Identifying Traces of Scientific Discoveries by Comparing the Content of Articles in Biomedical Sciences with Web Ontologies

Carlos Henrique Marcondes[1] and Luciana Reis Malheiros[2]

[1] *marcon@vm.uff.br*
Federal Fluminense University, Department of Information Science, R. Lara Vilela, 126 – São Domingos, CEP 24210-590, Niterói – Rio de Janeiro (Brazil)

[2] *malheiro@vm.uff.br*
Federal Fluminense University, Department of Physiology and Pharmacology, R. Prof. Hernani Melo, 101 – Centro, CEP 24210-130, Niterói – Rio de Janeiro (Brazil)

Abstract
This paper reports a methodological proposal consisting of comparing the content of scientific articles represent in machine-processable format as triads Phenomenon-Semantic_Relation-Phenomenom with the content of Web public ontologies in order to identify traces of scientific discoveries reported by the article. Articles which content is poorly represented in those ontologies are strong candidates to report discoveries. The methodological proposal described will be the basis to the future development of an automatic procedure.

Introduction
Even today, with the advances of information technology (IT) scientific communication is a slow social process which largely depends on discourse, text producing and reading/interpreting/inquiring/citing these texts by scholars until new knowledge is incorporated to the corpus of Science. Howeversome papers reporting important scientific discoveries stay uncitted for many years as "sleeping beuties" in science (Van Haan, 2004). Through scientific journals new knowledge, results and benefits of scientific activity have been systematically incorporated by society. Before the raise of the Web, what constitutes the assented humanity scientific knowledge was fuzzy, lacks formalization and was scattered across journals collections throughout libraries.
Ontologies are one of the foundations of the Semantic Web (Berners-Lee, 2001) and have been used to formally record scientific knowledge in specific domains. De Roure (2001) stresses the importance of knowledge integration from different sources, including scientific articles Web published, to e-Science environments. To meet this requirement knowledge must be represented in machine-processable format.
Nowadays, electronic Web publishing is a common activity to scholars. Most scientific journals are now available through the Web. But IT is not yet used to directly process the knowledge embedded in the text of scientific articles. Electronic published articles are knowledge bases, but for human reading. There are two barriers to a large scale use of this knowledge: the amount of information available throughout the Web and the fact that knowledge is in textual format, in an unstructured way, not adequate for program processing. Today electronic journals are still based on paper print mode.
Scientific journal articles add something to the human stock of knowledge. How scientific discoveries can be identified? Can these features be identified in the text of scientific articles, especially in a Semantic Web publishing environment?
A criticism of bibliometric and scientometric approaches is that "*they do not take into account the semantic content of scientific publications*" (Niiniluoto, 2002). The indexing of articles is not done by the authors who know best what is being reported and its contribution to science but later, when articles are included in databases or repositories as Medline and PubMed.
This research is pursuing a new paradigm in electronic publishing. Articles, besides being published in textual format, have also their content identified, extracted, recorded and

Proceedings of ISSI 2009, pages 173–177.
Edited by B. Larsen and J. Leta.

published as an ontology instance in machine-processable format as a by-product of the process of self-publishing and self-describing when submitting articles to an electronic journal system. The model has two components: a knowledge representation model for recording the knowledge content of articles in a machine-processable format and a Web authoring/publishing system to be used in an electronic journal, repository or digital library.
Miller (1947) states that: "*The above remarks imply that science is a search after internal relations between phenomena*". Our approach to knowledge representation of the content of scientific articles is based on the fact that scientific knowledge consists in claims made by scientists in article texts expressing relations between phenomena or between a phenomenon and its characteristics. Relations are the basic units of scientific knowledge and synthesize articles content. Claims are extracted, marked up as relations and recorded in machine-processable format. This enables their processing by software agents thus providing scientists with new means to retrieve, compare and reason on this knowledge.
We have developed an ontology aimed at representing the knowledge embedded in scientific article text (Marcondes, 2009). Instances of this ontology corresponding to each article content are generated by the authoring/publishing software system as a by-product of the process of submitting articles by authors.
Besides knowledge management and retrieval another function to the proposed knowledge representation format could be the identification of traces of new discoveries. Once represented in machine-processable format this knowledge could also be compared by programs with the knowledge held in Web public ontologies thus revealing inconsistencies, faults and even new discoveries. Are the claims made by an author represented by concepts in a Web ontology in the same scientific domain? Is it possible that a scientific article, at the moment of its publishing in an electronic journal system and without even being refereed or cited, reveal formal and/or content traces that may indicate it reports a scientific discovery?
We hypothesize that there is a correlation between articles which content is poorly represented or represented just in a generic level in terminological data banks as UMLS – the Unified Medical Language System - and the fact that these articles report scientific discoveries. The aim of this paper is to demonstrate the feasibility of an automatic methodology of comparing the knowledge content of scientific articles represented in machine-processable format with Web ontologies as the basis to identify traces of discoveries.

Representing the content of scientific articles in machine-processable format

We propose a model which extends conventional bibliographic record models comprising elements as authors, title, source, publication date information and others descriptive elements together with article's content information as keywords or descriptors. The model adds to these elements the claims made by authors in their papers are represented as relations between two different phenomena or between a phenomenon and its characteristics. A relation has the form of an Antecedent (a concept referring to a phenomenon), a Semantic Relation and a Consequent (another concept referring to a phenomenon or a characteristic of the phenomenon in the Antecedent). A Semantic Relation may be a specific Type_of_relation as "causes", "affects", "indicactes", or a (has_as_)Characteristic relation. Examples of knowledge representation according to this schema are the following:

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

A complete description of the model can be found in Marcondes (2009).
We are giving the first steps towards the development of electronic journal system through which authors can submit/publish their articles. This system will develop an interactive dialog

with authors, making them questions, processing their answers and the text of their articles in order to extract and markup knowledge content of the article according to the model proposed and record it in machine-processable format as ontology instances. The system will also allow and assist authors to browse through public Web ontologies like UMLS and to annotate the relations representing the article knowledge content to these ontologies.

UMLS Semantic Network organizes UMLS MetaThesaurus terms in hierarchies of terms – semantic types. Other component of UMLS SN are relations encompassing formal semantic rules to relate 2 MetaThesaurus terms. Each UMLS hierarchy has as its top a semantic type. There are 134 semantic types and 53 possible relations linking two semantic types in UMLS SN. UMLS SN stores not only each of the 53 relations but also the rules governing which semantic type can be related by each of them.

The annotation process is a key step in identifying discoveries. It aims to identify if concepts in Antecedent, in Consequent or the Type_of_Relation/Characteristic do exists as terms in UMLS vocabulary. Once extracted relations and annotations are both coded in machine-processable format using OWL and recorded as ontology instances, comprising semantic richer bibliographic surrogates. They can be further processing by software agents to reason on the content of scientific articles and on annotations. As knowledge content of articles according to the model proposed and also many public ontologies on the Web are both coded in OWL, they can be automatically processed and compared.

Subsequent processing by a program of the annotated knowledge representation surrogate of each article can thus indicate the grade of mapping achieved in each article (full mapping, mapping just the Antecedent and the Consequent, mapping just one relata and the Type_of_Relation/Characteristic, mapping just the Type_of_Relation/Characteristic, mapping of neither of these elements). It can indicate also, in case of some mapping, if it is valid according the UMLS semantic rules. Not mapping at all or a low grade of mapping or incorrect mapping according UMLS semantic rules indicates that phenomenon represented by terms in article knowledge representation are new, not yet recognized, not yet incorporated as entries and rules to terminological data banks like ULMS. This fact can indicate that the corresponding article may report a scientific discovery.

Material and Methods

We manually analyze 75 Biomedical articles both to develop the model previously described and to work up the hypothesis that articles which content is poorly represented or represented just in a generic level in terminological data banks as UMLS may report scientific discoveries. Articles analyzed comprise 3 groups.

- articles from two outstanding Brazilian research journals, 20 articles from the Memórias do Instituto Oswaldo Cruz and 20 articles from the Brazilian Journal of Medical and Biological Research.
- 20 articles about stem cells were also analyzed. Stem cells, as an emerging research area in rapid development, were chosen expecting to find articles reporting important discoveries.
- 15 articles from the Albert Lasker Basic Medical Research Award 2006 key publications were analyzed. This last group is of special interest to the objectives of this research because the articles report, step by step, the rise of new scientific discovery, the discovery of telomerase enzyme since 1978 - the first article - to 2001 - the last article of this group.

The analysis process is developed in 2 steps and consists in identifying the main relations posit by the author in the text. The analysis procedure *simulates* the results to be obtained by the journal system in its dialog with authors in order to extract a knowledge content representation of each article. Here follows an example:

CÂMARA, G.N.L. 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,* 98(7), Oct. 2003.

- Step 1 –author claims are identified in the text:

'HPV causes pre-neoplastic and neoplastic cervical lesions'

Knowledge as a relation: Antecedent: HPV,
Type_of_relation: causes,
Consequent: pre-neoplastic and neoplastic cervical lesions.

- Step 2 - Each of these elements is tentatively mapped to concepts in UMLS/UMLS Semantic Network. This mapping is achieved by comparing terms in article's knowledge content representation extract in step 1 to PubMed records of each article, which includes MeSH/UMLS terms indexing the article:

'Papillomavirus, Human'
'Causes', UMLS Semantic Network Relation T147,
'Colonic Neoplasias'

In this example, all concepts in Antecedent, in the Type_of_relation and in the Consequent were successfully mapped to UMLS concepts.

Results

Among the 75 articles analyzed the groups of articles which reports discoveries – The Lasker Awards 2006 group of articles, followed by the Stem Cells group of articles - obtained the worse mapping rates to UMLS concepts. This is group of articles which in dead reports steps toward an important scientific discovery. This group presents the lowest rate of mapping as showed in the following Table. In this group 100% of the articles did not map at least one element of the knowledge representation format. Inside this group the partially mapped articles (6 in 15), achieved mapping of just the Type_of_relation to Relations in UMLS SN; none achieved full mapping.

Among the Stem Cells group of articles, 80% did not map at least to one element of the knowledge representation format; 16 in 20 partially mapped and just in 5 the Type_of_relation maps to UMLS SN Relations. This result, when added to the not mapping at all articles sums up 45%. Relations in UMLS SN are few (just 53 when compared with 1 million biomedical concepts and 5 million concept names in UMLS Metathesaurus), so more generic and more stable throughout time in comparison to concepts in a scientific area.

The two groups of articles reporting scientific discoveries have low rates of mapping. Any article in both groups presented fully mapped content representations.

Table I. Results of the mapping of concepts to UMLS per group of articles

Articles analyzed	MIOC	BJMBR	STEM CELLS	TELOMERASE	TOTAL
Fully mapped	7 (35%)	3 (15%)	0 (0%)	0 (0%)	10
Partially mapped	13 (65%)	11 (55%)	16 (80%)	6 (40%)	44
Not mapped	0 (0%)	6 (30%)	4 (20%)	9 (60%)	21
Total of articles	20	20	20	15	75

Discussion

These results indicate that the grade of successful/unsuccessful mapping achieved by article content representation to UMLS concepts may be associated to the fact that the article reports scientific discoveries. It seems methodologically feasible to propose a procedure that automatically process and compare annotated knowledge representations of articles as previously proposed with the terms found in an ontology. This procedure can attract attention

of scientists to traces of scientific discoveries. In the sample analyzed articles which presents low rates of mapping or mapping just the Type_of_Relation/Characteristic or no mapping in dead all report discoveries. The method here proposed is intended to be complementary to bibliographic and scientometric methods.

More research must be done to verify the feasibility of the methodology described including ontologies of a similar scope of the analyzed article corpora and a whole e-science environment. The processes and methods of ontology curation today are also new and still lack social endorsement and validation. A scientific discovery creates new concepts to which terms are not yet coined in terminological data banks like ULMS. There is a delay between the discovery of new phenomenon/concepts and the update of ontologies like UMLS with the terms representing these concepts. Interestingly telomerase enzyme was first reported in 1985 (Greider, 1985) while Mesh – Medical Subject Headings – entry for telomerase was just created at 1995/06/05[*].

With the raise of ontologies as new scientific artifacts (Smith, 2008) we are facing new processes of scientific validation/ratification which are specific (see e.g. OBI: Structures and Regulations, 2008). Ontologies are also evolving towards more formal devices and deserve new methods of curation (Williams, 2003).

The same can be said of scientific articles when published in digital format: as soon as they are published in a richer and formal content format, this enable the processing of these content and their comparison to public Web ontologies as proposed here.

When fully implemented and used in large scale, the publishing model here proposed can provide scientists with new tools for knowledge management, validation and identification of scientific discoveries as part of a new e-science environment.

References

Albert Lasker *Basic Medical Research Award 2006* (2008). Retrieved June 11 2008 from: http://www.laskerfoundation.org/awards/2006_b_keypub_blackburn.htm.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The semantic web. *Scientific American*.

Blackburn, E. H, Greider, C. W., Szostak, J. (2006). Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nature* 12(10), 1133-1138.

De Roure, D, Jennings, N., Shadbolt, N. (2001). *Research agenda for the Semantic Grid: a future e-Science infrastructure*, Report commissioned for EPSRC/DTI Core e-Science Programme (2001).

Greider, C.W., Blackburn, E.H. (1985) Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell*, 43, 405-413.

Miller, D. L. (1947). Explanation Versus Description. *Philosophical Review* 56(3) 306-312.

Marcondes, C. H; Mendonça, M. A. R.; Malheiros, L. R; Costa, L. C. da; Santos, T. C. P. (2009). Ontological and conceptual bases for a scientific knowledge model in biomedical articles. *RECIIS*, 3(1) 19-30.

Niiniluoto, I. (2002). Scientific progress. In: *Stanford Encyclopedia of Philosophy*. Retrieved December, 18, 2004 from http://plato.stanford.edu/entries/scientific-progress/.

OBI: Structures and Regulations. Retrieved December, 2, 2008 from http://sourceforge.net/project/downloading.php?groupname=obi&filename=OBI-GroupsRegs-0.3.doc&use_mirror=ufpr.

Smith, Barry. (2008). Ontology (science). *Nature Precedings* : hdl:10101/npre.2008.2027.2. Retrieve November 23, 2008 from http://precedings.nature.com/documents/2027/version/2/files/npre20082027-2.pdf.

Van Haan, Anthony F. Sleeping beauties in science. (2004). *Scientometrics*, 59(3), 467-472.

Williams, Jennifer; Anderson, William. Bringing ontology to the Gene Ontology. (2003). *Comparative and Functional Genomics* 4, 90–93.

[*] See information in http://www.nlm.nih.gov/cgi/mesh/2009/MB_cgi , accessed in Dec. 17, 2008.