Biotea

Alexander Garcia[1,*], Leyla Jael García Castro[2] and Casey McLaughlin[1]

[1]Institute for Digital Information and Scientific Communication, College of Communication and Information, Florida State University, Tallahassee, Florida, 32306-2651, USA.

[2]Temporal Knowledge Bases Group, Department of Computer Languages and Systems, Universitat Jaumé I, Castello de la Plana, Valencia, 12071, Spain.

ABSTRACT

Motivation: Scientific information is usually locked up in discrete documents that are not always interconnected or machine-readable. The connectivity tissue provided by RDF technology has not yet been widely used to support the generation of self-describing, machine-readable documents. In this paper we present our approach to machine-processable documents. We have semantically modeled and enriched the full-text open-access subset of PubMed Central. Our model delivers a highly interconnected and semantic dataset.

Introduction

In spite of technological advances, scientific publications remain poorly connected to each other as well as to external resources. Furthermore, most of the information remains locked up in discrete documents without machine-processable content. Such interconnectedness and structuring would facilitate interoperability across documents as well as between publications and online resources. Scholarly data and documents are of most value when they are interconnected rather than independent.

Methods

We use BIBO [1], DCMI Terms [2], and the Provenance Ontology (PROV-O) [3] to model the bibliographic metadata. BIBO provides classes and properties to represent citations and bibliographic references. BIBO can be used to model documents and citations in RDF or to classify documents within a hierarchy. Dublin Core (DC) [4] offers a domain-independent vocabulary to represent metadata; such vocabulary aims to facilitate cross-resource exploration. In order to identify biological terms, we use two entity recognition tools: Whatizit [5] and the NCBO Annotator [6]. Both tools are based on exact string matching and pre-defined dictionaries. By doing so, relevant biological identifiers such as UniProt accessions and ChEBI and GO identifiers are added. We are working with more than 20 biomedical ontologies.

The workflow that we followed to generate the RDF files for PubMed Central (PMC) articles is illustrated in Fig. 1. The main input for our process is the XML offered by PMC for open- access articles. We are also using available vocabularies to represent the metadata as well as the content in RDF; such vocabularies have been mapped to Java classes by using the RDFReactor. The article itself is modeled as bibo:Document; whenever it is possible, a more precise class is also added, e.g., bibo:AcademicArticle for research articles. Publisher metadata is modeled using BIBO, including publisher name, the International Standard Serial Number (ISSN), volume, issue, and starting and ending pages. Authors are modeled as a bibo:authorList, where each member is a foaf:Person. Abstract and sections are modeled as a doco:Section with a cnt:chars containing the actual text with formatting omitted. Well-known identifiers such as PubMed ids and DOIs are included in the output. In this way it is possible to track the original source of the article; the same principle is also applied to the references. In order to identify biological terms within the RDFized article, it is processed with Whatizit and the NCBO Annotator. Those terms are modeled as semantic annotations, i.e., annotations associated to ontological concepts such as proteins, components, drugs, diseases, and medical terms. Whenever it is possible, we also link to entities in Bio2RDF [7] and identifiers.org as well as to relevant web pages.

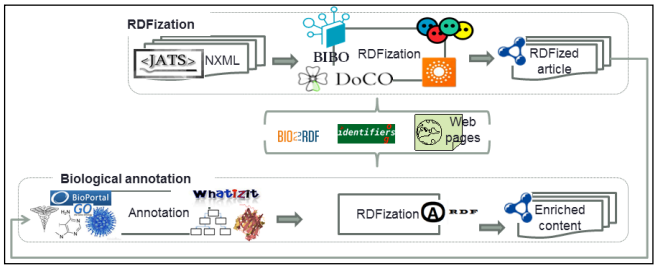

Fig. 1. Biotea workflow

Results

We have semantically processed the full-text, open-access subset of PubMed Central. Our RDF model and resulting dataset make extensive use of existing ontologies and semantic enrichment services. We expose our model, services, prototype, and datasets at http://biotea.idiginfo.org/. The semantic processing of biomedical literature presented in this paper embeds documents within the Web of Data and facilitates the execution of concept-based queries against the entire digital library. Our approach delivers a flexible and adaptable set of tools for metadata enrichment and semantic processing of biomedical documents. Our model delivers a semantically rich and highly interconnected dataset with self-describing content so that software can make effective use of it.

References

1. D'Arcus, B. and F. Giasson. *Bibliographic Ontology Specification*. Bibliographic Ontology 2009 [cited 2012; Available from: http://bibliontology.com/specification.
2. http://dublincore.org/. *DCMI Metadata Terms*. DCMI Metadata Terms 2012 [cited 2012; Available from: http://dublincore.org/documents/dcmi-terms/.
3. Belhajjame, K., et al. *PROV-O: The PROV Ontology*. W3C Recommendation 2013 [cited 2012; Available from: http://www.w3.org/TR/prov-o/.
4. http://dublincore.org/. *Dublin Core Metadata Initiative* [cited 2012; Available from: http://dublincore.org/.
5. Rebholz-Schuhmann, D., et al., *Text processing through Web Services: Calling Whatizit*. Bioinformatics, 2007. **24**(2).
6. Clement, J., S. Nigam, and M.A. Musen. *The Open Biomedical Annotator*. in *AMIA Summit on Translational Bioinformatics*. 2009. San Fransisco.
7. Callahan A, et al. *Improved dataset coverage and interoperability with Bio2RDF Release 2*. in *Semantic Web Applications and Tools for Life Sciences*. 2012. Paris, France.

*To whom correspondence should be addressed.