health Science are used due to their highly and standardized structured, the so-called IMRAD[4] structure.

As a second step in analyzing journal articles, the knowledge extracted from articles text, in the form of relations between phenomena, was mapped to the public knowledge base, in a similar process to scientists reading, validating, criticizing and citing other articles. We use the UMLS to simulate this knowledge base because the electronic journals analyzed are indexed using DECS, a translation to Portuguese of MESH – the

[3] http://www.nlm.nih.gov/pubs/factsheet/umls.html.
[4] http://www.icmje.org

Medical Subject Headings, which is a subset of UMLS.

3 RESULTS

An enhanced and richer model emerged from the analysis of the articles. The model classifies scientific articles as theoretical articles, which employ abductive reasoning and experimental articles which employ inductive or deductive reasoning. Depending on the type of reasoning the structure of their elements contained in the article text differs. These elements are: the PROBLEM, the HYPOTHESIS, in the form of a RELATION, a possible empirical MANIFESTATION of the phenomena described, divided in RESULTS, MESURE, CONTEXT (subdivided in ENVIRONMENT, PLACE, TIME and GROUP), and CONCLUSION.

Theoretical-abductive articles analysis different previous hypothesis, show their faults and limitations and propose a new hypothesis; reasoning is as follows:

- *a PROBLEM is identified, with the following aspects and data;*
- *the previous authors/HYPOTHESIS 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;*
- *on the basis of this new HYPOTHESIS the PROBLEM has the following empirical MANIFSTATION;*
- *we developed an experiment to test this MANIFESTATION 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 are neither the former, it just reports intermediate, not conclusive results toward the hypotheses corroboration.

Experimental-deductive articles use hypothesis proposed by other researchers 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 which has this empirical MANIFESTATION;*
- *we test, enlarge and re-contextualize this HYPOTHESIS;*
- *the test shows the following RESULTS in this new CONTEXT.*

The relation expressed as the knowledge contained in the article text is identified with the HYPOTHESIS and each of its components, the ANTECEDENT, the TYPE OF RELATION and the CONSEQUENT are then mapped to the UMLS.

Afterwards is showed a record in XML to explain the analysis performed in the following article:

Camara, Geni NL, Cerqueira, Daniela M, Oliveira, Ana PG *et al*. 2003. 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], 98(7), pp.879-883. Available from World Wide Web: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762003000700003&lng=en&nrm=iso>.

```
<?xml version="1.0" encoding="ISO8859-1" ?>
<knowledge_structure art-id="352387“ source=“Lilacs”>
    <problem>Which types of HPV are  prevalent in the group?</problem>
    <reasoning type=”deductive”
     <hypothesis type=”previous”>
            <citation> Lorincz et al. 1992</citation>
            <citation> IARC 1995 </citation>
            <citation> Muñoz 2000</citation>
            <citation> Sherman et al. 1994</citation>
            <relation>
```

```
                    <antecedent>human papillomavirus</antecedent>
                    <type_of_relation>causes ("causes", T147/UMLS SN)</type_of_relation>
                    <consequent>pre-neoplastic and neoplastic cervical lesions</consequent>
                </relation>
            </hypothesis>
        </reasoning>
        <manifestation>
            <results>Table 1, Table 2</results>
            <measure>Prevalence</measure>
            <methodology></methodology>
            <context>
                <group>Women</group>
                <space>Federal District</space>
                <space>Brazil</space>
            </context>
            <conclusion type="corroborates_hipothesis">
                HPV-16 is the most  prevalent in the group
            </conclusion>
        </manifestation>
        </knowledge_structure >
```

4 DISCUSSION

To map the relation established in an article to a public knowledge base is conceptually an important step in the model proposed. In the empirical experiment we developed to simulate the tasks that would be performed by a future authoring/publishing tool, the use of the UMLS was not completely satisfactory as the public knowledge base. Some concepts found in articles relations were too specific and we did not find the corresponding concepts in UMLS. Also the relations provided by UMLS Semantic Network are always binary relation and we find relations like "method" which as a ternary relation between two phenomena and its instrument or "method" of measuring. The UMLS is mainly a terminological device, not really an ontology or a knowledge base.

Almost all articles analysed were experimental-deductive articles. Within Kuhn's [22] vision of scientific revolutions, this kind of articles works inside the actual paradigm in a scientific domain. They do not question this paradigm. The model proposed can be a valuable tool to point out traces of new scientific discovery. To test this hypothesis we plan to analyse scientific articles in leading journals which traditionally address new discoveries, as Nature Genetics, in areas as stem cells and using the GeneOntology[5] as knowledge base.

As example of potentialities of the model knowledge marked up as described and recorded in a semantic information retrieval system will enable the following queries:

- which other articles have hypothesis suggesting HPV as the cause of cervical neoplasias in women?

- which articles have hypothesis suggesting other causes to cervical neoplasias different from HPV in woman?

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

- which articles have hypothesis suggesting HPV as the cause of other pathologies different from neoplasias?

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

5 CONCLUSION

The complete development of the research agenda proposed here will certainly need a highly interdisciplinary perspective, with approaches from Philosophy and Epistemology of Science, from Health Science and from Computer Science and Knowledge Representation besides Information Science,

[5] http://www.geneontology.org

Maybe OWL[6] is a best framework to represent the knowledge contained in articles than XML due to XML lack of formalism. We plan to develop an OWL ontology to articles knowledge.

The types of reasoning identified so far are typical of empirical sciences. Social Sciences and Humanities frequently employ qualitative methods quite different from the experimental methods of empirical sciences. To what extend can this model be applied or must be modified to be applied to Social Sciences and Humanities? Rojas [23] says that the typical reasoning in Social Sciences and Humanities is what he calls the Hermeneutic inference.

To publish scientific articles both as text and as machine readable knowledge bases seems to be a promising approach. As a generalize authoring/publishing environment as proposed is available to scholars it will enable the processing of knowledge contained in articles by program agents, thus improving critical inquiry, semantic querying and validation of scientific contributions to Science. This research also points toward the development and standardization of a Sm-ML - Scientific Methodology Markup Language.

REFERENCES

[1] Berners-Lee, Tim; Hendler, James; Lassila, Ora. 2001. The semantic web. Scientific American, May, 2001. Available at <http://www.scian.com/2001/0501issue/0501berners-lee.html>. Access in May 24 2001.

[2] Lévy, Pierre. 1993. As tecnologias da inteligência: o futuro do pensamento na era da informática. Rio de Janeiro : Ed. 34, 208 p. (Coleção Trans).

[3] Bunge, Mario. Philosophy of science. 1998. New Brunswick; London : Transaction Publishers.

[4] Alves_Mazzotti, Alda; Gewandsznaider, Fernando. 2002. O Método nas Ciências naturais e sociais: pesquisa quantitativa e qualitativa. São Paulo : Pioneira Thomson Learning.

[5] Marconi, Marina de Andrade; Lakatos, Eva Maria. 2004. Metodologia científica. São Paulo : Editora Atlas.

[6] Mattar Neto, José Augusto. 2002. Metodologia científica na era da informática. São Paulo : Saraiva.

[7] Kintsh, Walter; Van Dijk, Teun A. 1972. Towards a model of text comprehension and production. Psycological Review, 84(5), pp.363-393.

[8] Gardin, Jean-Claude. 2001. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. Proc. Colloque ISKO-France Filtrage et résumé automatique de l' information sur les réseaux., Conference invitee, Univesité de Nanterre – Paris X. (Conference proceedings).

[9] Smit, Johanna. 1987. Análise documentária: análise da síntese. Brasília : IBICT.

[10] Kobashi, Nair. 1994. A elaboração de informações documentais: em busca de uma metodologia. Tese (doutorado), Escola de Comunicação e Artes, USP. São Paulo.

[11] Ziman, John. 1979 Conhecimento público. Belo Horizonte : Itatiaia, São Paulo : Ed. da Universidade de São Paulo.

[12] De Roure, David; Jennings, Nicholas; Shadbolt, Nigel. 2001. Research agenda for the Semantic Grid: a future e-Science infraestructure. (Report commissioned for EPSRC/DTI Core e-Science Programme).

[13] Brooks, B. C. 1980. The foundations of Information Science. Part I: Philosophical aspects. Journal of Information Science, 2, pp.125-133.

[14] Sowa, John. 2000. Knowledge representation**:** logical, philosophical and computacional foudations. Pacific Grove : Brooks/Cole.

[15] Bacon, Francis. Novum organum. 1973. São Paulo : Abril Cultural. (Coleção Os pensadores, 13).

[16] Descartes, René. 2005. Discurso do método. São Paulo : Martin Claret. (Coleção Obra prima de cada autor).

[17] Popper, Karl. A lógica da pesquisa científica. 2001. São Paulo : Ed. Cultrix, Ed. USP, 2001.

[18] Marcondes, Carlos H. 2005. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. Proc. 9th ICCC ElPub - International Conference on Electronic Publishing, Leuven, Belgium, 2005, 9, pp.119-27. (Conference Proceedings). Available from <http://elpub.scix.net> .

[19] Gross, Alan G. 1990. The Rhetoric of Science. Cambridge, Massachusetts; London: Harvard University Press.

[20] Hutchins, John. 1977. On the structure of scientific texts. Proc. 5th. UEA Papers in Linguistics, Norwich..Norwich, UK: University of East Anglia, 1977. p. 18-39.(Conference Proceedings) Available at: <http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>. Access in Mar. 30, 2006.

[21] Hoffmann, Michael. 1997. Is there a "Logic" of Abduction? Proc. 6th. Congress of the IASS– AIS International Association for Semiotics Studies, Guadalajara, Mexico. (Conference Proceedings) Available at <http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html>. Access in Dez. 14, 2005.

[22] Kuhn, Thomas. 1970. The structure of scientific revolutions. In: Foundations of the unity of Science, vol. 2. Chicago : the University of Chicago Press.

[23] Rojas, Miguel Angel Rendón. 2005. Relación entre los conceptos: información, conocimiento y valor. Semejanzas e diferencias. Ciência da Informação, 34(2), pp52-61.

[6] OWL, Ontology Web Language, http://www.w3c.org/owl