A Model to Represent and Process Scientific Knowledge in Biomedical Articles with Semantic Web Technologies†

Carlos H. Marcondes* and Leonardo C. da Costa**

Post-graduate Program of Information Science, Federal Fluminense University,
R. Tiradentes, 148, Ingá, CEP 24210-510, Niterói, RJ, Brazil
*<marcon@vm.uff.br>, **<leo.cruz@yahoo.com.br>

Carlos Henrique Marcondes is Professor in the Department of Information Science and in the Information Science Postgraduate Program at Federal Fluminense University, Rio de Janeiro, Brazil. He is an associate researcher of CNPq, The Brazilian Council for Scientific Development. His research interests are conceptual modeling, scientific communication, theory of classification and ontological analysis.

Leonardo Cruz da Costa is Professor in the Department of Computer Science and in the Information Science Postgraduate Program at Federal Fluminense University, Rio de Janeiro, Brazil. His research interests are conceptual modeling, scientific communication, text mining and semantic annotation.

Marcondes, Carlos H. and Leonardo C. da Costa. 2016. "A Model to Represent and Process Scientific Knowledge in Biomedical Articles with Semantic Web Technologies." *Knowledge Organization* 43 no. 2: 86-101. 60 references.

Abstract: Knowledge organization faces the challenge of managing the amount of knowledge available on the Web. Published literature in biomedical sciences is a huge source of knowledge, which can only efficiently be managed through automatic methods. The conventional channel for reporting scientific results is Web electronic publishing. Despite its advances, scientific articles are still published in print formats such as portable document format (PDF). Semantic Web and Linked Data technologies provides new opportunities for communicating, sharing, and integrating scientific knowledge that can overcome the limitations of the current print format. Here is proposed a semantic model of scholarly electronic articles in biomedical sciences that can overcome the limitations of traditional flat records formats. Scientific knowledge consists of claims made throughout article texts, especially when semantic elements such as questions, hypotheses and conclusions are stated. These elements, although having different roles, express relationships between phenomena. Once such knowledge units are extracted and represented with technologies such as RDF (Resource Description Framework) and linked data, they may be integrated in reasoning chains. Thereby, the results of scientific research can be published and shared in structured formats, enabling crawling by software agents, semantic retrieval, knowledge reuse, validation of scientific results, and identification of traces of scientific discoveries.

Received: 5 August 2015 Revised: 21 December 2015; Accepted 29 December 2015

Keywords: scientific knowledge, models, knowledge units, biomedical research articles, semantic web, linked open data, RDF, PDF

† This article is an extended version of the paper presented in the Workshop on Knowledge Organization and Semantic Web, German ISKO e.V., SEMANTiCS, Leipzig, Germany, 1st Sept. 2014. Our thanks to the Brazilian grant agencies CNPq and CAPES.

1.0 Introduction

Since the rise of the first scientific journal—*The Proceedings of the Royal Society*—in the seventeenth century, scientific articles have become privileged channels of scientific communication. The content of scientific articles is submitted to critical reading, inquiry, and citation through a long social process until it becomes public knowledge. The current scholarly Web publishing environment is still an electronic metaphor of the print publishing system used throughout the twentieth century and is still based on linear text formats such as portable document format (PDF). The textual format of articles suitable for human reading limits the possibilities for automatic processing of their contents. Tasks such as knowledge reuse, discovery, gap analysis, contradictions and agreements in knowledge, and validation of scientific results, all demand human effort. The availability of automatic tools to assist such tasks

is increasingly important as the number of scientific articles published in digital formats increases and scientists in their daily work have to process results from different articles and sources. For example, the PubMed repository currently holds over 23 million articles. Scientific articles are also spread across various information resources such as digital libraries, electronic journal systems, and repositories. According to Renear and Palmer (2009), scientists are increasingly using what they call strategic reading to cope with the amount of literature being published. Research tasks demand new tools for information discovery, retrieval, and content comparison in very specific, precise, and meaningful ways.

Although modern bibliographic information retrieval (IR) systems exploit the potential of information technology, they are not yet used to directly process the knowledge embedded in the text of scientific articles. Since the MARC (MAchine-Readable Cataloging) record was established in 1960, bibliographic record models have changed little. A typical bibliographic record is comprised of a flat set of unconnected database fields for content description, keywords or descriptors, or other attributes such as author, journal title, publication date etc., each having an equal weight for retrieval purposes. Content access to documents in modern bibliographic information retrieval systems is still achieved by matching user queries comprised of keywords connected by Boolean operators to keywords or attribute values presented in the bibliographic records, a technology similar to early bibliographic retrieval and library automation systems of the 1960s and 1970s. The conventional bibliographic records do not show explicit relations between elements comprising the content of documents they represent. This state of affairs did not change even with the rise of formats such as Dublin Core, developed to manage electronic scientific publications.

Scientific knowledge aims at universality and necessity. Such characteristics make this kind of knowledge susceptible to heavy reuse. Miller (1947, 310) states that "science is a search after internal relations between phenomena." Scientific knowledge, as it appears in the text of scientific articles, consists of claims made by authors throughout the article text, synthesized in the article's conclusions. These claims comprise the knowledge in scientific articles. They are highly reliable knowledge units, as they are validated by the peer-review process and are the result of an experiment described and tested in the article.

Compared to the poor expressiveness of the three Boolean operators, AND, OR, and NOT, the Unified Medical Language System (UMLS) Semantic Network (SN), the classification schema of the UMLS National Institutes of Health Metathesaurus, organizes every concept in hierarchy trees, each having as its root a top level semantic type. The UMLS SN uses fifty-four relation types to express the semantic relations between concepts in semantic type hierarchies. The UMLS SN determines the allowed relations between semantic types. Although this semantically rich schema of representing the content of articles by relations is supported by the UMLS, the bibliographic record models in databases such as Medline are not able to explore its full potential. For example, two of the UMLS semantic types are "Pharmacologic Substance" (UMLS unique identifier T121) and "Pathologic Function" (UI T046). Relationships between these two concepts can be quite different, a pharmacologic substance can "cause" (UI T147) a pathologic function, or a pharmacologic substance can "prevent" (UI T148) a pathologic function. These two different relationships can only be expressed in one way in Boolean algebra as "Pharmacologic Substance AND Pathologic Function."

Many different relationships involving scientific articles can be identified: bibliographic or citation relations; relations with datasets holding raw results of scientific experiments or databases such as GenBank, DrugBank, ArrayExpress, PhenomicDB; internal relations between parts of an article—semantic components—such as a problem, question, hypothesis, methodology, results or conclusion; relations with terminological knowledge bases or ontologies such as UMLS or GO; relations with grant agencies that support the research; relations with claims within an article and across articles; relations within two different bibliographic sets (literature-related discovery methodologies) (Swanson et al., 2006); and relations with annotations or comments made about an article. Now the semantic web and linked open data environments provide means of making explicit all these relationships.

The current citation-based information retrieval systems and the print model of publication constitute closed systems where scientific articles are isolated from the mainstream Web and thus are barriers to data reuse, sharing, integration, and synthesis. Within such environments, those relations are implicit, informal, and are not coded in machine-processable formats. Scientific publications were first recorded in bibliographic databases. Based on these databases, citation models and citation networks were developed as tools to understand and manage the development of science (Garfield et al., 1964). This situation can now be overcome within the scope of the semantic web linked open data platform. These technologies offer the possibility of developing a richer and more multifaceted scientific knowledge environment where navigating throughout a citation network will be only one of the many possibilities.

The semantic web (Berners-Lee et al., 2001) and linked open data technologies (Bizer et al., 2009) constitute a step forward from conventional information retrieval en-

vironments. The content of a Web document is no longer a matter of a simple keyword match, as in conventional computational environments since the 1960s, but instead comprises structured sets of concepts connected by precise meaning relations expressed in RDF (Resource Description Framework) and RDF schema directly published and available through the Web. Such a rich knowledge representation schema enables software agents to perform "inferences" and more sophisticated tasks based on article content.

These technologies provide new opportunities for communicating, sharing, reusing, interlinking, and integrating scientific knowledge published in digital formats that may overcome the limits of the current print format used for the publication of scientific results, which is only suitable for reading and processing by people. We are now beginning to use these technologies for sophisticated tasks such as knowledge discovery, knowledge comparison, and integration of multiple sources, which facilitates inference capabilities among different and autonomous information resources.

The aim of this article is to propose a semantic model of scientific articles that goes beyond conventional metadata formats such as MARC and Dublin Core. The proposed model enables the management of article content not just for retrieval purposes by humans, but also for automatic reasoning. That means not only enhanced possibilities for semantic retrieval by human users, but also the enabling of the content of scientific articles to be integrated in inference chains by computer programs and to be queried as a database. The model also addresses and outlines a wholly new scientific publishing environment, one that exploits the possibilities created by semantic web and linked open data technologies (Shadbolt et al., 2006). The model aims also at providing guidelines for the development of technological solutions that can, partially or completely, implement the model's components. The article is organized as follows: following the introduction, section 2 discusses the theoretical foundations of the model and related works; section 3 presents and explains the model; section 4 discusses model implementations so far and the model's potential to enhance information retrieval, knowledge discovery, identification of discoveries in science; and section 5 presents concluding remarks and perspectives of research development.

2.0 Materials and methods

The insight to develop such a model came from literature on philosophy of science and scientific methodology. Literature on rhetoric of scientific papers and different reasoning patterns found in them (Bezerman 1988; Gross 1990; Hutchins 1977; Nwogu 1997; Skelton 1994) also were inputs to the model. The different types of scientific articles according to their reasoning patterns found in this literature were classified and incorporated in the model. The model was tested and adjusted by analysing eighty-nine articles in biomedical sciences with the aim of identifying the semantic components of scientific methodology, how they appeared in the texts of the articles, and reasoning patterns and sequences that combine these elements. Details of this empirical analysis are related in Marcondes et al. (2014). The model is graphically presented as a UML (Unified Modeling Language) class diagram, developed using the NClass software tool. Semantic records in RDF were validated using W3C Validator services (http://www.w3.org/RDF/Validator/); screens with graphs generated from RDF records were captured from the same service.

3.0 Bases of the model

In this section, the theoretical and empirical bases used for developing the model are presented and discussed. To achieve these goals, first we discuss how to identify a knowledge unit in the text of scientific articles; after that we discuss the question of how to model a context where such knowledge units can be safely used in valid reasoning chains.

What are the methods for achieving truth in science? In the modern age, an important contribution to this discussion was the proposal of the scientific method, postulated by Francis Bacon (1973) (among others). In opposition to medieval scholastics, Bacon emphasized the importance of observational experiments to set general laws in science. His reasoning method of deriving general statements from a particular number of observational cases was called induction.

Today's version of scientific method is called the hypothetico-deductive method, largely used in experimental sciences such as the biomedical sciences. The method consists of giving a problem, proposing a feasible hypothesis, deducing consequences of the hypothesis and developing experiments to test it. The results of the experiments confirm or reject the hypothesis.

An hypothesis is an essential component of the hypothetico-deductive method. An hypothesis expresses a contingent relation between phenomena (Marconi and Lakatos 2004, 137). Research in library and information science, especially in domains such as indexing languages, coordinated indexing systems, and information retrieval, gives special attention to relationships as keys for representing meaning. Farradane's (1980) relational indexing approach proposed "meaning, considered as relations between terms." According to Brookes (1980), "knowledge is a structure of concepts linked by their relations and in-

formation is a small part of such a structure." Sheth et al. (2003) state, "relationships are fundamental to semantics—to associate meaning to words, items and entities. They are a key to new insights. Knowledge discovery is about discovery of new relationships."

Although a complex phenomenon, scientific reasoning, as recommended by the scientific method and expressed in the texts of scientific articles, plays an essential role in communication science, i.e. the validation of the knowledge contained in any article, thus enabling a scientist to reproduce the steps taken by the author in any experiment. The need for this rigid protocol when communicating research results is stated by The International Committee of Medical Journals Editors (2003):

> 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.

In addition to the IMRAD structure, in 1987 the Ad Hoc Working Group for Critical Appraisal of Medical Literature recommended the use of structured abstracts to improve the information of clinical articles. Since then an increasing number of biomedical journal editors have adopted the structure for abstracts. Structured abstracts are one of the roots of the model proposed here. Their adoption can improve scientific communication and reliability of the results reported in each paper. In the Structured Abstract Labels Research Dataset, a large number of structured abstract labels used in different biomedical journals can be found. Their variety stresses the importance of making explicit the decisions made in the process of scientific inquiry to improve scientific communication and reliability of the results reported in the paper (Salager-Meyer 1991). The National Library of Medicine, too, adopted a five label categories standard for structured abstracts: background, objective, methods, results, and conclusion. Structured abstracts are a step towards a semantic format for scientific papers. A further step in this direction would be to represent these semantic components in a machine-readable format so they can be processed by programs.

In the Structured Abstract Labels Research Dataset previously mentioned, there are some suggested labels used in structured abstracts that are found in a variety of medical journals; e.g., such as "problem addressed," "basic problem and aim of study," "basic problems and objectives," "questions of the study," "questions under study," "clinical questions," "hypothesis," "hypothesis tested," "clinical questions,'" etc. All of these labels are related to fundamental issues in scientific inquiry.

In the model proposed, scientific knowledge units consist of the establishment of new relationships between phenomena. A phenomenon can be defined as a "perceptible fact, a sensible occurrence" (Bunge 1998, 173). Relationships in their simplest form could be modeled as triples of <Antecedent> <Type_of_relation> <Consequent>. In a scientific article relationships may appear in different semantic components throughout the article text, such as within the problem that the article addresses as a contingent relationship; as a question, in which either one of the two relata or the type of relation is unknown; in the hypothesis as a hypothetical relationship; or in the conclusion as an empirically tested relationship. Frequently, the conclusion also poses new questions. "Questions," "Hypothesis," and "Conclusion" are the semantic components of the proposed model that synthesize these issues. Such elements, according to the scientific method, are also related in a reasoning chain. A question evokes a tentative hypothesis of how to solve it; the hypothesis must be tested by a practical experiment which confirms or denies it; the results of the experiment are synthesized in a conclusion. These semantic components converge to justify, support and guarantee the article's conclusion.

The conclusion is the essential semantic component, as it synthesizes the knowledge content of an article and its contribution to a scientific domain. According to Samwald (2009) "The conclusion sections of biomedical abstracts seem like a gold-mine for automated key assertion identification, since the relevant portion of text can be identified easily." However, in the scope of a recently-published article, the conclusion is a provisional knowledge unit, as it is at least validated by the experiment reported in the article and by the peer-review process by which the article was accepted for publication. Semantic components such as questions and hypotheses are important as they permit the evolution of a claim to be traced. Other elements have rhetorical functions, as extensively discussed in Skelton (1994) and Nwogu (1997), or serve to describe methodological options, the experiment performed, its context, or display more clearly the results obtained. Other elements include the proposed objectives of the article and the description of the experiment performed.

4.0 Proposal

The idea of a bibliographic record model is the result of a long tradition that harkens back to Panizzi, Jewett, Cutter, and Lubetzky (Bianchini and Guerrini 2009). Although these pioneers posed the differences between a work as an abstract creation and the book that encompasses it, tech-

nological limitations of the earlier bibliographic systems imposed a simplified representation restricted to a brief description and the main access points. This representation encompasses different independent entities that were mixed and simplified as mere record fields. This scenario began to change in 1998 with the publication of IFLA's Functional Requirements for Bibliographic Records (1998).

Functional Requirements for Bibliographic Records, better known as FRBR (IFLA 1998), is a conceptual model, based in Chen's (1976) entity-relationship (ER) model. An ER model comprises the entities "a 'thing' which can be distinctly identified," the relationships "an association among entities," and the attributes of entities or relationships in a specific domain. "The information about an entity or a relationship is obtained by observation or measurement, and is expressed by a set of attribute-value pairs" (Chen 1976, 12). In the bibliographic domain entities could be a work such as *Hamlet*, and its author, William Shakespeare, a relationship would be the authorship of Hamlet by William Shakespeare, an atribute of this work entity would be its title, attributes of the author entity might be his name and birth and death dates, respectively 1564 and 1616. Also, FRBR identifies the relations between the work and its various expressions such as the text of a play or a film; their various manifestations, such as the 1994 English edition by Penguin Books or the 2007 Portuguese edition by L&PM Editora; and the several items available in a library's holdings.

The model presented here was first proposed in 2005 and has been continuously improved since then (Marcondes 2005). It addresses a related but a slightly different issue from that addressed by the FRBR model. It extends conventional bibliographic record models comprised of descriptive elements such as authors, title, abstract, bibliographic source, publication date, and content information such as keywords or descriptors and references to cited papers. In addition to these bibliographic elements, the model makes explicit and formalizes the claims made by authors throughout the article text. These claims are the basic knowledge units that comprise the scientific contributions of an article. They are not explicitly represented nor coded in conventional bibliographic records and are hard to find in article texts. The model proposes the coding of those claims in a machine-processable format, enabling their use in tasks that demand intelligent processing. Once explicit and coded, they may be integrated into reasoning chains, enabling semantic retrieval, knowledge reuse, validation of scientific results, tracing of scientific discoveries, new scientific insights, and identification of knowledge contradictions or inconsistencies. Such a model is designed to be implemented in the semantic web linked open data platform.

Since the 1970s, many initiatives have been undertaken to organize and standardize knowledge in biomedical sciences. Initially those initiatives took into account the terminological knowledge necessary to cope with indexing biomedical literature. Recently, many models of scientific knowledge in biomedical sciences have been proposed to address biomedical knowledge directly (OBO Foundry). To address the aims of these models, or ontologies, built on the principles of OBO Foundry, models are created with the epistemic approach of scientific realism, which means that all instances of entities represented in such ontologies must have real existence in time and space (Cleusters and Smith 2006).

The model proposed makes quite a different assumption. It is committed to modeling and discovering the knowledge units that comprise scientific reasoning. According to Shultz and Jassen (2013, 8), domain models in biomedical sciences may include four kinds of statements: universal statements, terminological statements, statements about particulars, and contingent statements. We do not intend to model universal statements in biomedical sciences. Accordingly, the model includes contingent or hypothetical knowledge units and knowledge units that are not yet largely accepted in a scientific domain, such as the conclusions of an article. Indeed these semantic components that comprise scientific reasoning are the core of the model.

Scientific claims in articles take the form of relations between phenomena or between a phenomenon and its characteristics. They are expressed linguistically through propositions (Dahlberg 1995, 10):

a- "telomere shortening (Phenomenon) causes (Type _of_relation) cellular senescence (Phenomenon)" (Hao et al. 2005);
b- "telomere replication (Phenomenon) involves (Type_of_relation) nontemplate addition of telomeric repeats onto the ends of chromosomes (Phenomenon)?" (Shampay et al. 1984); or
c- "tetrahymena extracts (Phenomenon) show (Type _of_relation) a specific telomere tranferase activity (Characteristic)" (Greider and Blackburn 1985).

In formal ontology and knowledge representation literature, both characteristics and relations are called properties. We opt to differentiate characteristics (also called attributes) as properties, which depend only on one entity instance, and relations, as properties that depend on two or more entity instances. This decision makes our model compatible with OWL (Web Ontology Language), which has properties as one of its basic building blocks, although it distinguishes data type properties (attributes, characteris-

tics) from object properties (relations). Furthermore, this decision has to do with the proposed classification of patterns of reasoning in scientific articles. One type of article, experimental-exploratory articles (EE), is characterized as describing a new phenomenon by proposing relations, not between two different phenomena, but between a phenomenon and its characteristics, as are shown in the following paragraphs.

As explained in the previous section, claims such as basic knowledge units represented in the model, may appear in different semantic elements that comprise the model, such as in a question within the problem, in the hypothesis and in the conclusion. Semantic components such as questions and hypotheses are important because they enable the tracing of a claim over time. However, the conclusion is the core semantic component of the proposed knowledge representation schema. There are other elements in the model which are not represented as relations, such as the objective and the experiment. The semantic components that comprise the proposed model are described below.

- The semantic article is a complex digital object with conventional bibliographic metadata and with links to its full-text and to other articles that are cited or that cite one specific article. Its components are the following:
  - the problem the article is addressing and the question derived from it;
  - an original or new hypothesis, aimed at addressing the question proposed;
  - sometimes authors develop experiments in order to corroborate or negate a previous hypothesis, one which was proposed by some other author;
  - an empirical experiment is also described with the aim of observing the phenomenon described it relations with other phenomena and specific characteristics which is comprised of:
  - results: tables, figures, and numeric data reporting the observations made;
  - measurement used;
  - the specific context where the empirical observations take place with the following components:
    - environment, e.g., a hospital, a daycare center, a high school;
    - geographical location where the empirical observations occur;
    - time when the empirical observations occur;
    - specific population in which the phenomenon occurs, e.g., pregnant women, early born babies, mice;
  - conclusions: a set of propositions made by authors as a result of the findings. A conclusion corroborates, totally or partially, the hypothesis of an article or negates it. A conclusion may also be conclusive or not yet conclusive;

- Semantic components such as question, hypothesis, previous hypothesis, and conclusion are comprised of:
  - an antecedent;
  - a type_of_relation (holding the semantic of the relation in a domain, for example, in biomedical sciences); and
  - the consequent.

The antecedent and consequent may be two different phenomena or one single phenomenon and its characteristics.

Articles differ in the way they are built around previously-stated hypotheses, or those stated by authors other than the authors of the current article, or new, original hypotheses, i.e., those stated by the authors of the current article. Articles may also differ in documenting an experiment or just discussing theoretical considerations when comparing previously-stated hypotheses. In the model proposed, four types of articles are identified by the patterns of reasoning they contain: theoretical articles and experimental articles, which may be just exploratory articles or employ inductive or deductive reasoning. The four patterns of reasoning are described as follows:

- Theoretical-abductive (TA) articles analyze different, previous hypotheses, showing their faults and limitations and proposing a new hypothesis; the reasoning is as follows:
  - A problem is identified, with the following aspects and data … ;
  - The previous hypotheses (from other authors) are not satisfactory to solve the problem due to the following criticism … ;
  - Therefore, we propose this new (original) hypothesis, which we consider a new pathway to solve the problem.

- Experimental-inductive (EI) articles propose a hypothesis and develop experiments to test and validate it; the reasoning is as follows:
  - A problem is identified, with the following aspects and data … ;
  - A possible solution to this problem can be based on the following new hypothesis … ;
  - We developed an experiment to test this hypothesis and obtained the following results.

In experimental-inductive articles, a conclusion may be mainly one of these alternatives: it corroborates the hypothesis, refutes it, or partially corroborates the hypothe-

sis. However, in some cases, the conclusion is not one of the former; it simply reports intermediate but not conclusive results toward the hypothesis corroboration.

- Experimental-deductive (ED) articles use a hypothesis proposed by other researchers, cited by the articles' author, and apply it to a slightly different context; the reasoning is as follows:
  - A problem is identified, with the following aspects and data … ;
  - In the literature, the previous hypotheses (by other authors) have been proposed … ;
  - We choose the following previous hypothesis … ;
  - We enlarge and recontextualise this hypothesis; we develop an experiment to test it in this new context … ;
  - The experiment shows the following results in this new context.

- Experimental-exploratory (EE) articles usually are not hypothesis driven; their objective is to acquire knowledge about a poorly-understood scientific phenomenon by performing an experiment; the reasoning is as follows:
  - There is a phenomenon that is poorly understood in a scientific domain.
  - We developed an experiment that permits the identification of the following characteristics of this phenomenon.

These basic semantic components of scientific articles are interrelated and structured. Together with the corresponding bibliographic metadata and article text, they form richer article surrogates able to be coded in machine-understandable formats. Once coded, they constitute single digital objects which may be stored in a digital library or electronic journal publishing system. All these features are formalized in the semantic model of articles (SMA) illustrated in Figure 1.

5.0 Model implementation and potentialities

This section describes partial implementations of the model proposed, developed with the aim of demonstrating its feasibility.

Nowadays considerable effort has been expended in extracting and formalizing knowledge from the unstructured text of scientific articles. Within this scope a main focus is the article's conclusion and its feasibility to be represented as a relationship using RDF coding (Samwald, 2009). This trend could be enhanced if the article conclusion as a knowledge unit could be captured as a structured RDF triple as in the model proposed.

Recently, different ontologies have been proposed with the aim of aggregating semantics processable by programs to scientific articles in digital format. Among others are CITO, the citation ontology, which aims to make explicit the reasons for citation in a scientific article, and EXPO, ontology for biomedical experiments, which is an intermediate-level ontology between the top-level ontology SUMO and domain-specific level ontologies. It aims to formalize knowledge about scientific experiments' designs, methodologies procedures, and results.

The use foreseen for these ontologies is the annotation of scientific papers or scientific experiments (Soldatova and King 2005). In practical use cases, annotation is a problematic task, as it may be considered as an additional task among many others performed by scientists in their everyday work. Although these ontologies are very expressive in formalizing scientific knowledge, their use in annotation is a strong limitation to their practical use.

The components described in the model, once coded in program-understandable formats using semantic web standards, constitute rich knowledge representations, which can enable direct management of knowledge contained in scientific articles, their use in automatic reasoning and inference tasks applied to different and unpredicted contexts, thus enlarging the possibilities for automatic processing of the rich digital content now available throughout the Web.

However, the semantic components provided by the model are hard to capture within the scope of the current scholarly electronic publishing environment. We avoid using them for annotation. Thus, we propose to engage authors in developing a richer content representation of their articles. To take full advantage of the facilities provided by the model, a future scholarly electronic publishing framework should be developed, a scholarly electronic article editor and submission system able to capture, formalize, and code the elements provided by the model.

We propose here some initial steps toward this framework. Researchers are accustomed to self-describing their papers when submitting them to a digital library, to a conference, to a digital repository or to a journal system. The submission of an article to a journal system is a privileged process during which authors are particularly motivated to clarify and disambiguate questions about their articles. In our proposal we take advantage of this moment. We have developed a prototype system of a Web author's submission interface to a journal system, which partially implements the model (Costa 2010). In the prototype developed, authors use a Web submission interface to a journal system to type, in addition to standard metadata, the article conclusions at the moment of submission and upload of the article text. The system

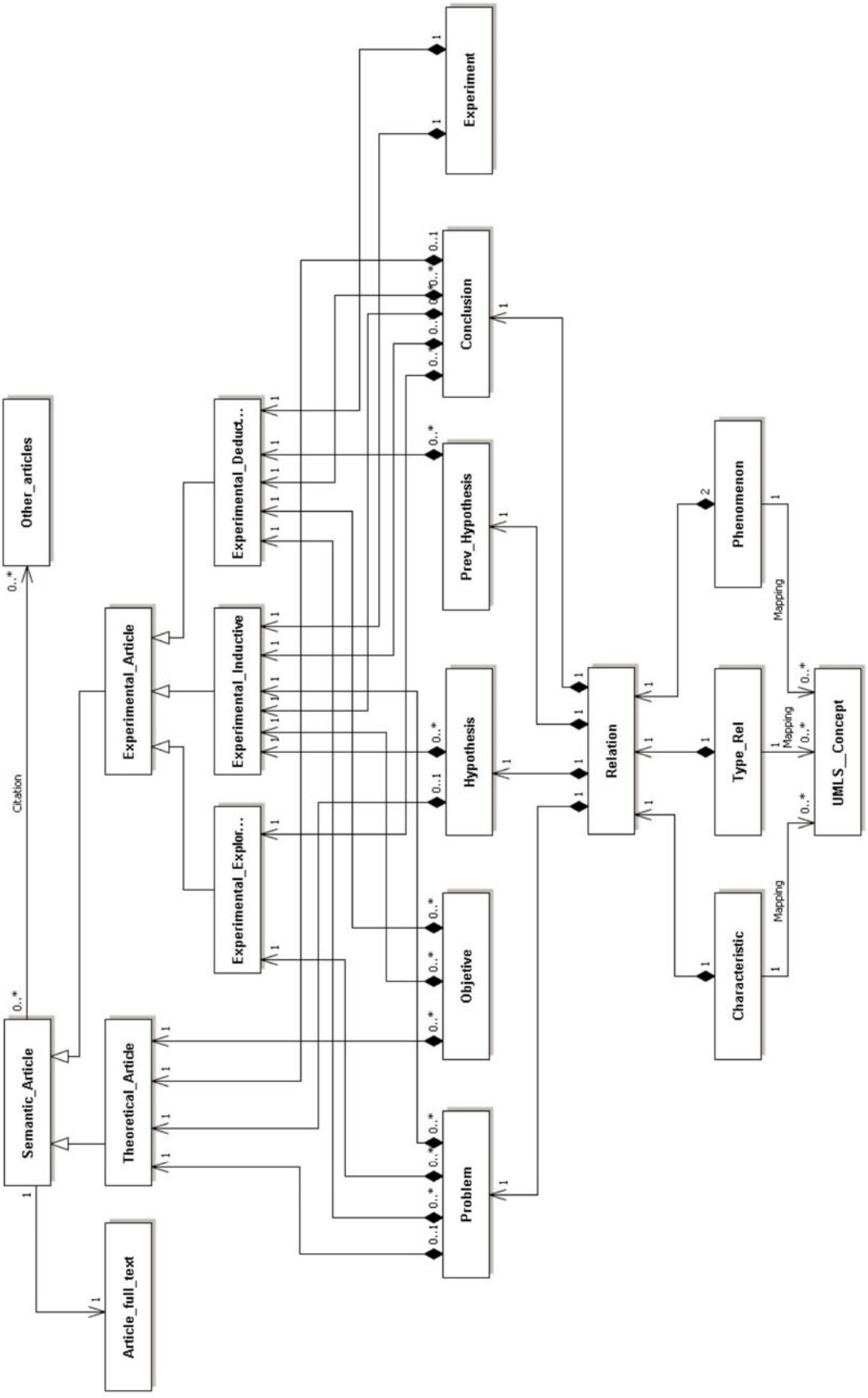

Figure1. Semantic Model of Articles. The dashed line delimits the portion of the model that was implemented.

	Questions	Author 1	Author 2	Author 3	Author 4	Author 5	Author 6
1	*I think that I would like to use this system frequently*	5	3	4	5	4	4
2	*I found the system unnecessarily complex*	1	1	1	1	2	2
3	*I thought the system was easy to use*	5	2	4	5	4	4
4	*I think that I would need the support of a technical person to be able to use this system*	1	1	1	1	1	1
5	*I found the various functions in this system were well integrated*	4	3	4	5	4	3
6	*I thought there was too much inconsistency in this system*	1	3	1	1	1	1
7	*I would imagine that most people would learn to use this system very quickly*	5	4	4	4	4	4
8	*I found the system very cumbersome to use*	1	1	1	1	3	1
9	*I felt very confident using the system*	4	3	5	4	3	4
10	*I needed to learn a lot of things before I could get going with this system*	4	1	1	1	1	1

Table 1. Results of system usability test.

performs natural language processing (NLP) of the conclusion, breaking it into short pieces of text as it is typed, formatting it as a relationship. The system interacts with authors, asking them to validate the relation extracted and the mapping done by the system of concepts found in the conclusion to concepts in a domain-specific, terminological knowledge base. In the case of the prototype developed, the terminological knowledge base used is the UMLS. The results of this processing are recorded as a bibliographic record, rich, in semantic content, in which scientific claims made by authors throughout the articles are expressed by relations. Each article, in addition to being published in textual format, has its claims also represented as structured relations and recorded in program-understandable format using semantic web standards such as RDF and OWL. These semantic records are also formally related by the author, i.e. mapped and annotated to concepts in a standard terminological knowledge base expressing the author's own view and judgment of how the conclusion of the article might be represented in such terminology.

Authors are asked to validate the automatic mapping made by the system, even choosing other terms from a list displayed by the system or deciding that there is not any satisfactory mapping among the options offered; in this case, the system assigns "no mapping" to this specific element of the relation. The article conclusion, formatted as a relation and with terms of its "antecedent," "type_of_ relation" and "consequent" annotated by the author to terms in the UMLS is then recorded as a rich article surrogate.

Despite the difficulties in engaging authors to test the prototype a test was performed with six authors, professors and a researcher of the Biomedical Institute of our university. An interview was performed with each author, relative to an article of his or her authorship. Authors were asked to identify within the article text the question, the hypothesis and the article conclusion. Afterward, authors were asked to submit the article's conclusion to the prototype. Authors spent an average time of twelve minutes and twenty-three seconds to interact with the prototype to submit an article. After using the prototype, authors were asked to respond to a questionnaire to evaluate system usability (Usability.gov). The System Usability Scale (SUS) provides a quick and dirty reliable tool for measuring the usability. It consists of a ten-item questionnaire with five response options for respondents; from 5 (Strongly agree) to 1 (Strongly disagree). The prototype average usability was high, 83.75. Results are as follows and are shown in Table 1.

Besides the conventional bibliographic metadata the only additional metadata that the prototype interface asks to the authors to type are the article's objective and conclusion. In all test cases the prototype was able to formalize the article conclusion as a relationship and the formalized conclusion was approved by the author.

Figures 2-4 illustrate some of the steps involved in the processing of the following conclusion (Segundo et al. 2004): "The results presented herein emphasize the importance to accomplish systematic serological screening during pregnancy in order to prevent the occurrence of elevated number of infants with congenital toxoplasmo-

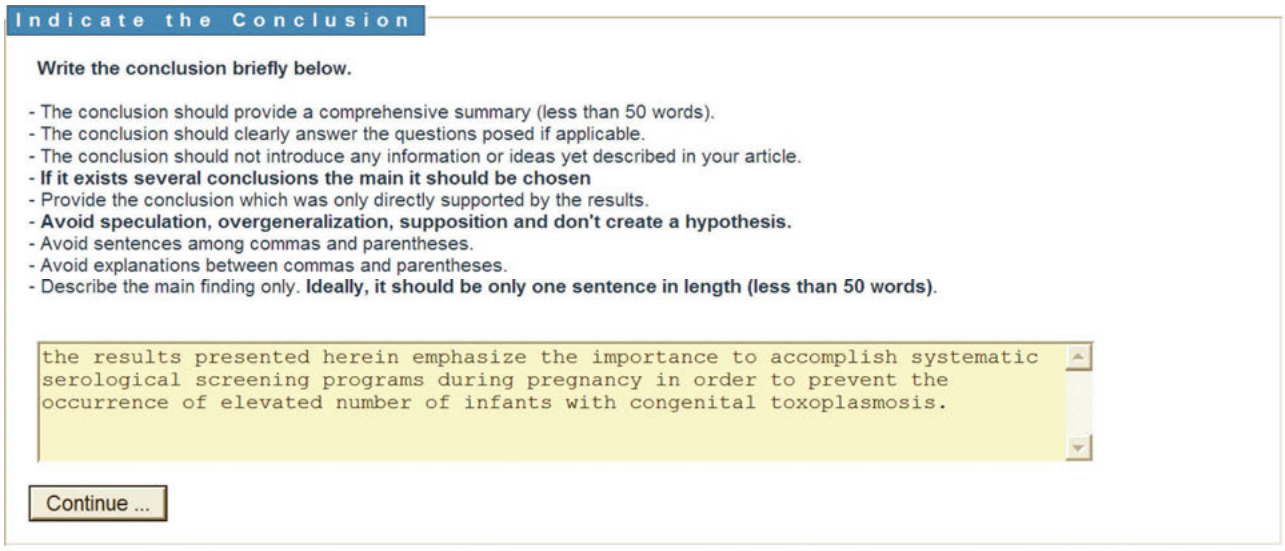

Figure 2. The author specifies the article conclusion.

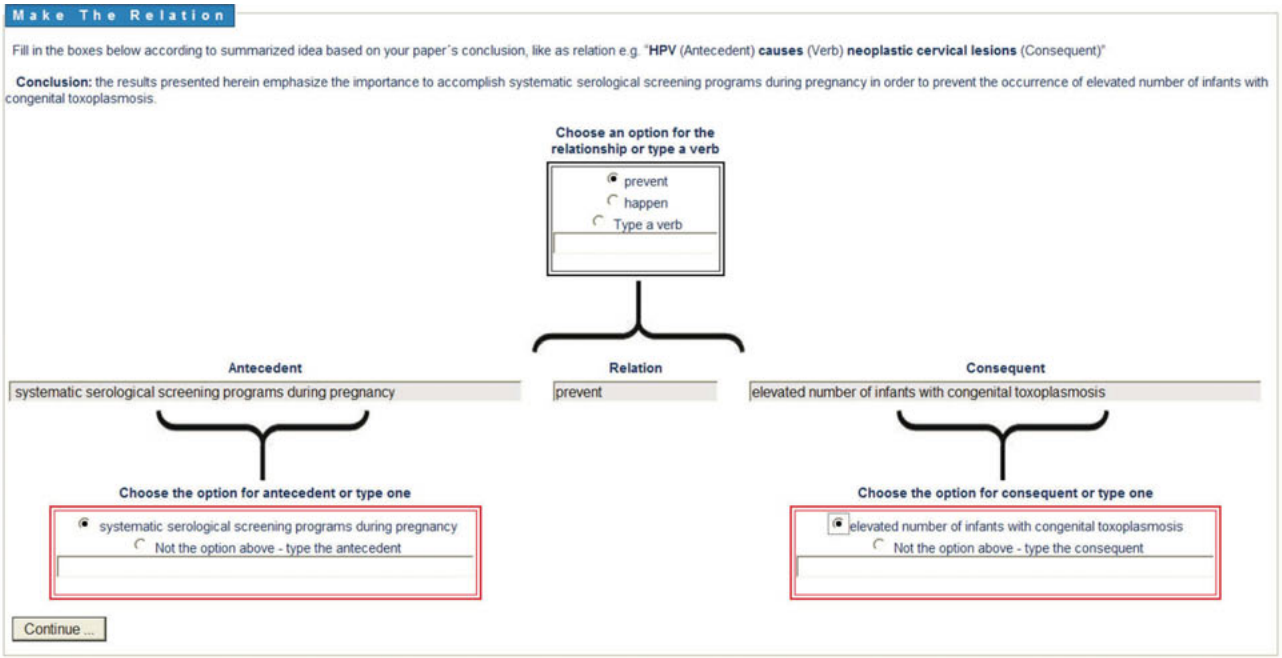

Figure 3. The article conclusion is formatted as a relation.

sis." The screens were captured from the prototype system developed in Costa (2010).

This framework enables the posterior use of these surrogates to compare their content to terminologies like the UMLS in order to identify related claims in different articles or traces of discoveries at the moment of article publication, which may be advantageous when compared to methods such as article citations.

Figures 5-6 show how the conclusion "telomere replication [Antecedent] involves [Type_of_relation] a terminal transferase-like activity [Consequent]," found in (Greider and Blackburn. 1987), may be formated in RDF.

Both RDF documents in Figure 5 are named-graphs, meaning that their triples are identified by a URI. Even a partial implementation of the record model proposed in RDF, where the only semantic component captured is the conclusion, will facilitate more expressive semantic retrieval from a knowledge network enabling queries such as the following:

– Which other articles have hypotheses suggesting HPV as the cause of cervical neoplasias in women?
– Which articles have hypotheses suggesting other causes of cervical neoplasias different from HPV in women?

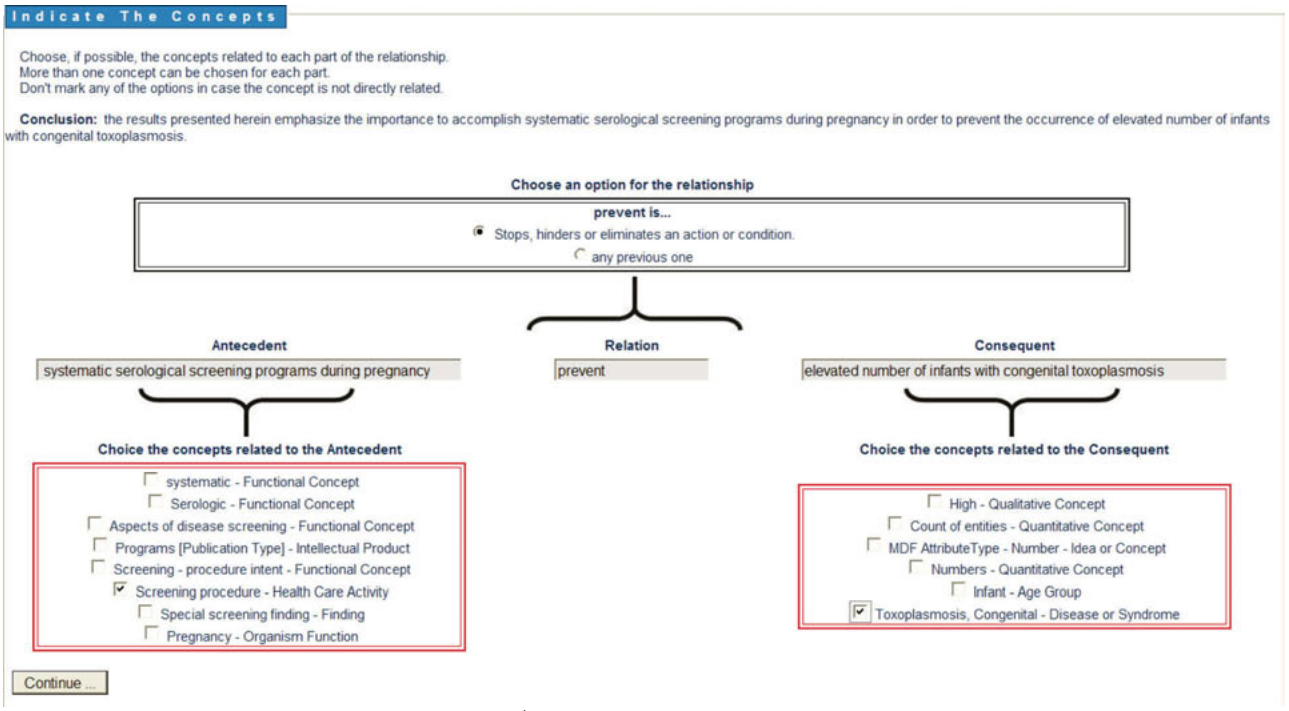

Figure 4. Authors are asked to map/annotate concepts in the article's conclusion to UMLS terms.

– Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in groups different from women?
– Which articles have hypotheses suggesting HPV as the cause of pathologies different from neoplasias?
– Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in different contexts (not in women from Federal District, Brazil)?

The model also enables queries that may indicate new discoveries, for example, new causes of cellular senescence:

– Which experimental-inductive articles propose (Antecedent?) causes (Type_of_relation) to cellular senescence (Consequent) that are not mapped to UMLS concepts?
– Is there any confirmation of the hypothesis that "Several aspects of both the structural and dynamic properties of telomeres (Antecedent) led to the proposal that telomere replication involves (Type_of_relation) non-template addition of telomeric repeats onto the ends of chromosomes (Consequent)" (Shampay et al. 1984)?
– Who first maintained, and when, that "the RNA component of telomerase (Antecedent) may be directly involved in (Type_of_relation) recognizing the unique three-dimensional structure of the G-rich telomeric oligonucleotide primers (Consequent)" (Greider and Blackburn 1987)?

The article's types and corresponding reasoning patterns described in the model enabled classification of the 89 articles analyzed. The results are the following: 27 articles were classified as as experimental-inductives (EI), 44 as experimental-deductives (ED), 15 as experimental-exploratories (EE), and 3 as theoretical-abductives (TA). Details can be found at Marcondes et al. (2014).

These general frameworks may be used to identify discoveries reported in scientific articles based on two aspects: the evolution of their rhetorical patterns within a chronological series of articles reporting an important scientific discovery and by comparing the content of the articles' conclusions with terminological knowledge bases. We found a characteristic evolution of these patterns in the Lasker Medical Award 2006 articles group, as predicted by authors on scientific discovery. Articles within this group have a specific reasoning pattern when analyzed chronologically; additionally, the mapping/non-mapping of terms in the conclusions to terminological knowledge bases as an indicator, as proposed, also shows a specific pattern within the same chronological series. Articles published before the publishing of the article which marks the discovery of the telomerase enzyme and named it in 1985 were all of the experimental-exploratory (EE) type and achieved non-mapping (NM) of the concepts in their conclusions to *MeSH* (*Medical Subject Headings*) concepts. Experimental-exploratory (EE) articles seem to characterize first steps toward the discovery of new phenomena. After the discovery of telomerase and the coinage of its scientific name in 1985, the first non-experimental-exploratory (EI and ED) articles appear. Furthermore, just after 1986 appear the first partially

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:sa="http://example.org/semarticles/"
  xmlns:umls="http://www.nlm.nih.gov/research/umls/">
   <rdf:Description rdf:about="http://art_id/">
   <dc:title>title</dc:title>
   <dc:creator>creator</dc:creator>
   <dc:subject>subject</dc:subject>
   <dc:date>date</dc:date>
     <sa:conclusion>http://art_id/conclusion</sa:conclusion>
   </rdf:Description>
</rdf:RDF>



<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:sa="http://example.org/semarticles/"
  xmlns:umls="http://www.nlm.nih.gov/research/umls/">
   <rdf:Description rdf:about="http://art_id/conclusion">
    <sa:antecedent>telomere replication</sa:antecedent>
   <umls:antecedent_mapping>
      http://www.nlm.nih.gov/research/umls/CUI01
   </umls:antecedent_mapping>
    <sa:type_rel>involves</sa:type_rel>
   <mapping dc:contribuitor=" author">
   <umls:type_rel_mapping>
      http://www.nlm.nih.gov/research/umls/CUI02
   </umls:type_rel_mapping>
            <sa:consequent>terminal transferase-like activity</sa:consequent>
   <umls:consequent_mapping>
      http://www.nlm.nih.gov/research/umls/CUI03
   </umls:consequent_mapping>
   </mapping>
   </rdf:Description>
</rdf:RDF>
```

Figure 5. An article and its conclusion, represented as two RDF documents. CUI means "concept unique identifier"

mapped (PM) articles. Details can be found in Malheiros and Marcondes (2013).

The model might also enable researchers to find articles with related claims, from which new knowledge may be inferred, as in the following example. Suppose an article's conclusion claims that "telomere shortening causes cellular senescence," while another article's conclusion claims that "telomerase activity is associated with cancer." The concepts "telomere shortening" and "telomerase activity" are both mapped, i.e., linked, to the same UMLS concept, which is identified by its Concept Unique Identifier (CUI) as "telomerase activity," which is a generic term relative to the first; a software agent might infer a new claim, i.e., that (maybe) "telomere shortening" "is associated with" "cancer." The claim is trusted based on the evidence presented in the experiments described in both articles and by the judgment of journal referees, who certified that both articles had sufficient scientific quality to merit publication. Figure 7 illustrates this case.

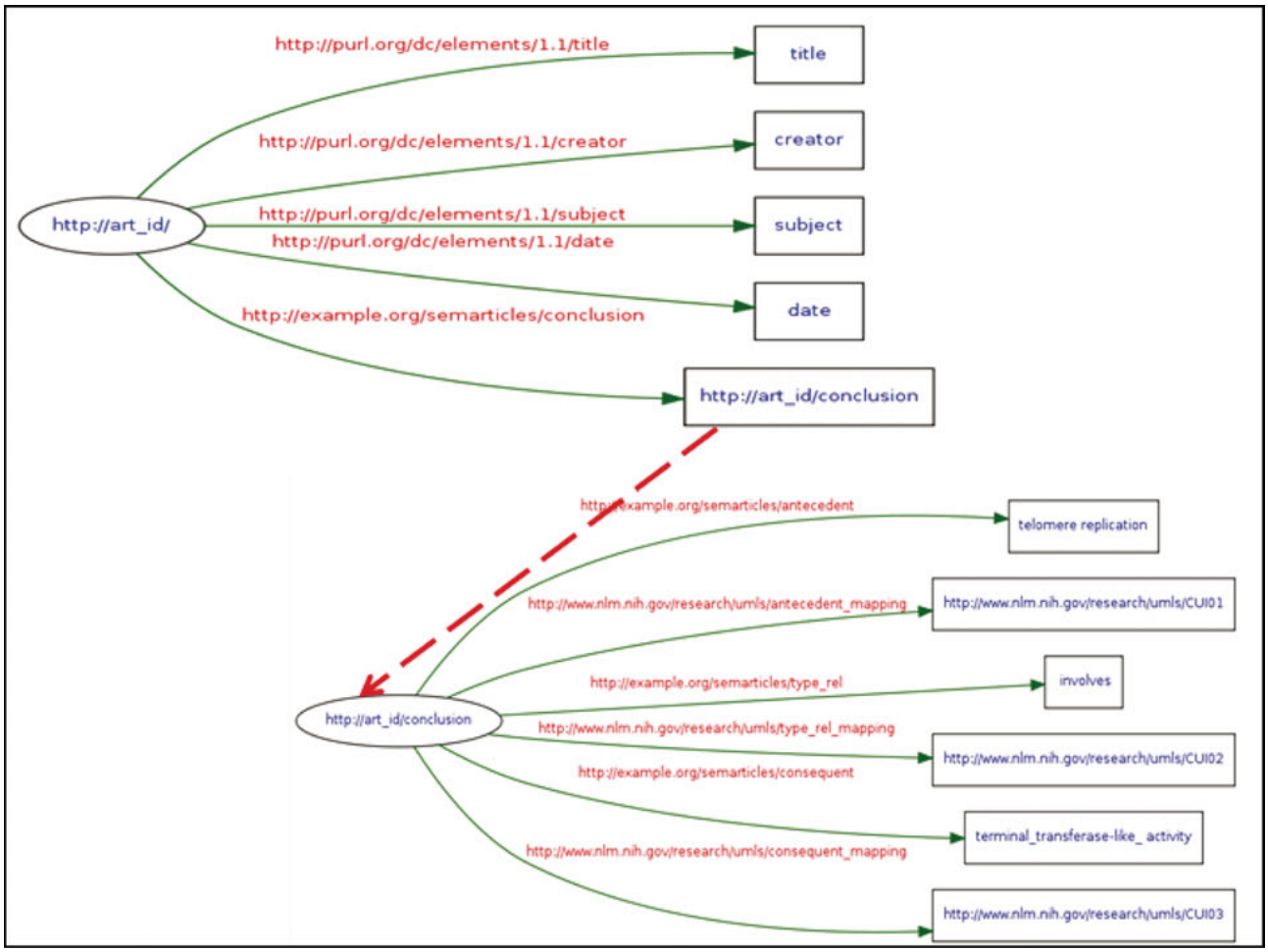

Figure 6. RDF data set representing a semantic article with its conclusion.

These examples show how the knowledge representation schema proposed may improve semantic retrieval and the use of knowledge in different and unpredicted contexts.

The model proposed is broad and aims to encompass the complete scientific publishing and information retrieval environment throughout the Web. For this reason, it is very difficult to achieve a comprehensive test which could validate the model as whole. Such a test also depends on the development of software tools that have not been developed yet. More tests and experiments must be achieved to arrive at full potencialities. Our research group has not been able to fully develop the model to the potentialities outlined here. We consider the model outlined a starting point that can be discussed and built upon by the scientific community.

6.0 Concluding remarks

Several signs indicate that digital scientific articles are overcoming the limitations of the paper-print model that prevailed for centuries in scientific journals and now are heading towards a semantic format. The adoption of structured abstracts by many biomedical journals is a move in this direction. Within such a context, in light of the new computer-driven scientific methods, the question arises of whether the semantic components of scientific methodology, such as problem, question, and hypothesis, are still valid as necessary stages in the construction of solid scientific claims and bases of scientific knowledge. The amount of scientific literature published throughout the Web is becoming increasingly vast and complex. It will be necessary for scientists to have enhanced software tools in order to process this content. The Web provides a wholly new platform for publishing, sharing, and interlinking scientific activities and data. At present, the distinction between electronic journals and databases also are blurring. This integrated knowledge network could be crawled by software agents, thus helping scientists in semantic retrieval, knowledge reuse, validation of scientific results and identification of traces of scientific discoveries, new scientific insights, and knowledge contradictions or inconsistencies.

In this paper we propose a semantic model of scholarly electronic articles in biomedical sciences that can

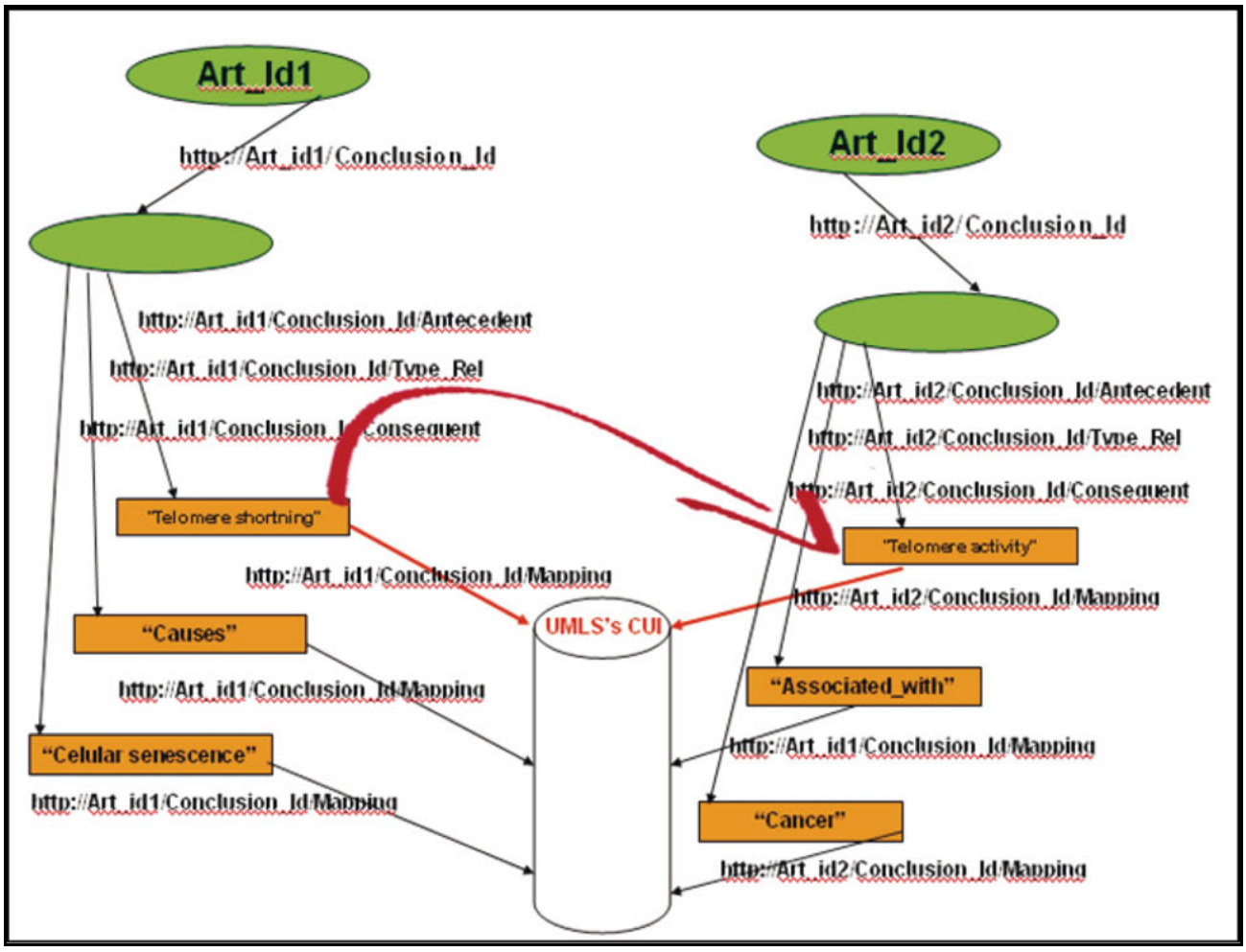

Figure 7. Related claims in two semantic articles

overcome the limitations of traditional flat records formats. We also describe some partial implementations of the model that demonstrate its feasibility and utility to manage scientific knowledge. Knowledge organization can go beyond just using conventional indexing techniques to provide quick access to full-text scientific articles. It can help scientists to directly process the knowledge content of scientific articles and to recognize the reasoning that leads to a scientific discovery. The model proposed also points to the standardization of a scientific knowledge markup language encompassing the knowledge content of Web-published scientific articles, taking a step forward to proposals like those of Murray-Rust and Rzepa (1999; 2002) and Hucka et al. (2003). This opens a new perspective in scientific electronic publishing, knowledge acquisition, storage, processing, and sharing.

References

Ad Hoc Working Group for Critical Appraisal of Medical Literature. 1987. "A Proposal for More Informative Abstracts of Clinical Articles." *Annals of Internal Medicine* 106, no. 4: 598-604.

ArrayExpress. http://www.ebi.ac.uk/arrayexpress/

Bacon, Francis. 1973. *Novum Organum.* São Paulo: Abril Cultual.

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284, no. 5: 29-37.

Bazerman, Charles. 1988. *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science.* Madison: University of Wisconsin Press.

Bianchini, Carlo and Mauro Guerrini. 2009. "From Bibliographic Models to Cataloging Rules: Remarks on FRBR, ICP, ISBD, and RDA and the Relationships Between Them." *Cataloging & Classification Quarterly* 47, no. 2: 105-24.

Bizer, Christian, Tom Heath and Tim Berners-Lee. 2009. "Linked Data-The Story So Far." *International Journal on Semantic Web and Information Systems* 5, no. 3:1-22.

Brookes, Bertram C. 1980. "The Foundations of Information Science. Part I. Philosophical Aspects." *Journal of Information Science* 2, nos. 3-4: 125-33.

Bunge, Mario. 1998. *Philosophy of Science*. New Brunswick, N.J.: Transaction Publishers.

Ceusters, Werner and Barry Smiths. 2006. "A Realism-Based Approach to the Evolution of Biomedical Ontologies." In *Biomedical and Health Informatics: From Foundations to Applications to Policy: AMIA Annual Symposium Proceedings*, edited by David W Bates, John H. Holmes and Gilad J. Kuperman. Bethesda, Md.: American Medical Informatics Association, 121-5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839444/

Chen, Peter Pin-Shan. 1976. "The Entity-Relationship Model—Toward a Unified View of Data." *ACM Transactions on Database Systems* 1, no. 1: 9-36. http://csc.lsu.edu/news/erd.pdf

CITO, The citation ontology. http://speroni.web.cs.unibo.it/cgi-bin/lode/req.py?req=http:/purl.org/spar/cito

Costa, Leonardo Cruz da. 2010. *Uma proposta de processo de submissão de artigos científicos à publicações eletrônicas semânticas em ciências biomédicas*. Tese doutorado, Programa de Pós-graduação em Ciência da Informação UFF-IBICT, Niterói., Brazil.

Dahlberg, Ingetraut. 1995. "Conceptual Structures and Systematization." *International Forum on Information and Documentation* 20, no. 3: 9-24.

DrugBank. http://www.drugbank.ca/

Dublin Core. http://dublincore.org/

EXPO. Ontology for Biomedical Experiments. http://expo.sourceforge.net/

Farradane, J. 1980. "Relational indexing. Part I." *Journal of Information Science* 1, no. 5: 267-76.

Garfield, Eugene, Irving H. Sher and Richard J. Torpie. 1964. *The Use of Citation Data in Writing the History of Science*. Philadelphia: The Institute for Scientific Information.

GenBank. http://www.ncbi.nlm.nih.gov/genbank/

GO – Gene Ontology. http://geneontology.org/

Greider, Carol W. and Elizabeth H. Blackburn. 1985. "Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts." *Cell* 43: 405-13.

Greider, Carol W. and Elizabeth H. Blackburn. 1987. "The Telomere Terminal Transferase of Tetrahymena is a Ribonucleoprotein Enzyme with Two Kinds of Primer Specificity." *Cell* 51: 887-98.

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

Hao, Ling-Yang, Mary Armanios, Margaret A. Strong, Baktiar Karim, David M. Feldser, David Huso and Carol W. Greider. 2005. "Short Telomeres, Even in the Presence of Telomerase, Limit Tissue Renewal Capacity." *Cell* 123: 1121–31.

Hucka, Michael, et al. 2003. "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models." *Bioinformatics* 19, no. 4: 524-31.

Hutchins, John. 1977. "On the Structure of Scientific Texts." In *Proceedings of UEA Papers in Linguistics, September 5th 1977, Norwich, UK*: University of East Anglia, 18-39.

International Committee of Medical Journals Editors. 2003. http://www.icmje.org

IFLA Study Group on Functional Requirements for Bibliographic Records.1998. *Functional Requirements for Bibliographic Records: Final Report*. München: K. G. Saur.

Malheiros, Luciana Reis and Carlos Henrique Marcondes. 2013. "Identificación de indicios de descubrimientos científicos en artículos biomédicos mediante análisis de contenidos." *Revista Española de Documentación Científica* 36, no. 2. doi: http://dx.doi.org/10.3989/redc.2013.2.915

MARC Standards. http://www.loc.gov/marc/

Marconi, Marina de Andrade and Eva Maria Lakatos. 2004. *Fundamentos de Metodologia Científica*. 5th ed. São Paulo: Atlas.

Marcondes, Carlos Henrique. 2005. "From Scientific Communication to Public Knowledge: The Scientific Article Web Published as a Knowledge Base." In *From Author to Reader: Challenges for the Digital Content Chain: Proceedings of the 9th ICCC International Conference on Electronic Publishing, June 8-10, 2005, Katholieke Universiteit Leuven, Belgium,* edited by Milena Dobreva and Jan Engelen. Leuven, Belgium: Peeters, 119-126. http://eprints.rclis.org/bitstream/10760/7389/1/ELPUB_2005-Marcondes.pdf

Marcondes, Carlos H., Luciana R. Malheiros and Leonardo C. da Costa. 2014. "A Semantic Model for Scholarly Electronic Publishing in Biomedical Sciences." *Semantic Web Journal* 5, no. 4: 313-34.

Miller, David L. 1947. **"**Explanation versus Description." *Philosophical Review* 56, no. 3: 306-12.

Murray-Rust, Peter and Henry S. Rzepa. 1999. "Chemical Markup, XML and the World Wide Web. I: Basic principles." *Journal of Chemical Information and Computer Science* 39, no. 6: 928-42.

Murray-Rust, Peter and Henry S. Rzepa. 2002. "STMML. A Markup Language for Scientific, Technical and Medical Publishing." *Data Science Journal* 1 no. 2: 128-93.

Nwogu, Kevin Ngozi. 1997. "The Medical Research Paper: Structure and Functions." *English for Specific Purposes* 16, no. 2: 119-38.

Open Biomedical Ontologies (OBO) Foundry. http://www.obofoundry.org/

Ontology for Experiment Self-Publishing. http://www.w3.org/wiki/HCLS/ScientificPublishingTaskForce

OWL. http://www.w3.org/2001/sw/wiki/OWL

PhenomicDB. http://www.phenomicdb.de/

Renear, Allen H. and Carole L. Palmer. 2009. "Strategic Reading, Ontologies and the Future of Scientific Publishing." *Science* 325, no. 5942: 828-32.

Salager-Meyer, Françoise. 1991. "Brief Communication Medical English Abstracts: How Well Are They Structured?" *Journal of the American Society for Information Science* 42: 528-31.

Samwald, Matthias. 2009. "Extracting Conclusion Sections from Pubmed Abstracts for Rapid Key Assertion Integration in Biomedical Research." *Nature Proceedings* 3775, no.1. doi:10.1038/npre.2009.

Schulz, Stefan and Ludger Jansen. 2013. "Formal Ontologies in Biomedical Knowledge Representation." *Yearbook of Medical Information* 8, no. 1: 132-46.

Segundo, Gesmar Rodrigues Silva Deise Aparecida Oliveira Silva, José Roberto Mineo and Marcelo Simão Ferreira. 2004. "A Comparative Study of Congenital Toxoplasmosis Between Public and Private Hospitals from Uberlândia, MG, Brazil." *Memórias do Instituto Oswaldo Cruz* 99, no. 1: 13-7.

Shadbolt, Nigel, Tim Brody, Les Carr and Stevan Harnad. 2006. "The Open Research Web: A Preview of the Optimal and the Inevitable." In *Open Access: Key Strategic, Technical and Economic Aspects, edited by Neil* Jacobs, 10-24. Oxford: Chandos.

Shampay, Janis, Jack W. Szostak and Elizabeth H. Blackburn. 1984. "DNA Sequences of Telomeres Maintained in Yeast." *Nature* 310: 154-7.

Sheth, Amit, I. Budak Arpinar and Vipul Kashyap. 2004. "Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships." In *Enhancing the Power of the Internet*, edited by Masoud Nikravesh, Ben Azvine, Ronald Yager and Lotfi A. Zadeh, 63-94. Berlin; New York: Springer-Verlag.

Skelton, John. 1994. "Analysis of the Structure of Original Research Papers: An Aid to Writing Original Papers for Publication." *British Journal of General Practice* 44: 455-9.

Soldatova, Larisa N. and Ross D. King. 2005. "Are the Current Ontologies in Biology Good Ontologies?" *Nature Biotechnology* 23, no. 9: 1095–8. doi:10.1038/nbt0905-1095

Structured Abstract Labels Research Dataset. http://structuredabstracts.nlm.nih.gov/downloads.shtml

Swanson , Don R., Neil R. Smalheiser and Vetle I. Torvik. 2006. "Ranking Indirect Connections in Literature Based Discovery: The Role of Medical Subject Headings." *Journal of the American Society for Information Science and Technology* 57: 1427–39.

UMLS Semantic Network. http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

UMLS. http://www.nlm.nih.gov/research/umls/

Usability.gov. System Usability Scale (SUS). http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

W3C. Resource Description Framework (RDF). http://www.w3.org/RDF/

W3C. Resource Description Framework (RDF) Schema Specification 1.0. http://www.w3.org/TR/2000/CR-rdf-schema-20000327/