RECIIS
Electronic Journal of Communication Information & Innovation in Health
[www.reciis.cict.fiocruz.br]
ISSN 1981-6286

Original Articles

Ontological and conceptual bases for a scientific knowledge model in biomedical articles

DOI: 10.3395/reciis.v3i1.240en

Carlos Henrique Marcondes
Information Science Department, Federal Fluminense University, Niterói, Brazil
marcon@vm.uff.br

Marília Alvarenga Rocha Mendonça
Information Science Department, Federal Fluminense University, Niterói, Brazil
marilaalvarenga@terra.com.br

Luciana Reis Malheiros
Physiology and Pharmacology Department, Federal Fluminense University, Niterói, Brazil
malheiro@vm.uff.br

Leonardo Cruz da Costa
Computing Department, Federal Fluminense University, Niterói, Brazil
leo@dcc.ic.uff.br

Tatiana Cristina Paredes Santos
Biomedical Institute, Federal Fluminense University, Niterói, Brazil
tatianacps@biof.ufrj.br

Abstract
Scientific articles published in electronic format are knowledge bases, specially in Medicine. An obstacle to semantic processing of this knowledge by computers is that, in spite of their digital format, articles are in text format for human reading and processing. A model is proposed to electronic publishing scientific articles both in textual format and in machine "understandable" format, in ontology format. The model is based in insights from Philosophy and Methodology of Science and in the results of analysis of 75 scientific articles in Medicine. Software agents can process the content of an article thus enabling semantic retrieval, consistence checking and the identification of new discoveries.

Keywords
medical knowledge; knowledge representation; ontologies; electronic publishing; scientific communication

Introduction

Since the period of oral culture was left behind by the invention of writing, for many centuries human knowledge, especially scientific knowledge, was recorded in documents.

We are at the threshold of a deep transformation in the means that humanity has at its disposal for recording, keeping and disseminating information. Nowadays we can not only record this knowledge digitally, stored in mass memory devices, we can also disseminate it on a large scale with computer networks. A device such as the pen drive can store more than 4 GB of information, making it possible to carry a gigantic library in one's pocket. More significant than these questions, however, is the fact that this knowledge is no longer only coded in a text format, which people can read, as in text documents in WORD or PDF formats, but also in the real novelty of formats "intelligible" to programs, giving these increasing capacity to make "inferences", "decisions" and "reasoning" about the content of these documents. This is the proposal of the Semantic Web project (Berners-Lee 2001).

One of the bases of the Semantic Web is formed by ontologies. An ontology is an informational model describing and representing a domain of specific knowledge, through concepts corresponding to the relevant objects in this domain, representing its structure and its inter-relationships; this model must be understood by a community of users. The concepts are organized in class hierarchies in ascending levels of generality and they possess attributes and relationships among themselves. An ontology is represented in language "intelligible" to "software agent" programs and used by these to make inferences about the concepts of this domain. When representations of individual objects of a specific knowledge domain are added to the classes of that domain which form an ontology, there is a knowledge base. There are many user communities developing or using ontologies, especially in the biomedical area.

The passage from recording knowledge in a text format to recording it in formats "intelligible" to programs presupposes new ways of looking at it. Obvious and immediate questions that arise are: what is knowledge? What is it made up of? How can it be systematized? How can it be recorded in a format other than text? Although most of these questions have been asked by philosophy for centuries, the last question is a totally new problem and it wasn't practically asked while humanity only knew the text format, even when it was digital.

Scientific research, especially in the biomedical area, increasingly uses the computer as a tool. Growing quantities of data about genetic sequencing, proteomics, etc, are kept in computer databases (Stein 2008).

Coded knowledge in a format "intelligible" to programs allows them to be employed in tasks where computers/programs are clearly more efficient than we are. For large-scale use of "software agent" technology, scientific knowledge that is already recorded digitally but is still in a text format must be extracted and also represented in a format "intelligible" to programs.

The traditional and institutionalized way through which contemporary society records and disseminates scientific knowledge is by publishing articles in periodicals. Scientific articles form bases of knowledge, but to be read and processed by scientists, due to their text format. Processing this knowledge, by reading these articles, reviewing and quoting them, and by reproducing the experiments reported in them, including their content in lessons, teaching texts, manuals and treatises, is a slow social process. For years we have been working on the project of recording the content of electronically-published scientific articles in a format "intelligible" to programs. We proposed a "software" Web environment that allows the simultaneous publishing of articles electronically and conventionally, as text, and in an ontology format, as illustrated in Figure 1 below. This "software" Web environment interacts with the author through a structured dialogue and analysis of the article's text, extracting and representing knowledge contained therein as an ontology.

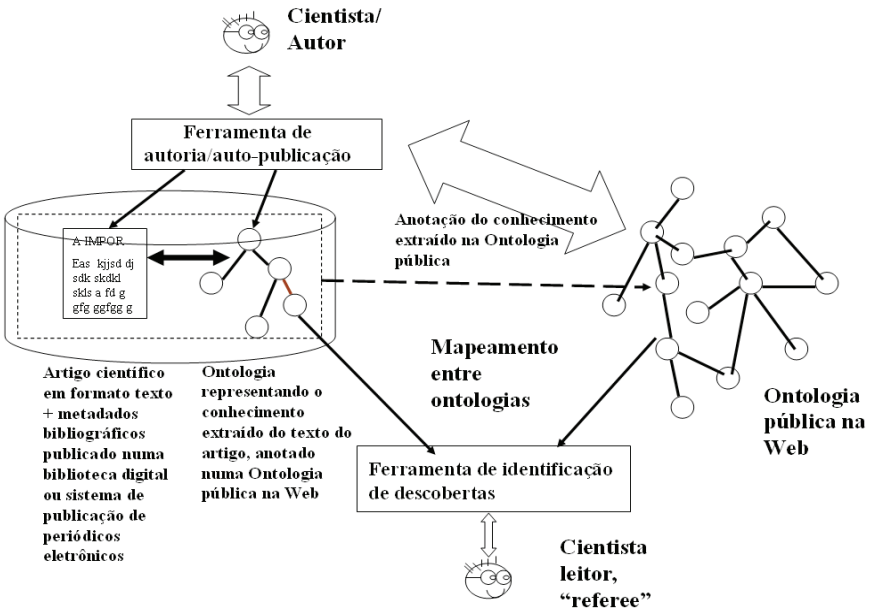

Figure 1 – Web authorship/self-publication environment.

We have created a model for the semantic elements that make up the knowledge content of an article based on the elements of the scientific method, as they appear in scientific articles. The aim of this model is to serve as a base for a new form of publishing articles in ontology formats, so as to make their content "intelligible" to programs, thus allowing this content to be processed in a much more sophisticated and useful way than in conventional information recovery, databases, "data mining" or statistics programs, helping researchers with the semantic recovery of information, the assessment of coherence, the identification of gaps in scientific knowledge and new discoveries.

The initial proposal of the model was based on the concept that scientific knowledge consists in proposing and proving the existence of *relations between phenomena*, unknown until then. That is how Miller (1947) defines scientific knowledge: "The above remarks imply that science is a search after internal relations between phenomena". A phenomenon can be defined as "… an event or a process such as it appears to some human subject: it is a perceptible fact, a sensible occurrence" (Bunge 2004, p. 173).

This model was constructed based on the strong theoretical references of philosophy of science and scientific methodology. This was complemented and validated by a set of 75 health sciences articles, extremely rich material that made it possible to perfect the model and verify in practice the nuances of how scientific knowledge in this area is recorded and communicated in scientific articles. This model was formalized in a Knowledge Content Ontology in articles[1]. Future research developments will consist of using the model as a base for constructing a self-publication Web base (Cost 2006) of scientific articles and for developing programs that compare the content of scientific articles recorded according to the model with knowledge recorded in medical ontologies, such as the UMLS – Unified Medical Language System[2] or the Gene Ontology[3], with a view to identifying new scientific discoveries (Malheiros 2005). The most complete version of this initial proposal is described in Marcondes (2007).

The initial model, however, had to be perfected following the conclusions of the analysis of the last set of articles analyzed. It is a set of 15 articles which, unlike the previous sets, has an internal coherence and visible inter-relations within itself. They are the so-called "key publications" by the group of three researchers, Elizabeth Blackburn, Carol Greider and Jack Szostak, who received the Albert Lasker Award for Basic Medical Research in 2006 (http://www.laskerfoundation.org/index.html). The articles cover the period from 1978 to 1999 and, as a whole, report the sequence of steps in the discovery of the telomerase enzyme (Blackburn 2006), its fundamental role in cell reproduction, its influence in processes such as cell ageing and the appearance of cancer. The articles have a strong relationship to each other, clearly showing the process of *initial identification of a new phenomenon, the aggregation of new knowledge so as to characterize it scientifically, until its complete identification and the investigation of its possible relationships with other phenomena.* The results of the analysis of these articles show facts that we had not encountered until then with the analysis of the previous material and that our model did not handle. The analysis of this set of articles, therefore, imposed the need for review and perfection of the initial model, so as to make it more complete, broad and robust. This is the purpose of this article.

It is structured thus: after this introduction, section 2 discusses the conceptual setting that forms the base for the proposed model; then, section 3 discusses the theoretical aspects that form the base for the model; then, section 4 discusses the new characteristics of the set of articles that won the 2006 Lasker Award and that led to review of the model; section 5 then presents the reviewed model; finally, section 6 presents conclusions and future directions of research.

Materials and methods

For the proposal of the model, theoretical support was sought for in disciplines such as information science, especially scientific communication, science methodology, philosophy of science and computer science. Seventy-five medical articles were analyzed, subdivided into the following groups: 20 articles from the periodical Memórias do Instituto Oswaldo Cruz, 20 articles from the periodical Brazilian Journal of Medical and Biological Research, both available on the SciELO portal and chosen from the list of most consulted articles from both publications, available on the publications' sites; also, 20 articles about stem cells were researched, chosen based on three important articles that review the theme.

Finally, 15 articles were analyzed out of the so-called key publications by the group of researchers who won the Albert Lasker Basic Medical Research Award in 2006. The text by Charlton (2006) shows that scientific awards can help to identify the "revolutionary science" elite, in Kuhn's words, which was especially interesting for this research.

For this last group, a methodological procedure was to order the articles chronologically (see Attachment 1) and use the 2006 article where the three winners of the Lasker Award (Backburn 2006) comment the trajectory of the research which culminated with the discovery of the enzyme telomerase as a reading guide for the 15 articles that make up the key publications.

The analysis sought to identify phenomena described in the article or to establish relationships between phenomena. After the identification of phenomena and their relations in the text of each article, it was verified whether the concepts corresponding to the phenomena and their relations existed in the UMLS. The results of the analysis were recorded in a specific form.

The medicine area was chosen due to the fact that its scientific articles followed a strict formal model in their texts, with sections defined according to the IMRAD – Introduction, Method, Results and Discussion – model, recommended by the International Committee of Medical Journals Editors[4] for scientific articles in biomedical periodicals, thus facilitating analysis.

Conceptual setting

Scientific knowledge, as conveyed by periodical articles, consists in formulating "scientific statements" about phenomena (Bunge 2004, p. 173) or about relations between phenomena (Miller 1947). Scientific statements, as will be seen in the examples in section 4, reflect an increasing degree of certainty, appropriation and framing of the scientific phenomenon and its interrelations in the conceptual setting or system of concepts that makes up knowledge in a certain scientific domain. This increasing degree of certainty moves in the direction of what Bungue (2004, p. 3) considers to be the aim of science, that is, to answer "why questions", to seek explanations for phenomena.

The classical views of how science is done, presented by scientific methodology manuals and the philosophy of science in Popper (2001) and Hempel (1995), separate the procedures and reasoning employed in scientific discoveries from the methodological explanation of scientific facts. There is an emphasis on the linguistic, logical and formal aspects of science, which originates in logical positivism in the 19th century (Marcondes 2004).

The views situated within the "justification logic", whether they belong to the logical positivists or their critics, like Popper, are excessively formal and the examples analyzed show that, in a state of the art research area such as cellular biology, they provide little help in understanding the research and discovery practices that led to the scientific constructions leading to the discovery of telomerase. Thus, the material we analyzed, as will be shown in section 4, has much in common with the positions of Aliseda (2004), Klahr and Simon (1999) and Thargard (1993).

The distinction between "discovery logic" as opposed to "justification logic" is emphasized in Atocha (2004), in a criticism of the positivist views, which state that: "The context of discovery is taken to be purely psychological" (Aliseda 2004, p. 340).

Or:

> Reinchenbach's distinction between the contexts of justification and of discovery has left out of its analysis – especially from a formal point of view – a very important part of scientific practice, that which includes issues related to the generation of new theories and scientific explanations, concept formation as well as aspects of progress and discovery in science (Aliseda 2004, p. 341).

Thargard (1993, p.176) criticizes both the inductive method, preferred by logical positivists, and Popper's hypothetical-deductive method, calling both myths. He states that, on the contrary:

> In well-trod areas of investigation, it may be possible to form a Sharp hypothesis and then test it. But when novel topics are being pursued, researchers in psychology and other fields cannot always start with hypotheses sharp enough to be tested. Often some vague ideas will lead to the collection of some data, which then suggested a refinement of an existing hypothesis. Or results are very different from what was expected may spur abductive formation of a new hypothesis that can then be subject to further test. (Thagard 1993, p. 177).

Klahr and Simon (1999, p.8) also criticize Popper and give special emphasis to the discovery context:

> In science there is an important, and extremely common, form of experiment, at times referred to somewhat dismissively as "exploratory," that is guided by no specific hypothesis to be tested, and no clear control condition, but only a vague and general direction of inquiry. The goal of exploratory experiments is to permit phenomena to appear that will invite exploration or suggest whole new forms of representation or generate new hypotheses.

> The contemporary literature on research methodology is dominated by the notion, promulgated by Popper (1959) among others, that the purpose of observation in general, and experiment in particular, is to test hypotheses in order either to falsify or validate them. In contrast to this position, we have argued that much of the important empirical work in science is undertaken - to use Reichenbach's phrase - in the context of discovery rather than the context of verification (see Simon, 1973). That is, a major goal of empirical work in science is to discover new phenomena and generate hypotheses for describing and explaining them, and not simply to test hypotheses that have already been generated. Indeed, theories cannot be tested until they have been created, and creation takes place in the context of discovery, not verification.

We are interested in seeking a model to represent the knowledge contained in the text of scientific articles in a format "intelligible" to programs, allowing these programs increasing degrees of "inference", "decisions" and "reasoning" about the content of these articles.

Historically, humanity's intervention in its environment is increasingly indirect, using different tools. Especially in science, the tools used are cognitive tools. Formally representing scientific knowledge in general, and in the health sciences in particular, in a format "intelligible" to programs means, in practical terms, developing ontologies. These ontologies will not only be more useful but they will also more accurately reflect the scientific discoveries in this area. That is, the more they correspond or are analogous to reality in this area, the more they can be used as scientific models of this reality, models that can be processed by computers to test hypotheses, make comparisons, diagnosis, identify inconsistencies, etc, becoming cognitive tools and instruments for the advancement of research and scientific knowledge.

In logic, the term inference means the process (also known as "reasoning") of deriving true consequences from premises that are true or held to be true. Simulating inference processes in a computing environment is when information is fed into a system and this returns other information that is in any way related to the supplied information.

Scientific knowledge, as conveyed by periodical articles, consists in formulating, through language, propositions containing *scientific statements* about phenomena (Bunge 2004, p. 173) or relating phenomena to each other (Miller 1947), or relating a phenomenon to its characteristics.

We sought conceptualizations of "phenomenon" that could serve to formally represent scientific knowl-

edge in medicine, the object of this research. A definition of phenomenon used in philosophy and science methodology is: "… an event or a process such as it appears to some human subject: it is a perceptible fact, a sensible occurrence" (Bunge 2004, p. 173).

Scientific statements reflect an increasing degree of certainty, appropriation of the scientific phenomenon and its interrelations, in the conceptual setting or system of concepts that make up knowledge in a certain scientific domain.

Two forms of scientific knowledge as relations were identified in the analysis of the literature that made up the empirical material:

The first form is the appropriation of a phenomenon with its progressive characterization through the systematic collection of "scientific statements" (Bunge 2004, p. 173) in the form of propositions *relating the phenomenon to its characteristics* (Dalhberg 1977, p. 16). More specifically, Dahlberg calls "essential characteristics" those which characterize or give identity to a certain phenomenon and without which this phenomenon would lose its identity (GUARINO 1997). These are characteristics that make up what Aristotle calls essence or essential attributes of the substance (Chaui 2005). Through the systematic collection of its scientifically tested characteristics a phenomenon is progressively identified and integrated to the system of concepts of a scientific domain.

The second form of scientific knowledge is the identification and establishment of relations between distinct phenomena, which had been unknown until then.

Relations, in the form proposed above, also reflect scientific propositions' increasing degrees of certainty, from a question or problem – what is the mechanism that determines the synthesis of the telomeres' ends? – where one of the *relata* is unknown, passing through a hypothesis, where the relation between the *relata* is hypothetical – an enzymatic activity determines the synthesis of the telomeres' ends – until a conclusion – the telomerase enzyme determines the synthesis of the telomeres' ends – where the relation between the *relata* is proved by an experiment.

However, the word phenomenon is full of subjective connotations, which are necessarily incompatible with scientific knowledge due to their connection with phenomenology (Chaui 2005). This discipline of philosophy studies phenomena as perceived by an individual observer, in opposition to the true nature of things, the true being, that is, appearances in opposition to reality.

If the observation of a phenomenon can be distorted by the observer, as has already been widely discussed, if this is socially and historically conditioned and, in scientific terms, as Kuhn has shown, paradigmatically conditioned, if knowledge is progressively constructed by the individual, as Piaget says (1978), then scientific knowledge is an eminently social construction (Ziman 1979). As an institution, science has mechanisms that guarantee a high degree of consensus for a certain stage of knowledge at a certain historical moment.

Of course, this stage of knowledge is provisional, surpassable, limited, it was (socially) constructed. However, it corresponds to what is consensually identified as reality, to a stage of knowledge of reality. As this knowledge is shared and agreed socially, as we can use it to intervene in reality, use it as a cognitive tool to make predictions about reality, this knowledge – a mental representation or a record, inscription or document capable of being intersubjectively appropriated – *corresponds to reality.* Thus, scientific research, by observing and studying phenomena, supplies the elements – scientific knowledge – for the construction of an ontology that is always provisional, always being constructed, always unfinished. This must correspond as much as possible to a certain stage of knowledge of a given science, to reality itself.

Barry Smith (2002, p. 2), discussing the relation between science and ontology (as a domain of knowledge concerned with the nature of beings), states that ontology cannot "explain" nature as science. Its role is to come after explanations in order to describe, organize and systematize the knowledge obtained by scientific discoveries. This seems to be a place also occupied by information science.

Results

The process of increasing characterization and scientific appropriation of a phenomenon and the subsequent identification of relations between this phenomenon and others is illustrated in the following examples and can be accompanies by the titles of the articles in the 2006 Lasker award group, ordered chronologically in Attachment 1:

In the oldest analyzed article in the 2006 Lasker award group, Blackburn, E.H. and Gall, J.G. (1978), this aspect of gradual characterization of a new scientific phenomenon can be seen, not just in the title, but also in the article's aims, stated in its abstract: The extrachromossomal genes coding for the ribossomal RNA in the ciliated protozoan Tetrahymena thermophila we studied with respect to sequences occurring at their termini" (Blackburn 1998, p. 33).

- In a review article that shows a current setting of telomerase research, Cech (2004) states that the purpose of the research that led to the discovery of telomerase was to identify the *entity* (a term, by the way, often used in ontology modeling) responsible for the replication of the chromosomes' ends, which had previously been unknown and not characterized or integrated to the existing frame of knowledge: "Carol Greider, a graduate student in Liz Blackburn's group at the University of California, Berkeley, had chosen an ambitious PhD thesis project: identify the molecular entity responsible for replicating chromosome ends." (Cech 2004, p. 273).

- In the same article Cech (2004) explains the aim of Greider and Blackburn's research in the article that marks the discovery of telomerase: "The identification and characterization of this new enzymatic activity was the subject of Greider and Blackburn (1985)."

Another example taken from an article in the same group shows that telomerase was not yet clearly identified as an enzyme ("a terminal transferase-like

activity") nor was it clear what relation it had with the process of telomere complementation ("which adds the host cell telomeric sequence repeats onto recognizable telomeric ends"): "Based on all these considerations, the proposal was made that telomere replication involves a terminal transferase-like activity which adds the host cell telomeric sequence repeats onto recognizable telomeric ends". (Shampay et al. 1984), quoted in Greider (1985, p. 405).

- In a later article from the same set, an increasing scientific appropriation of the same phenomenon can be seen. Enzymatic transferase activity is identified, as seen in the title of the article: Greider, C.W. and Blackburn, E.H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. Cell v. 43, p. 405-413, 1985.

- In the same article, as commented in Cech's revision, the authors propose, still with a low degree of certainty, that telomerase activity should be framed within the known conceptual setting: "The authors made the reasonable proposal that the activity might be related to known terminal transferases, such as the enzyme that adds CCA to the 3_ ends of transfer RNAs." (Cech 2004, p. 273).

- Later, this enzymatic activity is identified, that is, framed within the system of concepts of this specific scientific domain, that is, in a classification of substances, and this enzyme is finally named telomerase as follows:

> Greider, C.W. and Blackburn, E.H. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell v. 51, p. 887-898, 1987.

- Later, the activity, function or role of telomerase as a catalyst and mold in the syntheses and complementation of telomere extremities is identified:

> Greider, C.W. and Blackburn, E.H. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. Nature v. 337, p. 331-337, 1989.

- Or: "Our results indicate the involvement of such sequence-specific telomeric DNA-protein interaction in cell or nuclear division" (Yu 1990, p. 131).

Once the phenomenon of telomere end complementation by the enzyme telomerase is characterized, proposals begin to arise establishing relations between this phenomenon and two others: cell senescence, that is, the finite number of times a cell is capable of reproducing, as a consequence of the progressive shortening of telomeres at each duplication and of the cell's incapacity to complement them through telomerase action, which leads to cell death; and the relationship of telomerase with cancer, identified as an uncontrolled process of cell duplication. These cases are described below:

- "These mutations also lead to nuclear and cell division defects, and senescence, establishing an essential role for telomerase in vivo" (Yu 1990, p. 126).

Or, in the article:

> - Allsopp, R.C., Vaziri, H., Patterson, C., Goldstein, S., Younglai, E.V., Futcher, C.W., Greider, C.W., and Harley, C.B. Telomere length predicts the replicative capacity of human fibroblasts. Proc. Natl. Acad. Sci. USA, v. 89, p. 10114-10118, 1992.

Or in the following excerpts and article: "This shortening has been proposed to play a role in signaling the cell cycle exit characteristics of senescent cells (14, 15), although a causal role has not been demonstrated." (Prowse 1993, p. 1493). In this article, the authors propose the existence of what is here called a "weak" relation by stating that a causal relation has not yet been demonstrated. In the other excerpt and in the article a relation is proposed between telomerase activity and cancer: "It has been proposed that the finite cell division capacity of human somatic cells is limited by telomere length (10). This is consistent with reports that telomerase activity is often high in cancer and immortalized tissue culture cells". (Mceachern 1995, p. 403).

And also in the article:

> - Rudolph, K.L., Chang, S, Lee, H.W., Blasco, M., Gottlieb, G., Greider, C.W., and DePinho, R.A. Longevity, stress response, and cancer in aging telomerase deficient mice. Cell v. 96, p. 701-716, 1999.

In the oldest analyzed article from the 2006 Lasker award group, Blackburn, E.H. and Gall, J.G. (1978), we see this aspect of the gradual characterization of a new scientific phenomenon, not only in the title, but also in the article's aims, stated in its abstract: "The extrachromossomal genes coding for the ribossomal RNA in the ciliated protozoan Tetrahymena thermophila we studied with respect to sequences occurring at their termini" (Blackburn 1998, p. 33).

A model for medical knowledge in scientific articles

Below, a model to represent, in a format "intelligible" to programs, the knowledge contained in scientific medical articles is proposed. The bases of the model are shown, from the content of articles that make up the empirical field and their analysis within the conceptual setting described above. As has already been mentioned, the earlier version of the model (Marcondes 2007) emphasized the role of hypotheses as relations among phenomena, following a more conventional view of science, based on the hypothetical-deductive method; the current version of the model, presented here, incorporated elements that show scientific knowledge being progressively constructed, through the characterization and incorporation of a new phenomenon.

Scientific articles are distinguished by the type of reasoning they employ when they conduct the argument about the phenomena discussed. There are theoretical articles and experimental articles. This classification is based on Hutchins (1997) and Gross (1990) and on texts

that have Pierce's (1977) view of abduction as a process of discovering new "insights" in science (Hoffman 1997, Magnani 2001, Paavola 2004, Aliseda 2004).

Theoretical-abductive articles are characterized by their discussion of broader questions. They critically analyze several previous hypotheses, showing their fragilities. These articles are the ones with the greatest potential for making contributions to science, as they discuss or question the existing paradigm (Kuhn 2003). Their contribution is a new hypothesis, indicating a new research route. The type of reasoning employed is abductive, that is, "insight" on solving questions unexplained by science and formulating new hypotheses to solve them.

Experimental articles necessarily contain an empirical experiment. They are divided into exploratory, deductive and inductive articles. They are characterized by discussions of questions in a limited scope. They do not discuss the directions of a scientific theory, but are limited to confirming and perfecting it. They always bring experimental results.

Experimental-exploratory articles have an exploratory character as they try to solve and characterize a phenomenon, working in the direction proposed by Dahlberg and formulating and proving propositions that characterize a phenomenon.

Experimental-deductive articles are based on relations between phenomena that have already been formulated, using quoted references, applying them, testing them and validating them in a specific context. **Experimental-deductive articles** are characterized by proposing and testing new relations between phenomena.

The structural text of health science articles follows the IMRAD model, as has already been mentioned. This structure corresponds to Chomsky's (1981) "surface structure" and the microstructure of "surface structure". On the other hand, the semantic components of an article, which make up the proposed model, correspond to Chomsky's "deep structure" and to Kintsh and Van Dijk's macrostructure: they are described below, identified in capital letters.

A PROBLEM expresses a lack, dissatisfaction or conceptual deficiency in the current state of knowledge in a domain. A PROBLEM can become research AIMS and, occasionally, the more precise formulation of a QUESTION that addresses the conceptual deficiency; this QUESTION can refer to a PHENOMENON (in the EXPLORATORY articles), or to two or more PHENOMENA involved in a RELATIONSHIP_BETWEEN_PHENOMENA or HYPOTHESIS. A HYPOTHESIS relates two or more PHENOMENA through a TYPE-OF-RELATION.

In an article, an author can formulate an original hypothesis – HYPOTHESIS(o) or take the previous hypothesis – HYPOTHESIS(p) – by other authors; in this case one or more quotations related to the HYPOTHESIS(p) – QUOTATIONS(h) – are made. An author can also analyze several HYPOTHESEs(p) to show that they are unsatisfactory as solutions for the PROBLEM and formulate his HYPOTHESIS(o). A theoretical article is justified simply by proposing a new HYPOTHESIS(o).

From the hypothesis, an EXPERIMENT capable of being empirically observed must be formulated. In an EXPERIMENTAL scientific article, this means having RESULTS observed according to a certain MEASUREMENT, in a certain CONTEXT, according to a certain METHODOLOGY. This CONTEXT in which the PHENOMENON(a) listed in the HYPOTHESIS is/are observed can take place in an ENVIRONMENT – a community or institution where the phenomenon occurs – SPACE – the place where the phenomenon occurs – TIME or era when the phenomenon occurs and the GROUP of individuals in which the phenomenon occurs. Every article also brings a CONCLUSION, in the form of a proposition about a phenomenon or about RELATIONS_BETWEEN_PHENOMENA.

The development of reasoning in an **abductive-theoretical article** follows this model:

- given a PROBLEM, with the following aspects and information

- the following authors/previous HYPOTHESES for its solution are not satisfactory

- therefore, we propose the following original HYPOTHESIS

The development of reasoning in a **deductive experimental article** follows this model:

- given a PROBLEM, with the following aspects and data

- the following authors formulated previous HYPOTHESES for its solution

- therefore, we chose the following (one of the previous HYPOTHESES);

we expanded and re-contextualized this previous HYPOTHESIS; we developed the following EXPERIMENT to test this previous HYPOTHESIS;

- the EXPERIMENT yielded the following RESULTS.

The development of reasoning in an **inductive experimental article** follows this model:

- given a PROBLEM, with the following aspects and data,

- a solution for this PROBLEM can be based on the following HYPOTHESIS,

- we developed the following EXPERIMENT to test this HYPOTHESIS,

- these tests yielded the following RESULTS.

The development of reasoning in an **exploratory experimental** article follows this model:

- given a PROBLEM or PHENOMENON that is not yet well characterized,

- we developed the following EXPERIMENT that allows the following characteristics of this PHENOMENON to be identified.

These schemes resulted in the current model or **Ontology of Knowledge in Scientific Articles – OCCAC,** also illustrated in Figure 2:

Classes: THEORETICAL articles
Have as components
PROBLEM
HYPOTHESIS(a)
HYPOTHESIS(o)
CONCLUSION(s) and
EXPERIMENTAL articles
Subclasses: EXPLORATORY articles
Have as components
PROBLEM
PHENOMENON
EXPERIMENT
CONCLUSION(s)
NDUCTIVE articles
Have as components
PROBLEM
HYPOTHESIS(o)
EXPERIMENT
CONCLUSION(s)
and
DEDUCTIVE articles
Have as components
PROBLEM
HYPOTHESIS(a)
EXPERIMENT
CONCLUSION(s)
semantic COMPONENTS of articles
PROBLEM
Subcomponents: AIMS
Research QUESTION
HYPOTHESIS (previous or new)
Subcomponents: PHENOMENON(A)
TYPE-OF-RELATION
REFERENCES (only in previous HYPOTHESES)
PHENOMENON (one, in EXPLORATORY articles)
Subcomponents: CHARACTERISTICS
EXPERIMENT
Subcomponents: RESULTS (quantitative data)
MEASUREMENT
CONTEXT
Subcomponents:
SPACE
TIME
Social GROUP
CONCLUSION(s)

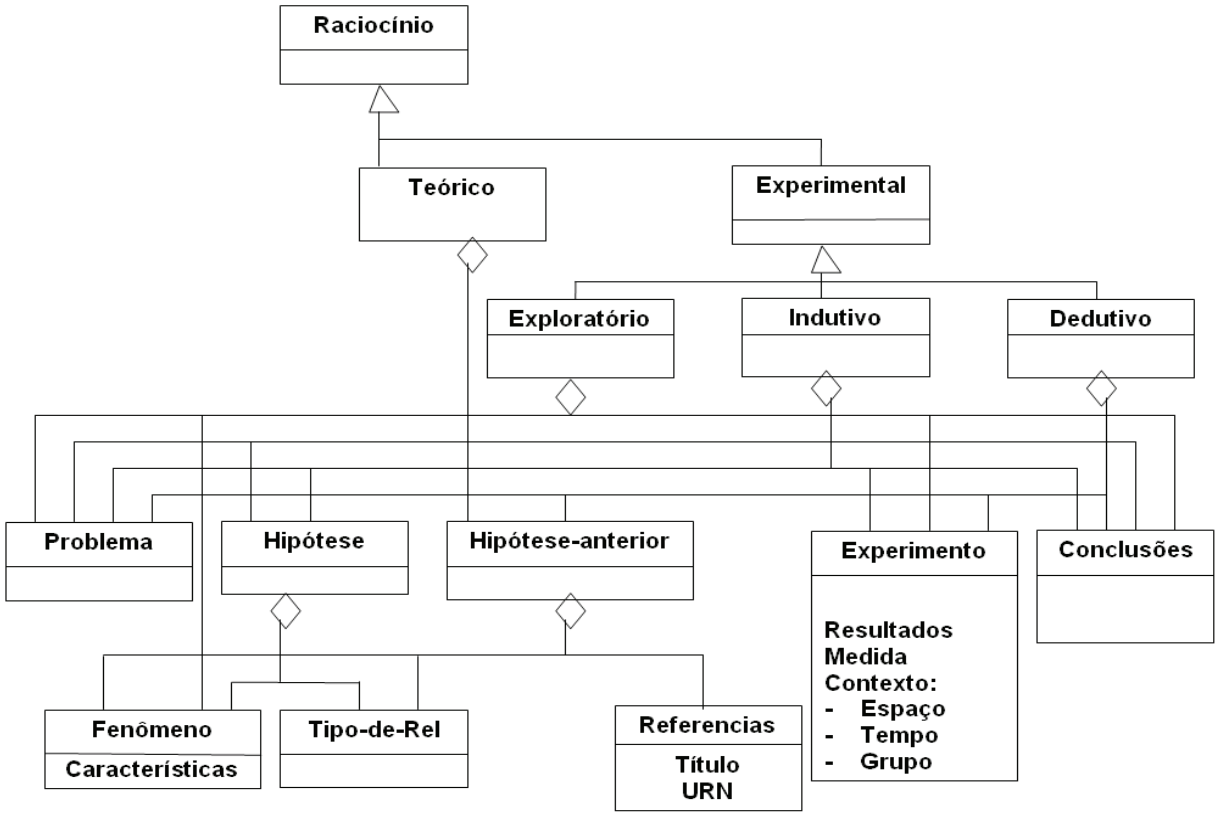

Figura 2 - Ontology of knowledge content in scientific articles

The analysis of the following article shows how semantic components of an article (in this case the hypothesis) are identified and recorded according to the model.

- 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], v. 98, n. 7, Oct. 2003.

3 steps:

Step 1 – Type of reasoning identified: deductive-experimental, that is, the article performs an experiment to show the prevalence of HPV, a hypothesis previously formulated by another author.

Step 2 – Semantic elements of knowledge are identified in the text, such as the hypothesis formulated by the author:

Hypothesis (previous)

Antecedent: HPV

Type of Relation: cause

Consequence: cervical pre-neoplastic and neoplastic lesions

Step 3 – Each of these elements is mapped in terms or relations of the public knowledge base, UMLS Semantic Network

Papillomavirus, Human

"Causes", UMLS Semantic network relation R147

Colonic Neoplasms

Conclusions

Information science comes from a long theoretical, methodological and practical tradition that converges on many of the current questions raised by the Semantic Web proposal and for the construction of ontologies. An area of recurring research in information science is the semantic processing of information by computers, with already historic contributions such as those by Shera (1957), Luhn (1960) and, more recently, by Gardin (2001) and Kajikawa (2006).

Information science can and should go beyond providing techniques and methodologies to allow access to the complete text of scientific articles in digital libraries and scientific deposits. The current indexing methodologies consist of assigning key words or isolated terms from a controlled vocabulary to bibliographical records, with no relation or semantic role among themselves. But the research in IC about the importance of relations and their semantic role is also historic, as the recent revision by Khoo (2007) shows.

The proposed model, by identifying types of relations (with their own semantics) and characteristics of the phenomena described in an article, allows automatic inferences to be made and, for example, sophisticated consultations to be resolved, such as:

- what (other) articles (also) contain hypotheses naming HPV as a cause of pre-neoplastic and neoplastic lesions in women?

- what articles contain hypotheses relating other factors, apart from HPV, as a cause of pre-neoplastic and neoplastic lesions in women?

- what articles identify other characteristics related to the structure of the ends of linear rDNA molecules?

- what articles identify characteristics of the telomere replication phenomenon that can be linked to enzymatic activity?

The practical implementation of the model for recording content of scientific articles as described presupposes the development of a whole set of "software" tools that process structured contents as proposed. It is, in reality, a research program. We believe that a solid starting point for this is to establish the model proposed here. Two of these applications in early stages of visualization are described in Malheiros (2005) and Costa (2006), and represented in Figure 1.

The model shows the benefits of a semantically-rich format for recording the content of scientific articles, based on relations, making it possible for software agents to make inferences through them. Forsythe (1989), talking about the first experiments in constructing specialist systems in the 1970s and 1980s, calls the acquisition of knowledge *the bottleneck of the construction of specialist systems*. Acquiring the construction of knowledge directly from authors/researchers and their scientific articles, which already have a high degree of knowledge formalization, as in the present proposal, can prove to be a promising alternative. With the support of adequate software tools, the content of scientific articles can be extracted as a sub-product of the authorship/self-publication of scientific articles by the authors themselves, which is fairly common nowadays when they submit their works to deposits, electronic publications or digital libraries. The vision outlined here (Figure 1) will result in electronic publications with a much richer potential for semantic treatment of their content by software agents than what it possible in current electronic text publications.

Notes

1. Available at www.professores.uff.br/marcondes/Scientific_reasoning.owl

2. Available at www.nlm.nih.gov/pubs/factsheet/umls.html

3. Available at m www.nlm.nih.gov/pubs/factsheet/umls.html

4. Available at www.icmje.org

Bibliographic references

Aliseda A. Logics in scientific discovery. Foundations of Science. 2004; 9: 339-63.

Berners-Lee T, Hendler J, Lassila O. The semantic web. Sci Am. 2001 May. Disponível em <http://www.scian.com/2001/0501issue/0501berners-lee.html>. Acesso em 24 maio 2001.

Blackburn EH, Greider CW, Szostak J. Telomeres and telomerase: the path from maize, Tetrahymena and yeast

to human cancer and aging. Nature. 2006 October; 12(10).

Bunge M. Philosophy of Science. New Brunswick, London: Transaction Publishers; 1998.

Cech TR. Beginning to understand the end of the chromosome. Cell. 2004 Jan.; 116:273-9.

Charlton BG. Editorial. Scientometric identification of the elite 'revolutionary science' research institutions by analysis of trends in Nobel prizes 1947-2006. Medical Hypotheses. 2007; 68:931-4.

Chauí M. Convite à Filosofia. São Paulo: Érica; 2005.

Chomsky N. Regras e representações: a inteligência humana e seus produtos. Rio de Janeiro: Ed. Zahar; 1981.

Costa LC da. Uma ferramenta para edição, extração e representação do conhecimento contido em artigos científicos publicados na Web. Projeto de Tese de Doutorado para ingresso no PPGCI UFF/IBICT. Niterói; 2006.

Dahlberg I. Ontical structures and universal classification. Bangalore: Sarada Ranganathan Endowment for Library Science; 1978.

Forsythe DE, Buchanan BG. Knowledge acquisition for expert systems: some pitfalls and suggestions. IEEE Transactions on Systems, Man and Cybernetics. 1989 May/Jun.; 19(3):435-42. Disponível em: <http://ieeexplore.ieee.org/iel1/21/1336/00031050.pdf?tp=&isnumber=&arnumber=31050>. Acesso em: 21 abr. 2008.

Gardin J-C. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. In: Colloques Isko-France. Filtrage et résumé automatique de l' information sur les réseaux, Conference invitee, Univesité de Nanterre – Paris X; 2001. (Conference proceedings).

Gross AG. The Rhetoric of Science. Cambridge, EUA; London, Inglaterra: Harvard University Press; 1990.

Guarino N. Some organizing principles for a unified top-level ontology. New version of paper presented at AAAI Spring Symposium on Ontological Engineering, Stanford University; March 1997.

Hempel K. Aspects of scientific explanation: and other essays in the philosophy of science. New York:Free Press; 1965.

Hutchins J. On the structure of scientific texts. In: UEA Papers in Linguistics, 5 th., 1977, Norwich. **Proceedings.** Norwich, UK: University of East Anglia, 1977. p. 18-39. Disponível em: <http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>. Acesso em: 30 mar. 2006.

Kajikawa Y, Abe K, Noda S. Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords. J Inform Sci. 2006; 32: 511-24.

Khoo C, Na JC. Semantic Relations in Information Science. Ann Rev Inform Sci Technol. 2007; 157-228.

Kintsh W, Van Dijk TA. Towards a model of text comprehension and production. Psychol Rev. 1972; 84(5):363-93.

Klahr D, Simon HA. Studies of scientific Discovery: complementary approaches and convergent findings. Psychol Bull. 1999; 125(5): 524-43.

Kuhn TS. A estrutura das revoluções científicas. São Paulo:Perspectiva; 2003. (Série Debates Ciência).

Luhn H. Keyword in Context Index for Technical Literature. American Documentation. 1960; 11(4): 288-95.

Malheiros L. A identificação de novas descobertas científicas através da análise do conhecimento contido em artigos científicos. Projeto de Tese de Doutorado para ingresso no PPGCI UFF/IBICT. Niterói; 2005.

Marcondes D. Filosofia analítica. Rio de Janeiro:Jorge Zahar; 2004. (Coleção Passo a passo).

Marcondes CH, Mendonça MAR, Malheiros CLC da, Santos TCP, Pereira LG. Representing and coding the knowledge embedded in texts of Health Science Web published articles. In: Chan L, Marten B. Ed. ICCC ElPub - International Conference on Electronic Publishing, Viena, Austria; 2007. Disponível em <http://elpub.scix.net>.

Miller DL. **Explanation versus description.** *Philosoph Rev.* 1947; 56(3): 306-12.

Piaget J. Psicologia e Epistemologia: por uma teoria do conhecimento. Rio de Janeiro: Forense; 1978.

Popper K. A lógica da pesquisa científica. São Paulo: Ed. Cultrix, Ed. USP; 2001.

Shera JH, Kent A, Perry JW, editors. Information systems in documentation. New York: Interscience Publishers; 1957. (Advances in Documentation and Library Science, v. 1).

Smith B. Beyond concepts: ontology as reality representation. In: FOIS - International Conference on Formal Ontology and Information Systems, Turin; nov. 2004. Disponível em: http://ontology.buffalo.edu/bfo/BeyondConcepts.pdf. Acesso em: 16 mar. 2007.

Smith B. Ontology and information systems. 2002. Disponível em: <http://ontology.buffalo.edu/ontology(PIC).pdf>. Acesso em: 26 maio 2008.

Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nature Rev Gen. 2008 Sept.; 9:678-88.

Thagard P. Computational Philosophy of Science. Cambridge: The MIT Press; 1993.

Ziman J. Conhecimento público. Belo Horizonte: Itatiaia, São Paulo: Ed. da Universidade de São Paulo; 1979.

Attachment 1 – "Key publications" – Lasker Prize, 2006, chronological order

Year	Article
1978	Blackburn, E.H. and Gall, J.G. (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. ***J. Mol. Biol.*** 120: 33-53.
1987	Szostak, J.W. and Blackburn, E.H. (1982) Cloning yeast telomeres on linear plasmid vectors. ***Cell*** 29: 245-255.
1983	Murray, A.W. and Szostak, J.W. (1983) Construction of artificial chromosomes in yeast. ***Nature*** 305: 189-193.
1984 JAN	Shampay, J., Szostak, J.W., and Blackburn, E.H. (1984) DNA sequences of telomeres maintained in yeast. ***Nature*** 310: 154-157.
1984 MAIO	Dunn, B.L., Szauter, P., Pardue, M-L., Szostak, J.W. (1984) Transfer of telomere-adjacent sequences to linear plasmids by recombination. ***Cell*** 39: 191-201.
1985	Greider, C.W. and Blackburn, E.H. (1985) Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. ***Cell*** 43: 405-413.
1987	Greider, C.W. and Blackburn, E.H. (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. ***Cell*** 51: 887-898.
1988 NOV	Greider, C.W. and Blackburn, E.H (1987) A telomeric sequence in the RNA of Tetrahumena telomerase required for telomere repeat synthesis. ***Nature*** 337: 331-337.
1989 JAN	Lundblad V. and Szostak, J.W. (1989) A mutant with a defect in telomere maintenance leads to senescence in yeast. ***Cell*** 57: 633-643.
1990	Yu, G.L, Bradley, J.D., Attardi, L.D. and Blackburn, E.H. (1990) In vivo alteration of telomere sequences and senescence caused by mutated Tetrahymena telomerase RNAs. ***Nature*** 344: 126-132.
1992	Allsopp, R.C., Vaziri, H., Patterson, C., Goldstein, S., Younglai, E.V., Futcher, C.W., Greider, C.W., and Harley, C.B. (1992) Telomere length predicts the replicative capacity of human fibroblasts. ***Proc. Natl. Acad. Sci. USA*** 89: 10114-10118.
1993	Prowse, K.R., Avilion, A.A., and Greider, C.W. (1993) Identification of a nonprocessive telomerase activity from mouse cells. ***Proc. Natl. Acad. Sci. USA*** 90: 1493-1497.
1995	McEachern, M.J. and Blackburn, E.H. (1995) Runaway telomere elongation cause by telomerase RNA mutations. ***Nature*** 376: 403-409.
1999	Rudolph, K.L., Chang, S, Lee, H.W., Blasco, M., Gottlieb, G., Greider, C.W., and DePinho, R.A. (1999) Longevity, stress response, and cancer in aging telomerase deficient mice. ***Cell*** 96: 701-716
2001	Kim, M.M., Rivera, M.A., Botchkina, I.L, Shalaby, R., Thor, A.D., and Blackburn, E.H. (2001) A low threshold level of expression of mutant-template telomerase RNA is sufficient to inhibit tumor cell growth. ***Proc. Natl. Acad. Sci. USA*** 98: 7982-7987

About the authors

Carlos Henrique Marcondes

http://www.professores.uff.br/marcondes

Professor at the Department of Information Science at UFF (Fluminense Federal University), where he acts in the undergraduate courses Archivology and Biblio-economics, professor and current coordinator of the PPGCI/UFF-Master's – Post-graduate program in Information Science – professor of the Scientific Information and Health Technology Specialization Course, at the Icict/Fiocruz, researcher at CNPq, developing research in modeling of scientific articles as ontologies, adhoc consultant for Capes and CNPq, member of the editorial council and referee of various scientific periodical, member of the Technical Chamber of Digital Documents of the Conarq/National Archive, acts in the information technology area applied to the treatment of information, is author of various scientific articles in this area; acted as consultant in projects such as the Digital Library of Theses and Dissertations – BDTD – of the Ibict and in the development of the SciELO/Open Archives server.

Marília Alvarenga Rocha Mendonça

Graduated in Biblio-economics from the School of Biblio-economics of the Federal University of Minas Gerais (1996), specialist in Planning, Organization and Direction of Archives (1985) and Organization and Administration of University Libraries (1986) by the Fluminense Federal University, Master in Business Administration from Fluminense Federal University (2001). Acted as librarian in the management of the Education Department Library/UFMG (1972/79), in the implementation and management of the Microfilming Center/UFMG (1979/82), in the Documentation Nucleus/UFF as General Archive manager and in the Pharmaceutical Department Library (1982/94). Is currently assistant teacher at the Department of Information Science of Fluminense Federal University. Researcher of the Research Group "Information, Knowledge and Information Technology", having as a line of research "management of digital contents and information technologies". Participates, since 2003, in the research projects CNPq/UFF, with results that have been presented in national and international events and published in annals and in the form of scientific articles in national and international periodicals. Has published a book and three book chapters.