Conceptual Exploration of Documents and Digital Libraries in the Biomedical Domain

Leyla Jael Garcia Castro[1], John Gómez[2], Alexander Garcia[2]

[1]biotea.ws Project
leylajael@gmail.com
[2]Florida State University, School of Library and Information Science,
Tallahassee, Florida, USA
alexgarciac@gmail.com

Abstract. In this demo we present our approach to the use of Semantic Web technology in scholarly communications; it entails the understanding of the research paper as an interface to the Web of Data. We are using the connectivity tissue provided by RDF technologies in order to facilitate semantic retrieval as well as to improve the user-experience when interacting with biomedical literature. Availability: http://biotea.idiginfo.org, http://199.102.237.69:8888/sparql/beta7/

Keywords: Semantics publications, linked data, biomedical visualization.

1 Introduction

As biomedical literature grows exponentially, so do the connections across documents. Research articles are deeply interconnected to one another and to resources in the web –e.g., biomedical databases. Such interconnectedness makes the paper an ideal interface to the Web of Data (WoD). Connections may be as evident as those due to bibliographic references; they may also be as complex as those due to similarities in topics, research questions, materials and methods, etc. Making use of such connectivity tissue requires a semantically processed dataset, delivering a self-descriptive document fully interoperable with the Web. We are working with RDF4PMC [1]; it comprises and makes available (i) a set of RDF files generated from the open access subset of PubMed Central (PMC) and enriched with semantic annotations, (ii) a Web Services API for querying the RDF data set, (iii) a SPARQL Protocol and RDF Query Language (SPARQL) endpoint containing a subset of the RDF files as a proof of concept, (iv) an article-centric prototype that acts as an interface to the WoD, and (v) an implemented transformation process from our RDF files to Bio2RDF (http://bio2rdf.org). As our model delivers self-descriptive documents we are able to focus on the user experience by building browsers that "*know*" the data types presented in the documents –biomedical entities. Also, as topics and sections are identified, search and retrieval can go beyond available functionality delivered by string-based engines. For instance, it is possible to retrieve articles including information related to cancer genes in sections whose titles contain "Methods" or "Re-

sults". Our RDF4PMC browser also makes it possible for users to examine the network of relationships across documents.

2 Browsing RDFized Biomedical Literature

RDF4PMC orchestrates ontologies such as DoCO (http://purl.org/spar/doco), BIBO (http://purl.org/ontology/bibo), DC (http://dublincore.org), and FOAF (http://xmlns.com/foaf/0.1) to model the metadata and content of the article. Meaningful terms in the content are represented as annotations modeled with the Annotation Ontology (AO) [2]. The resulting dataset provides a semantically linked version for PMC articles including metadata, references, and content, as well as enriched content in the form of annotations linked to biological entities –proteins, genes, drugs, diseases, among others. Our RDF4PMC browser makes it possible for users to search for a human gene, and to retrieve related papers; content as well as relevant terms are available –i.e., pieces of text associated to biological entities from specialized vocabularies. The retrieval also includes graphical tools that vary depending on the nature of the selected entity. In this way, for instance, sequences and 3D structures are displayed for proteins while chemical structures for chemical compounds. We are using a layer-based architecture as presented in Fig.1. The Presentation layer consists on a web-based Search & Retrieval interface; besides regular HTML, it also uses JavaScript technologies, particularly JQuery (http://jquery.com) and BioJS (http://code.google.com/p/biojs). Once the text to be searched for has been typed in the Presentation layer, the Communication layer is used to retrieve the required information. Once the information related to the searched text is retrieved, the Presentation layer is in charge of organizing this data. More search options will be powered by ontology mapping and ontology indexes. In this early version of our browser, users initiate the search by providing the name of a human gene. From the gene name, the corresponding protein accession is retrieved from the GeneWiki RDF; GeneWiki [3] is a Wikipedia project comprising about 10000 pages on human genes, including mappings to proteins, diseases, amongst others.

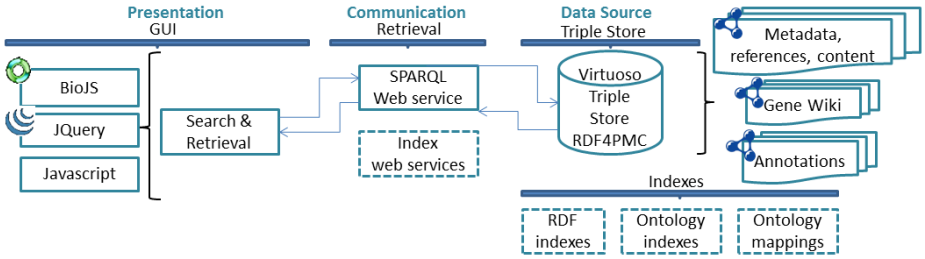

Fig.1. Architecture for RDF4PMC browser; data flow is shown by arrows between layers. Current components are shown in continuous-line, future components are in dashed-lines.

As presented in Fig.2 a), articles annotated with the protein accession corresponding to the insulin human gene are retrieved and organized in a list; results are alphabetically ordered. Metadata –title, authors, and abstract, as well as links – GeneWiki,

PubMed, PMC, and DOI, are presented. Whenever possible, links to identifiers.org (http://identifiers.org) –a resolvable persistent system for biological related information, and Bio2RDF are also provided. Additionally, a cloud of tags is displayed; this cloud contains the biological relevant terms identified in the article. The weight of each term in the cloud depends on the number of biological entities associated. Whenever a term in the cloud is selected, the vocabularies and subsequently the biological entities are displayed; different colors are used for different vocabularies. The interactive zone, Fig.2 b), changes depending on the selection in the cloud: (i) whenever a term is selected, paragraphs containing that term are displayed; a simple navigation bar allows user to move from one paragraph to the other; similarly (ii) when a biological entity is selected, relevant information is displayed, *e.g.*, sequences and 3D structures for proteins, structures for chemicals, or images for species.

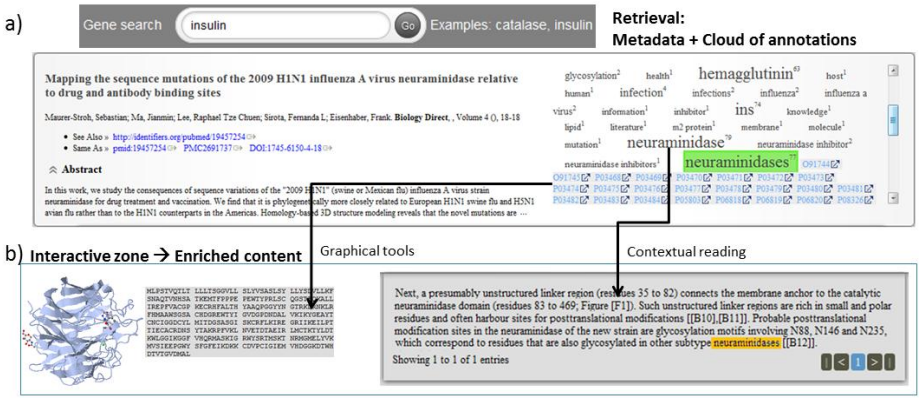

Fig.2. a) Search and retrieval based on human gene names. b) Enriched content based on annotations is displayed in the interactive zone.

A term-based browsing over the entire dataset is also possible; it facilitates navigation and filtering depending on terms contained in the articles. As articles are connected by shared terms, they are naturally arranged in a graph where articles are represented as nodes while terms are represented as arcs. Whenever two articles share a link, there will be an arc connecting them; the number of biological entities associated to a term will determine the weight of the arc. The term-browsing graph is created from a seed term defined by the user, for instance *catalase*. Once the seed term has been provided, the graph is generated with all articles containing that term; arcs are added depending on the other shared terms between any pair of articles. As documents and terms are retrieved, the graph is reorganized following a force-based algorithm. Fig. 3 shows the term-browsing graph for the term *catalase*; it also illustrates some features facilitating the navigation: mouse over events on nodes and arcs as well as filtering options. For nodes, mouse-over activates a close icon on the right upper corner so the article can be removed from the graph; in the figure it is possible to observe the close icon for the article with title "*Localization of the Carnation Italian ringspot virus* […]". For arcs, mouse-over displays the actual term, in the figure it is possible to observe the term *PROTEASE*. Terms can also be filtered out; as multiple terms can be shared amongst articles, an arc-based filtering feature has been defined:

depending on the weight, *i.e.*, number of biological entities, terms can be excluded from the graph; in the figure the minimum weight has been set to 30.

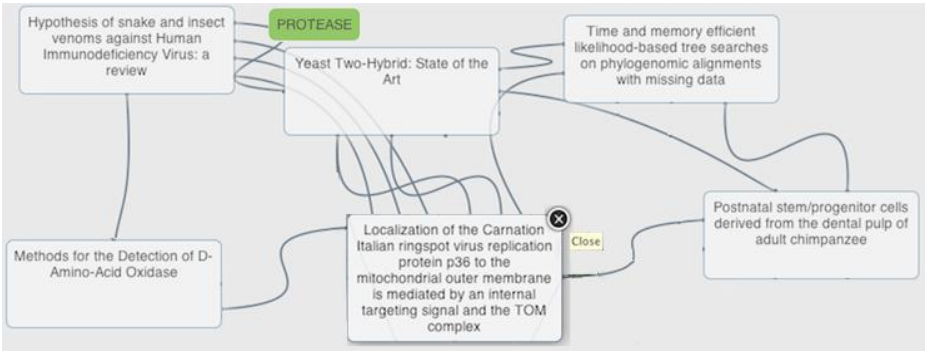

Fig.3. Partial connectivity graphs for the term catalase –some nodes are not displayed. Additional terms with more than 30 biological entities are displayed as arcs.

3 Conclusions and future work

The components for the RDF4PMC browser that have been described in this paper illustrate how the scientific article could be understood as an interface to the WoD. The RDFication delivers self-describing content that allows the implementation of semantic search, exploration and retrieval mechanisms throughout the digital library as well as a new reading experience when interacting with the document. The former is achieved by making extensive use of the interlinked nature of the RDF dataset; the latter is the consequence of using specialized visualization and manipulation gadgets that make it possible for the reader to browse the document focusing on those aspects he/she considers relevant for the research at hand. Exploring the network of interconnected documents facilitate the formulation of more precise queries; being able to post process queries based on information found in the documents deliver a new pivot from where to build queries. In the near future we are planning to explore the use of ontology mappings as well as ontology expansion; users should not need to know the exact name of a gene, and should be able to search for related information, e.g., protein name, diseases, drugs, etc. We will focus on knowledge that could be easily inferred as a consequence of the interrelated nature of the semantic model.

References

1. García A, García Castro LJ, McLaughlin C, Flager S: **RDFising PubMed Central**. In: *Bioontologies: 2012; Long Beach, CA, USA*; 2012.
2. Ciccarese P, Ocana M, Garcia Castro L, Das S, Clark T: **An open annotation ontology for science on web 3.0**. *Journal of Biomedical Semantics* 2011, **2**(Suppl 2):S4.
3. Huss JW, Lindenbaum P, Martone M, Roberts D, Pizarro A, Valafar F, Hogenesch JB, Su AI: **The Gene Wiki: community intelligence applied to human gene annotation**. *Nucleic Acids Research*, **38**(suppl 1):D633-D639.