Representing and organizing scientific knowledge in biomedical articles with Semantic Web technologies

Carlos H. Marcondes
Federal Fluminense University
R. Lara Vilela, 126, CEP 24210-590
Niterói, RJ, Brazil
tel 55 21 26299758
marcon@vm.uff.br

ABSTRACT

Currently, the conventional channel for reporting scientific results is the Web electronic publishing. Despite the advances in electronic publishing scientific articles are still published in paper print formats such as PDFs. The emergence of the Semantic Web and Linked Data environment provides new opportunities for communicating, sharing, and integrating scientific knowledge in digital formats that can overcome the limitations of the current print format, suitable only for reading by people. This paper explores the possibilities of this new environment proposing a semantic model of scholarly electronic articles in biomedical sciences. Thereby the results of scientific research can be published electronically, shared in structured, interlinked formats, enabling their crawling by software agents, facilitating semantic retrieval, knowledge reuse, validation of scientific results, identification of traces of scientific discoveries, new scientific insights, and identification of knowledge contradictions or inconsistencies.

Categories and Subject Descriptors

E.2 [**Data Storage Representations**]: Object representation
H.2 [**Data Management**]: Data Models
I.7 [**Document and Text Processing**]: Electronic Publishing

General Terms

Management, Experimentation, Human Factors.

Keywords

semantic publishing, Semantic Web, knowledge representation, biomedical knowledge, knowledge network, terminological knowledge bases, escience

1. INTRODUCTION

The conventional communication channel for reporting scientific results is the electronic publishing of scientific articles in formats, such as PDFs. The current scholarly Web publishing environment is still an electronic metaphor of the paper print publishing environment used throughout the twentieth century.

Despite numerous advances in information technology, Web electronic publishing is still based on the print text model; scientific results are still published in articles in textual format, which limits the possibilities for automatic reasoning on their contents for reuse, discovery, analysis and validation of scientific results. Knowledge is embedded in the text of scientific articles for human reading. Web electronic publishing is far from taking full advantage of the facilities offered by Semantic Web and Linked Open Data technologies.

Such tasks are 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. Today PubMed repository holds over 23 million articles. According to Renear and Palmer [1] scientists are increasingly using strategic reading to cope with the amount of literature being published. The tasks outlined demand new tools for information discovery, retrieval, and content comparison in very specific, precise, and meaningful manners.

Current information retrieval systems are based in Boolean operators which do not implement explicit meaningful relations between elements which comprise the surrogates of documents or resources they represent. Boolean operators are too general and lack the semantic expressiveness necessary for content retrieval in specific scientific domains. Relations expressed using Boolean operators are processed in the current information retrieval systems as extensive set operations using the keywords included in the bibliographic records, rather than as intensive semantic relations among concepts.

Many relations encompassing scientific articles can be identified: bibliographic/citation relations; relations with datasets holding raw results of scientific experiments or databases such as GenBank[1], DrugBank[2], ArrayExpress[3], PhenomicDB[4]; internal relations between parts of an article – semantic elements – such as problem, question, hypothesis, methodology, results and conclusion; relations with terminological knowledge bases or ontologies such as UMLS[5], GO[6]; relations with grant agencies; relations with claims within an article and across articles; relations

Conference'10, Month 1–2, 2010, City, State, Country.

[1] GenBank, http://www.ncbi.nlm.nih.gov/genbank/

[2] DrugBank, http://www.drugbank.ca/

[3] ArrayExpress, http://www.ebi.ac.uk/arrayexpress/

[4] PhenomicDB, http://www.phenomicdb.de/

[5] UMLS, http://www.nlm.nih.gov/research/umls/

[6] GO, http://www.nlm.nih.gov/research/umls/

within two different bibliographic sets (literature-related discovery methodologies) [2]; relations with annotations/comments made about an article.

The current information retrieval system/citation systems and the print model of publication constitute closed systems were scientific articles are isolated from web mainstream and thus are barriers to data reuse, sharing, integration and synthesis. This situation can now be overcome within the scope of Semantic Web/Linked Open Data platform.

Scientific knowledge aims at universality and necessity. Such characteristics make this kind of knowledge susceptible of large reuse. Miller [3] states that: ‘science is a search after internal relations between phenomena’. Scientific knowledge, as it appears in the text of scientific articles, consists in claims made by authors throughout the article text, synthesized in the article’s conclusions. These claims can be seen as units of 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.

2. PROPOSAL

Here is proposed a semantic model of scholarly electronic articles which *extends* conventional bibliographic record models, comprised of conventional descriptive elements such as authors, title, abstract, bibliographic source, publication date, content information such as keywords or descriptors and references to cited papers. Such a model is designed to be implemented in the Semantic Web/Open Linked Data plataform.

In addition to bibliographic elements the model includes also the *claims* made by authors throughtout the article text. These claims are not explicitly represented nor coded in conventional bibliographic records and are hard to find in article text. The aim of the semantic model is to enable the coding of independent claims made in biomedical articles as “knowledge units” in program “understandable” format such that this knowledge can be reused by software agents in task which demand intelligent processing. These claims take the form of relations between phenomena or between a phenomenon and its characteristics. They are expressed linguistically through propositions [4], e.g. a-“telomere shortening (Phenomenon) causes (Type_of_relation) cellular senescence (Phenomenon)” [5], b-“telomere replication (Phenomenon) involves (Type_of_relation) nontemplate addition of telomeric repeats onto the ends of chromosomes(Phenomenon)?” [6] or c-“tetrahymena extracts (Phenomenon) show (Type_of_relation) a specific telomere tranferase activity (Characteristic)” [7].

Relations may appear in different semantic elements throughout the article text, such as within the Problem that the article addresses, as a *Question*, in which either one of the two *relata* or the type of relation is unknown; in the *Hypothesis;* or in the *Conclusion*. Frequently, the Conclusion also poses new Questions. Such relations could be modeled as triples of <Antecedent><Type_of_relation><Consequent>.

Questions, *Hypothesis* and *Conclusion* are the semantic elements comprising the model proposed. They are the elements related to the knowledge content of an article. The *Conclusion* is the essential semantic element, as it synthesizes the knowledge content of an article. Although in the scope of a recently published article, the conclusion is a provisional knowledge unit, it is at least validated by the experiment reported in the article. Semantic elements such as *Questions* and *Hypothesis* are important as they permit to trace the evolution of a claim. Other elements have rhetoric functions, as extensively discussed in [8] and [9], or serve to describe methodological options, the experiment performed, its context, or display more clearly the results obtained.

Articles differ in the way they are built around the previous stated hypotheses—those stated by authors other than the author of the current article, or new, original hypotheses, i.e., those stated by the author of the current article. Articles may also differ by the existence of a documented experiment or just theoretical considerations comparing previously stated hypotheses. In the model proposed these patterns of reasoning define four types of articles: *theoretical articles* and *experimental articles,* which may be just *exploratory* articles or employ *inductive* or *deductive* reasoning. The four patterns of reasoning are described in the sequel.

Theoretical-abductive (TA) articles analyze different, previous hypotheses, showing their faults and limitations and proposing a new hypothesis. **Experimental-inductive** (EI) articles propose a hypothesis and develop experiments to test and validate it. **Experimental-deductive** (ED) articles use a hypothesis proposed by other researchers cited by the articles’ author and apply it to a slightly different context. **Experimental-exploratory** (EE) articles are not usually hypothesis driven; their objective is to acquire knowledge about a poorly understood scientific phenomenon by performing an experiment.

These basic semantic elements of scientific articles are interrelated and structured. Together with the corresponding bibliographic metadata and article full-text they form richer article surrogates in machine-understandable formats and constitute single digital objects stored in a digital library or electronic journal publishing system. All these features are formalized in the Semantic Model of Articles (SMA); a partial view of it comprising only the conclusion as a semantic element can be seen in Figure 1. A complete version of the model and the discussion of its features and potentialities can be found in [10].

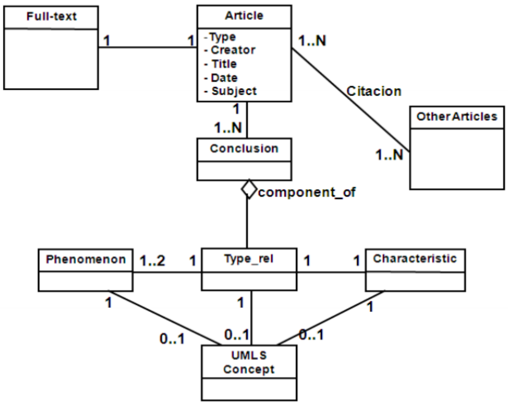

Figure1. Semantic Model of Articles

3. MODEL IMPLEMENTATION

The elements described in the model, once coded in program “intelligible” formats using Semantic Web standards, constitute rich articles surrogates which can enable direct knowledge management, their use in automatic reasoning and inference tasks applied to different and unpredicted contexts, thus enlarging of the possibilities of automatic processing of the rich digital content now available throughout the Web.

However the semantic elements provided by the model are hard to capture within the current scholar electronic publishing environment. To take full advantage of the facilities provided by the model a new scholar electronic publishing framework should be developed, a scholar electronic article editor and submission system able to capture, formalize and code the elements provided by the model. This framework should be driven by ontologies as the SMA proposed herein, but also by Ontology for Experiment Self-Publishing [11], IAO[7], OBI[8], CITO [12], etc., besides domain specific ontologies or terminologies like the UMLS, GO, SNOMED[9].

We propose some initial steps towards 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 [13]; the general framework proposed is used to identify discoveries in scientific papers based on two aspects: their rhetoric elements and patterns and by comparing the content of the articles conclusions with terminological knowledge bases [14].

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/upload of his/her article text. The system performs NLP in to short pieces of text as the conclusion typed, *formatting it as a relation*. 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 terminological knowledge base. In the case of the prototype developed, the terminological knowledge base used is the UMLS – Unified Medical Language System.

The result of this processing is recorded as a richer semantic content bibliographic record in which scientific claims made by authors throughout 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 a machine-understandable format using Semantic Web standards such as RDF (Resource Description Framework)[10] and OWL (Web Ontology Language)[11] and formally related by the author i.e. mapped, annotated, to concepts in a standard terminological knowledge base expressing his/her own view and judgement of how the conclusion of the article may be represented in such a terminology.

The author is asked to validate de automatic mapping made by the system, even choosing another terms of a list displayed by the system or deciding that there is not any satisfactory mapping between 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. We thus propose to engage authors in developing a richer content representation of their articles.

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

The following figure shows as the conclusion "telomere replication (Antecedent) involves (Type_of_relation) a terminal transferase-like activity (Consequent)," found in [15], may be formated in 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/">
   <rdf:Description rdf:about="http://art_id/">
     <dc:title>title</dc:title>
     <dc:creator>creator</dc:creator>
     <dc:subject>subject</dc:subject>
    <sa:conclusion>_:conclusion
       _:conclusion    sa:antecedent "telomere replication "
       _:conclusion    sa:type_rel   "involves"
       _:conclusion    sa:consequent "a terminal transferase-like
activity "
    </sa:conclusion>
    <sa:antecedent>_:antecedent
       _:antecedent    sa:mapping   "UMLS's CUI01"
    </sa:antecedent>
    <sa:type_rel>_:type_rel
       _:type_rel      sa:mapping    "UMLS's CUI02"
    </sa:type_rel>
    <sa:consequent>_:consequent
       _:consequent    sa:mapping    "UMLS's CUI03"
    </sa:consequent>
   </rdf:Description>
 </rdf:RDF>
```

Figure 2. Conclusion of an article, represented in RDF. CUI means concept unique identifier

Even a partial implementation of the record model proposed in RDF, where the only semantic element captured is the *conclusion,* will facilitate more expressive semantic retrieval from a knowledge network enabling queries like 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?

- 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 to cellular senescence:

- Which experimental-inductive articles propose (Antecedent?) causes (Type_of_relation) to cellular senescence (Consequent) that are not mapped to UMLS concepts?

[7] IAO, https://code.google.com/p/information-artifact-ontology/

[8] OBI, http://bioportal.bioontology.org/ontologies/OBI

[9] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[10] RDF, RDF, http://www.w3.org/TR/rdf-primer/

[11] OWL, http://www.w3.org/2001/sw/wiki/OWL

- 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) nontemplate addition of telomeric repeats onto the ends of chromosomes (Consequent)?" [16]?

- Who and when first maintained 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*)*" [17]?

The model may also find articles with related claims, from which new knowledge may be inferred, as in the following example. Suppose an article's conclusion which claims that "telomere shortening causes cellular senescence," while other article 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 infers 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 judgement of journal referees, who certified that both articles had sufficient scientific quality to merit publication.

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

4. CONCLUSION

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, share and interlinks scientific activities and data. This integrated knowledge network could be crawled by software agents thus helping scientist in semantic retrieval, knowledge reuse, validation of scientific results, identification of traces of scientific discoveries, new scientific insights, knowledge contradictions or inconsistencies.

Knowledge Organization can go beyond just using indexing conventional techniques to providing fast access to full-text scientific articles. It can help scientists to directly process the scientific articles knowledge content and to recover the reasoning that leads to a scientific discovery. The model proposed also points to the standardization of a SkML (Scientific Knowledge Markup Language) encompassing the knowledge content of Web published scientific articles, carrying on one step ahead proposals like those of [18], [19] and [20]. This opens a new perspective in scientific electronic publishing, knowledge acquisition, storage, processing, and sharing.

5. ACKNOWLEDGMENTS

Our thanks to the Brazilian grant agencies CNPq and CAPES.

6. REFERENCES

[1] Renear, A.H. and Palmer C. L.. 2009. Strategic reading, ontologies and the future of scientific publishing, Science 325, 828-832.

[2] Swason , D.R., Smalheiser ,N. R and Torvik , V. I. 2006. Ranking indirect connections in literature based discovery. The role of Medical Subject Headings, *JASIST* 57 11, 1427–1439.

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

[4] Ingetraut Dahlberg. 1995. Conceptual structures and systematization. International *Forum on Information and Documentation* 20, 3, 9-24.

[5] Hao, L.Y. et al. 2005. Short telomeres, even in the presence of telomerase, limit tissue renewal capacity. Cell 123 1121–1131.

[6] Shampay, J., Szostak, J.W. and Blackburn, E. H. 1984. DNA sequences of telomeres maintained in yeast. Nature 310, 154-157.

[7] Greider, C. W. and Blackburn, E. H. 1985. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. Cell, 43 405-413.

[8] Skelton, J. 1994. Analysis of the structure of original research papers: an aid to writing original papers for publication. British Journal of General Practice, 44, 455-459.

[9] Nwogu, K. N. 1997. The Medical Research Paper: Structure and Functions. English for Specific Purposes 16, 2, 119-138.

[10] C. H. Marcondes, L. R. Malheiros and L. C. da Costa. 2014. A semantic model for scholarly electronic publishing in Biomedical Sciences. Semantic Web Journal, 5, 4.

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

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

[13] Costa, L. C. da. 2010. Um 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.

[14] Malheiros, L. R. and Marcondes, C. H. 2013. Identificación de indicios de descubrimientos científicos en artículos biomédicos mediante análisis de contenidos. Revista Española de Documentación Científica 36, 2, abril-junio, doi: http://dx.doi.org/10.3989/redc.2013.2.915.

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

[16] Shampay, J., Szostak, J.W. and Blackburn, E. H. 1984. DNA sequences of telomeres maintained in yeast. Nature 310, 154-157.

[17] 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.

[18] Murray-Rust, P. and Rzepa, H. S. 1999. Chemical Markup, XML and the World Wide Web. I: Basic principles, Journal of Chemical Information and Computer Science 39, 928--942.

[19] Hucka, M., Finney, A., Sauro, H. and Bolouri, H. 2003. System Biology Markup Language (SBML) Level 1: Structures and facilities for basic model definitions.

[20] 38. Murray-Rust, P. and Rzepa, H.S. STMML. 2002. A markup language for scientific, technical and medical publishing, Data Science Journal 1, (2), pp. 128-193.