# FROM SCIENTIFIC COMMUNICATION TO PUBLIC KNOWLEDGE: THE SCIENTIFIC ARTICLE WEB PUBLISHED AS A KNOWLEDGE BASE

*Carlos H. Marcondes*

Information Science Department, Information Science Post-graduate Program, Universidade Federal Fluminense
Instituto de Arte e Comunicação Social, R. Lara Vilela, 126 – 24210-590, Niterói – RJ, Brazil
marcon@vm.uff.br

## Abstract

Linking Electronic published scientific articles to Web ontologies is a cognitive tool of which its impacts and possibilities are far from being evaluated. The objective of this research is to investigate the potential of Web published scientific articles, conceived not only as texts, but also as a machine readable knowledge base, explicitly and formally related to Web-based public ontologies, that represent the assented knowledge of a specific domain. A prospective survey is developed to identify similar proposals and innovative experiences in electronically publishing scientific articles, authoring tools and citation analysis. Scientific methodology is also reviewed looking for structural characteristics of the scientific method presented in the written text of scientific articles. Experiences in developing Markup Language for some specific areas of knowledge, such as Chemical Markup Language, Mathematics Markup Language and Biology Markup Language, are also reviewed. An electronic publishing process is outlined which would permit the electronic publishing of not only scientific articles as full-texts, but would also enables an author to formalize the "deep structure" of a scientific article, containing assumptions, hypotheses, methodology, citations, datasets used, conclusions and contributions. All these elements are published as a knowledge base, using XML language, thus outlining a Sm-ML, a Scientific methodology Markup Language. Concepts expressed in the different parts of a Scientific article are to be linked to public Web ontologies, thus enabling the establishment of a formal relationship between the Scientific article specific knowledge base to ontologies like the UMLS – the Unified Medical Language System (http://www.nlm.nih.gov/pubs/factsheet/umls.html). The citations of an article are also be linked to the cited Web published scientific articles as *qualified* citations, in which the reasons to cite and the relationship between this specific scientific article and its citations are made explicit. The proposed model can enhances the scientific communication process, permitting semantic retrieval, critical inquiring, semantic citation, comparison, coherence verification and validating of a scientific article against public Web ontologies, which express the assented knowledge of a scientific area. The model was also conceived as the base for developing enhanced authoring and retrieval tools.

**Key-words:** electronic publishing; scientific communication; scientific methodology; Semantic Web; XML; medical web ontologies; authoring tools; e-science

## 1 The problem

All scientific manuals agree that scientific knowledge must be communicable and verifiable. Scientific communication is a slow social process that largely depends on discourse, text producing by scholars and reading/interpreting/inquiring these texts by other scholars. What would be the impacts of information technology in the scientific communication process in order to accelerate the embodying of new research results to the corpus of public, assented knowledge in a specific domain? The potential of new information technology has been applied to modern bibliographic information systems to improve scientific communication, which provides fast notification and immediate access to full-text scientific documents. Despite these advances, the scientific communication process depends largely on text production, reading and interpretation, and the citation of scientific articles by researchers. Today, except for some pioneer initiatives, the majority of the Web-published scientific journals are strongly based on the paper print publishing model.

Nowadays, electronic 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. Web publishing scientific articles can be a cognitive tool in which its potentialities are far from being evaluated. A break from the print model wold be to permit researchers, in a Web publishing environment, an authoring tool, to electronically publish no only the text of a scientific article, but also assumptions,

hypotheses, conclusions and contributions as a knowledge base, formalized an coded in XML. This machine-readable formalized structure should be processed by intelligent software agents, that way, permitting semantic retrieval and validating new knowledge contained in the article. The authoring tool should also permit to formally relate the knowledge base to Web public ontologies, for example, the UMLS – the Unified Medical Language System (http://www.nlm.nih.gov/pubs/factsheet/umls.html) and also linking it to other electronically published scientific articles, not just as usual citations but also as *qualified* citations, in which the reasons to cite and the relationships between this specific scientific article and its citations are made explicit.

The objective of this research is to investigate the potential of Web published scientific articles, conceived not only as texts, but also as a machine readable knowledge bases, explicitly and formally related to a Web-based public ontology representing the assented knowledge of a specific domain. The goal is to enhance Web electronic publishing to embody new facilities provided by the Web environment in the Semantic Web initiative context. As a first step towards this objective a model of a article's scientific methodology structure in XML, is presented.

During World War II, Vannevar Bush, a prominent American scientist, was in charge of coordination hundreds of scientists working in the US military effort. The information overload and the scientists difficulties to deal with the growing number of scientific articles that were published grasped Bush's attention. In an article published in 1945, he proposed a mechanical device, which he called "Memex", to help scientists to deal with the growing amount of scientific articles that were published. The mechanism proposed by Bush worked like the human brain, "as we may think", by *relating* different scientific articles. The ideas of Bush, although never made concrete due to technological limitations back then, were considered as a foresight of hypertext and its cognitive potentialities.

In a new technological environment, some Bush's ideas were undertaken by Tim Berners-Lee. He created hypertext and the Web. Today, the cognitive potentialities of hypertext are outlined by Pierre Lévy (1993). He discusses the Web as a cognitive device which will aid humankind to enlarge science and culture.

The Web is now the preferred channel for information interchange among scientist. Lawrence (2001) emphasizes how Web published scientific articles enhance visibility. With the rise of the Internet it is actually very common for scientists to publish their research results on the Web. At the same time, there is a growing number of large, online, machine-readable, knowledge bases named ontologies, in different knowledge areas.

The Semantic Web Initiative (Berners-Lee, 2001) aims to develop new standards to cope with the growing amount of information available on the Web and the difficulties to retrieve relevant information for specific activities. This can be eased by publishing not only human readable but also machine readable documents and by mobilizing software agent to help humans to perform these tasks. One of the main features of this initiative is enable software agents to perform inferences basing themselves on coded knowledge embedded in digital documents.

Health Science literature is generally highly structured, as can be seen in the "instructions for authors" of many journals[*]; but this structure is the text structure. There may also be, embedded in the text structure, a specific structure corresponding to scientific method and scientific reasoning, a "deep structure", as stated by Noam Chomsky (1981), containing the statement of the problem, the assumptions, the hypotheses, the methodology, the conclusions etc, different from the text structure; a formal knowledge base could be derived from this structure. Nowadays in the Web environment there are several authoring tools to electronic publishing research results, to provide archiving and access as Web repositories and digital libraries, to provide metadata exchange and re-use. In contrast there are few experiences with the aim of providing authors with tools that permit a- electronic publishing research results both as text and as a knowledge base; b- to relate formally and explicitly this knowledge base to other scientific articles and to public Web ontologies, which store the established corpus of knowledge of a specific domain; and c- to permit other researchers to navigate throughout a semantically rich network of enhanced text/ontologies articles and to check their validity and coherence, to compare, comment and semantic query them. Web publishing of scientific articles as described can be a cognitive tool in which its potential is far from being explored. We envisage a network of enhanced digital Web Published scientific articles, in which different tools could be accomplished, as a mechanism to improve, formalize and speed the development of science.

[*] Uniform requirements for manuscripts submitted to Biomedical journals: writing and editing Biomedical publications, http://www.icmje.org.

This particular research, in this initial phase, is an exploratory research. A bibliographic and Web research was performed to identify similar experiences, projects and specific ontologies cases that permit outlining a conceptual framework to support the proposed model. This research emphasizes topics such as the expansion of science, the scientific communication process and social mechanisms to validate and check consistency of research results, "invisible colleges", their role and the mechanisms used to validate research results, the motivations of researchers to cite other authors.

## 2 Theoretic approaches, similar researches, projects and experiences

What are the methods to achieve the truth in Science? These questions date back to Greek Philosophy with Rhetoric, Dialectics and Sophistic, through Medieval Scholastics. A branch of this discussion with important contributions came at the Modern Age, with the establishment the scientific method by Francis Bacon and Descartes. Later on, Scientific Epistemology, the conditions of Truth, scientific paradigms and paradigm change were studied by Popper and Kuhn. Contemporarily, criticisms to the bases of the scientific method came from Maturana and Varela and from Morin and Freyrabend.

Another important branch of that discussion came about with the recognition that the language itself is a tool to reach knowledge, in Analytical Philosophy. Formal Logic has important contributions from Frege, Carnap and Wittgstein and more recently from Hempel. Pierce's Semiotic also addresses the question of reasoning modes.

The founding phenomenon of Information Science is the so-called "information overload", characteristic of the modern society. Scientific Communication and, specifically, bibliometry and citation analysis, were central themes studied by Information Science to deal with this phenomenon.

"Information overload" hit critical dimensions with the rise of the Internet in the 1990's. The Internet has become a vast net of few structured information, which restricts its usefulness. Semantic Web initiative (http://infomesh.net/2001/swintro/) is an answer to this situation. It promises digital documents with imbedded knowledge – the Web based ontologies – readable by intelligent software – known as "agents", "crawlers", "spiders" - to process this knowledge.

At the present date, the Internet made it possible for scholars themselves to electronically publishing research results, thus accelerating the publication cycle and increasing visibility. Citation analysis, in which pioneering work was done by Eugene Garfield in ISI, now with the Internet has acquired new dimensions. New possibilities of electronic publishing and crosslinking of research literature have been explored by Open Citation/Citebase (http://opcit.eprints.org/) and Citeseer (http://citeseer.isp.psu.edu/citeseer.html) Projects, providing scientists with new tools to analyze the research literature, backtracking contributions to a topic, based on bibliometry and citation analysis.

The point of view that the scientific article is a Rhetoric process is explored by Gross (1990): scientists make efforts to convince their peers of the accuracy of their findings. Citing reasons is another widely explored research topic (Case, 2000) in Information Science literature. Bruno Latour (2000) also develops a critical and detailed analysis of the reasons to cite, which can be related to Gross assertions. .

The Scholarly Ontologies Project (http://kmi.open.ac.uk/projects/scholonto) and the Trellis Project (http://www.isi.edu/ikcap/trellis/) (Oliveira, 2004) adds new features besides citation analysis, providing mechanisms to assign comments to research papers, to formalize comments and to navigate through a network of electronic published research papers and to the comments on them. De Roure (2001), in a Report commissioned for EPSRC/DTI Core e-Science Programme, describes a rich Web environment consisting in information resources and intelligent software, comprising a future e-Science infrastructure, in the context of the Semantic Web Initiative.

Bibliometric and citation methods, on which former projects are based, depend largely on human reading and writing. New methodologies, named ontologies, in the context of the Semantic Web Initiative, have been developed to record, exchange and use knowledge throughout the Web. Several web ontologies have been developed with focus for the Medical and Bio-medical areas (http://lhncbc.nlm.nih.gov/lhc/servlet/Turbine/template/research,langproc,MedicalOntology.vm, http://mged.sourceforge.net/, http://www.ifomis.uni-leipzig.de/). The UMLS (http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html) is very important to this research. It is initially conceived as a terminological device, resulting from the merge of different medical knowledge sources, as thesauri, classifications, code sets, lists of controlled terms, etc, UMLS is a semantic network that is developing

towards a knowledge base, allowing the semantic categorization of a wide range of semantic entities, such as, organism, anatomical structures, biologic functions, chemicals, events, physical objects and concepts.

Other experiments, related to the goals of this research, are the attempts to use XML – Extensible Markup Language – to Web publish scientific articles. There are different proposals as the pioneering CML – Chemical Markup Language (Murray-Rust, 1999), SBML - System Biology Markup Language (Hucka, 2003), MathML - Mathematical Markup Language and also more general approaches as STMML – Scientific Technical and Medical Markup Language (Murray-Rust, 2002). Another important experiment is TEI – Text Encoding Initiative, http://www.tei-c.org/, – which uses XML to mark-up scholarly texts in literature and linguistics in order to ease the retrieval and preservation of electronic publishing. The DDI - Data Documentation Initiative (http://www.icpsr.umich.edu/DDI/codebook/index.html) - aims to establish an international XML-based standard for the content, presentation, transport and preservation of documentation for datasets in social and behavioral sciences.

All these experiments, however, address to the specificity of these scientific discipline languages and try to take advantage of XML facilities to enhance the representation of mathematical formulas, chemical molecules and biological compounds in scientific texts. They are not concerned with the scientific methodology elements such as the parts of a scientific article related with problems, hypotheses, methodology, results, conclusions.

# 3 The outlined model

Science aims to state valid and general assertions. Those assertions states relations among some causes and a consequence which always follows from those causes. The main importance of Science to mankind is that scientific assertions permit *foreseeing* facts. The Scientific method comes from Aristotle's writings, which establish the deductive method. At the beginning of Modern age, Francis Bacon states the inductive method as a secure basis for Science. Other important contributions came from Descartes and Galileo Galilei, who, with Bacon, emphasized empirical observations and experimental methods, establishing the basis of the Scientific Method.

The highly structured text format of scientific articles in areas such as Health Science is mainly concerned with clarity in order to improve human reading and comprehension. Yet, it could help to identify in the text elements such as the problem, hypotheses and conclusions. As an analogy to the model proposed by Chomsky to languages, the scientific article text "surface structure" is generally comprised of, as stated in:

> "*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"* (http://www.icmje.org).

Since Popper (2001), hypotheses are central in scientific methodology. Scientific methodology manuals (Marconi, 2004) stress the importance of hypotheses formulation to guide scientific discovery. Hypotheses are temporary explanation for a research problem, guiding empirical data collection, tests and experiences, to then be confirmed or refuted. A hypotheses states that, under certain conditions, certain consequences always follow certain causes. Also hypotheses are not general statements, but are restricted explanations to a somehow specific context of reality. Specific knowledge areas have also specific terminology to describe phenomena concerned with them. The Hypothetical-deductive method is largely accepted as the proper method of science. It states the following steps in a scientific research:

- step 1- facts or, more precisely, problematic facts;
- step 2- formalization of a research problem or question;
- step 3 - development of a hypotheses which is an (temporary) answer to the research problem;
- step 4- empirical testing of the hypotheses;
- step 5 - analysis of the test results;
- step 6 - conclusion: hypotheses ratification or refusal

Klahr & Simon (1999), commenting the predominance of the Hypothetical-deductive method in science since the critical view of Popper (2001), attract attention to the fact that important research activities just collect problematic facts and formulates initial hypotheses to explain them, simply going through steps 1 to 3. On the other hand, most empirical researches are preoccupied only with testing hypotheses formulated by someone else, performing steps 4 to 6. This discussion just emphasizes the central role of the hypotheses to scientific

research. All these discussions lead to an initial model of the "deep structure" of a scientific article in XML, as proposed:

```
<scientific_article_deep_structure>
    <fact>... </fact> ...                          (new phenomena)
    <problem> ... </problem>                       (question)
    <method>
            <methodology> ... </methodology>
    </method>
    <hypotheses>                                   (provisory answer)
            <contextual_condition> ... </ contextual_condition > ...
            <cause> ...
                    <link to knowledge base> ... <link to knowledge base>
            </cause> ...
.           <consequence> ...
                    <link to knowledge base> ... <link to knowledge base>
            </consequence> ...
    </hypotheses>
    <result> ... </result> ...                     (data resulting of controlled experiences
                                                    or empirically collected – also a link to
                                                    datasets of results)
    <conclusion> ...                               (hypotheses ratification or refusal)
            <link to knowledge base> ... <link to knowledge base>
    </conclusion> ...
    <citation>
        <bibliographic_reference> ... </bibliographic_reference>
        <link to bibliographic reference> ... </link to bibliographic reference>
        <reason_to_cite> ... </reason_to_cite>
    </citation> ...
</scientific_article_ deep_structure>
```

An article "deep structure" model, as a XML document, is composed by structural elements, hierarchically organized and mapped to XML elements, and by two types of relations, expressed as links: from an article "deep structure" to other Web published articles cited in it and from an article "deep structure" to an available Web ontology: in any element of the proposed structure there might be links connecting specific terms contained in article's text to controlled concepts in ontologies like the UMLS. The way those links are is another research problem. As an additional feature, special emphases is giving to the <citation> element which aims, besides link the article to its citations, at capturing the <reason_to_cite> element. Since it is an initial model more elements may be added, such as, details on the method and methodology or the standard methodological procedures used.

All the outlined elements, the public Web ontology representing the assented knowledge of a scientific area and the different Web published scientific articles, containing full-text and "deep structures, form a rich Web environment. This permits, with the aid of intelligent software agents, browsing and navigation, semantic retrieval, critical inquiring, semantic citation, comparison, coherence verification and validating a scientific article against Web public ontologies, which express the assented knowledge of a specific scientific area.

## 4 Conclusions

To publish scientific articles both as text and as machine readable knowledge bases is a promising approach, in which its possibilities should be deeply explored. The outlined model is an initial approach of a general model of aiming to formalize Scientific knowledge in a machine readable format. The model points towards a Sm-ML – Scientific Methodology Markup Language as a standard to formalize and code the knowledge embedded in a scientific article. There are plans to develop empirical experiments analyzing scientific articles, initially in Health Sciences, in order to validate the model, verifying the existence of the "deep structure", identifying its elements, verifying its conformance to the model proposed and enhancing it.

Formal "deep structure" of scientific articles, in the form of machine readable knowledge bases, would improve critical inquiry, semantic querying and validation of scientific contributions to Science. The proposed model aims at proposing a discussion among other researchers. A Sm-ML should be a collective and a top-down

construction. Many questions remain open. Is such a general model feasible? Is the formal proposed "deep structure" common to all scientific disciplines or must there be specific models to specific disciplines? Is it feasible to merge article knowledge base Sm-ML elements to other specific elements of scientific disciplines, using the name spaces facilities of the XML language? Another challenge would be to develop an authoring tool based on this model to capture the "deep structure" of scientific articles, as a by-product of the Web publishing process of scientific articles. Is this feasible? Will researchers willingly use such an authoring tool while writing, describing, indexing and publishing scientific articles? At the present stage, all these and many others are still open questions to be answered and researches to be developed.

## References


Berners-Lee, Tim, Hendler, James, Lassila., Ora. (2001) The semantic web. Scientific American, May, 2001. Available in <http://www.scian.com/2001/0501issue/0501berners-lee.html>. Accessed in May 24 2001.

Bush, Vannevar. (1945). As we may think. The Atlantic Monthly, 176, 1, 101-108. Available in <http://wwwtheatlantic.com/unbound/fiasbks/computer/bushf.htm>. Accessed in Jan. 9 2005.

Case, Donald O.; Higgins, Georgeann M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in Communication. Journal of the American Society for Information Science, 51, 7, .635-645. Available at <http://www.periodicos.capes.gov.br>. Accessed in Sept. 9 2004.

Chomsky, Noam. (1981). Regras e representações: a inteligência humana e seus produtos. Ed.Zahar.

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

Gross, Alan G. (1990). The Rhetoric of Science. Harvard University Press.

Hucka, Michel; Finney, Andrew; Suro, Hubert; Bolouri, Hahmid. (2003). System Biology Markup Lnaguage (SBML) Level 1: structures and facilities for basic model definitions. <http://www.sbml.org/specifications/sbml-level-1/version-2/sbml-level-1-v2.pdf> . Accessed in Dez. 1, 2004.

Klahr, David; Simon, Herbert A. (1999). Studies of scientific discovery: complementary approaches and convergent fidings. Psycological Bulletin, 124, 5, 524-543. Available in <http://www.psy.cmu.edu/psy/faculty/Kands99.pdf>. Accessed in Mar. 05 2005.

Kircz J. G. (2002). New practices for electronic publishing 2: New forms of the scientific paper. Learned Publishing, 15, 1, 27-32. Available in <http://www.ingentaconnect.com/searching/Expand?pub=infobike://alpsp/lp/2002/00000015/00000001/art00004>. Accessed in Nov. 27 2004.

Klahr, David; Simon, Herbert A. (1999). Studies of scientific discovery: complementary approaches and convergent fidings. Psycological Bulletin, 124, 5, 524-543. Available in <http://www.psy.cmu.edu/psy/faculty/Kands99.pdf>. Accessed in Mar. 05 2005.

Latour, Bruno. (2000). Ciência em ação: como seguir cientistas e engenheiros sociedade afora. Editora UNESP.

Lawrence, Steve. Online or invisible. (2001) Nature, 411, 6837, 521. Available in <http://www.neci.nec.com/~lawrence/papers/online-nature01/>. Acessed in June, 13, 2004.

Lévy, Pierre. (1993). As tecnologias da inteligência: o futuro do pensamento na era da informática. Editora. 34. (Coleção Trans).

Li, Gangmin; Uren, Victoria, Motta, Enrico, Shum, Simon Buckinghan, Dominique, John. (2002). ClaiMaker: weaving a Semantic Web of research papers. Proceedings of the 1 INTERNATIONAL SEMANTIC WEB CONFERENCE, June 2002, Sardinia. Available at <http://kmi.open.ac.uk/projects/scholonto/papers.html>, accessed in June 5, 2004.

Marconi, Marina de Andrade; Lakatos, Eva Maria. (2004). Metodologia científica. Editora Atlas.

Murray-Rust, P; Rzepa H. S. (2002). STMML. A markup language for Scientific, Technical and Medical Publishing. Data Science Journal, 1, 2, 128-193. Available in <http://journals.eecs.qub.ac.uk/codata/journal/contents/1_2/1_2pdfs/ds121.pdf >. Accessed in Dez 1 2004.

Murray-Rust, P; Rzepa H. S. (1999). Chemical Markup, XML and the worldwide web. I: basic principles. Journal of Chemical Information and Computer Science, 39, 928-942.

Oliveira, Edgard Costa. (2004). Towards a new authoring environment: overview of some ontology-based systems. In: Jan Engelen (Ed.), Proceedings of the 8th ICCC INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, pp. 121-130. Dep. of Information Science and Documentation (CID)/UNB.

Rzepa H. S.; Murray-Rust, P. (2001) A new publishing paradigm: STM articles as part of the semantic web. Learned Publishing, 14, 177-182. Available in <http://www.ingentaconnect.com/searching/Expand?pub=infobike://alpsp/lp/2001/00000014/00000003/art00003>. Access in Nov. 30 2004.