Semantically: A Framework for Structured Biomedical Content Authoring and Publishing

Steve Fonin Mbouadeu, Fazel Keshtkar, and Syed Ahmad Chan Bukhari*

Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. John's University 8000 Utopia Parkway Jamaica NY US 11439-9000
{steve.mbouadeu19, keshtkaf, bukahris}@stjohns.edu

Abstract. We demonstrate *Semantically*, an open-source framework that enables biomedical researchers and authors at different expertise levels to write and publish biomedical semantic content. We showcase the workflow of *Semantically* with a real-world use case of biomedical semantic content authoring using biomedical ontologies. Additionally, we exhibit Web publishing steps of ontology-enriched content by expanding Schema.org to enhance biomedical content search. *Semantically* is available at: semantically.bukharilab.org

1 Introduction

Due to unprecedented growth in biomedical research and clinical practices, a huge volume of biomedical content such as research papers, clinical notes, biomedical reports, have been produced in recent years. According to the 2018 online molecular biology database collection, there are 1737 large-scale biological databases and over 30 million citations for biomedical literature available on PubMed with a tenfold increase annually [3]. Efficient practices for accessing biomedical publications are vital to allowing a timely transfer of information from the scientific research community to peer investigators and other healthcare practitioners. This explosive growth in the biomedical domain has introduced several access-level challenges for researchers and practitioners. Shah et al. define researchers struggle as ''*a researcher studying the allelic variations in a gene would want to know all the pathways that are affected by that gene, the drugs whose effects could be modulated by the allelic variations in the gene, and any disease that could be caused by the gene and the clinical trials that have studied drugs or diseases related to that gene* [5]. Despite this valuable information being available as content on the web, it still remains opaque to information retrieval and knowledge extraction search engines because of the missing machine-interpretable metadata (semantic annotations). Search engines require metadata to properly index contents in a context-aware fashion for the precise search of

biomedical literature and to foster secondary activities such as automatic integration for meta-analysis [2][4]. Incorporating machine-interpretable semantic annotations at the pre-publication stage of biomedical content and preserving them during online publishing is indeed desirable and will be a great value addition towards the broader semantic web vision [9]. However, both these processes are complex and require deep technical and/or domain knowledge. Therefore, a state-of-the-art freely accessible biomedical semantic content authoring and publishing framework would be a game-changer.

2 System Overview

Semantically is a web application that uses React JS and LAMP (Linux, Apache, MySQL, and PHP) architecture. The system is optimized by performing various processing on the client-side, enabled by React JS. As outlined in Fig. 1, the Semantically interacts with two external entities: BioPortal [1] and Schema.org [8].
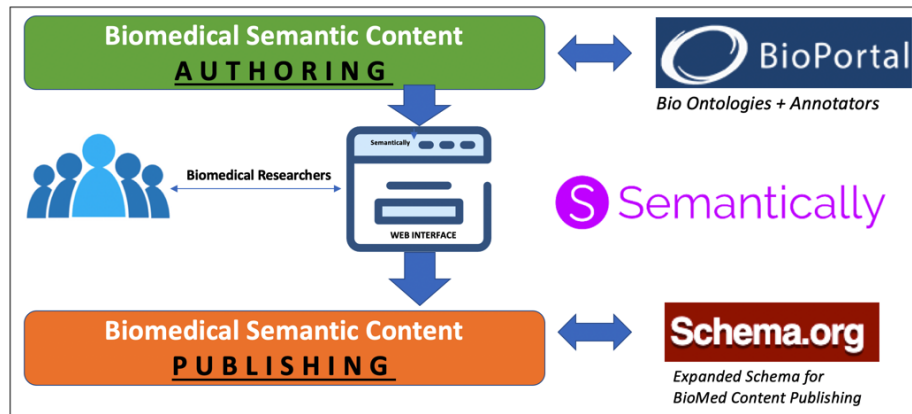

Fig. 1. A flow diagram of the Semantically framework.

2.1 Annotating Content

The preparation process before annotating includes basic NLP tasks such as biomedical entity identification, disambiguation, and extraction. Bioportal serves as the ontology repository and as a general-purpose biomedical annotator through its recommender API. Having general-purpose annotations at the beginning is crucial for the system to achieve acceptable real-time content annotation speeds. Afterward, the general annotations are further tailored for the content using Bioportal's context-aware recommendations. This specific combination and order were chosen to optimize our system for accuracy and speed, which we found

equally important. An additional advantage of utilizing general-purpose annotations first in our system was to maximize the field of content and terminologies it can recognize and therefore annotate. Once the annotations have been processed, they are clustered by terminology. During annotation selection, candidate annotations are organized by both relevancy based on the context and a frequency index compiled by Bioportal. Once the annotation candidates have been retrieved, the user is free to choose between annotations candidates for the ones that best aligns with the definitions they intended from the annotation management panel (Fig. 2).
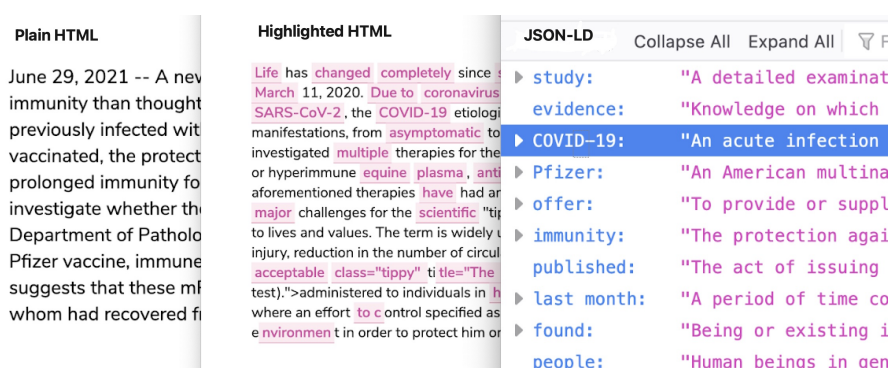
Fig. 2. (1) Plain HTML export view, (2) highlighted HTML export view, (3) JSON-LD export view.

2.2 Publishing Content

After annotating the content, our system provides various publishing format options to choose from, all of which utilize the schema.org structure. Our approach explicitly takes full advantage of the medical entity meta tags from Schema.org [8]. This structure is recognized and endorsed by widely used search engines such as Google and Yahoo for the search engine optimization (SEO). Semantically preserves semantic annotations made on the biomedical contents and embeds them into the HTML file's head tag as metadata following the schema.org structure. Search engines utilize the embedded semantic tags for context-aware indexing and retrieval of precise information. In addition, Semantically also offers a JSON-LD [6] export format that is based on the schema.org entity structure. JSON-LD is a standardized format for semantic annotations, allowing authors to convert it into other formats of their choice.

3 Demonstration Scenario: Annotating and Publishing a Biomedical Research Article

The demonstration session will consist of a run-through of semantically using a biomedical research paper. We will take a biomedical article text from Pubmed.org and will paste it to the Semantically editor. The user interface and controls will be explained to the conference attendees. Attendee volunteers may access the Semantically framework simultaneously and follow the steps below to annotate a biomedical text to observe the system in action.

– Copy and Paste any biomedical text from Pubmed.org to the Semantically editor. We will demonstrate with a Covid-19 paper [7]
– Choose the annotation mode **All** to select ontology list from NCBO Bioportal repository.
– Click on **Annotate**.
– Once the annotations are retrieved, users can edit annotations by clicking on each context in the text editor and clicking on the **edit** button.
  • *Completion of annotation retrieval is indicated by the loading icon terminating and the highlights appearing on the content.*
– Publish the content as plain HTML as shown in Figure 2 ready to be hosted as ontology-embedded Web page.
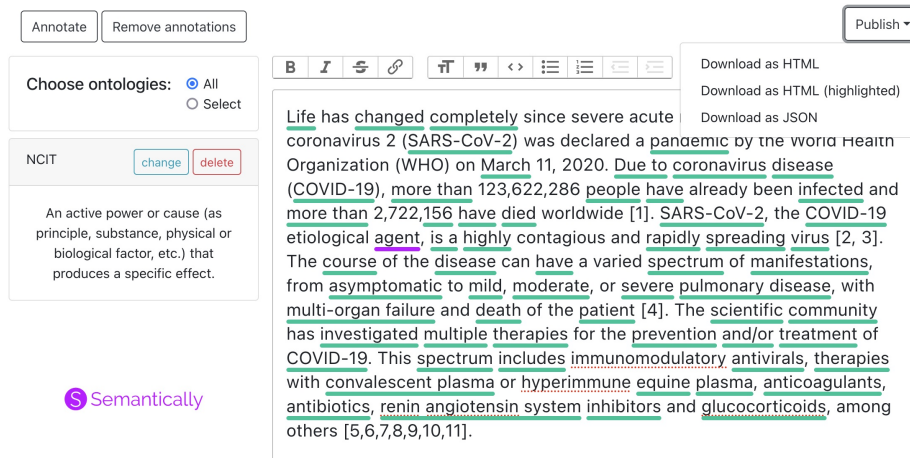
Fig. 3. A screenshot of Semantically exhibiting the semantic content authoring process of a Biomedical Article.

4 Conclusion and Future work

Semantic rich content authoring and sharing is the least explored aspect of the semantic content life-cycle and this research develops and advances state-of-the-art in semantic web research and systems. The proposed framework democratizes the process of semantic content authoring currently available to sophisticated users by extending cutting-edge semantic web infrastructure to research and education communities at large. The amount of accuracy that can be attained using machine learning techniques is limited as only the author(s) and domain experts truly know the intended meaning of the content. To address this shortcoming, the next phase of Semantically involves implementing a social-technical solution to complement annotation selection. Our system still remains in a prototype state where key features for convenience such as importing documents, persisting documents throughout sessions, and connectivity to popular external utilities are still missing. We look forward to add such features in future as Semantically moves towards production.

References

1. Ontology for general medical science — NCBO BioPortal. `https://bioportal.bioontology.org/ontologies/OGMS`, accessed: 2017-9-6
2. Bukhari, S.A.C.: Semantic Enrichment and Similarity Approximation for Biomedical Sequence Images. Ph.D. thesis, University of New Brunswick (Canada) (2017)
3. Daniel J Rigden, X.M.F.: The 2018 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res. **46**(Database issue), D1 (Jan 2018)
4. Jovanović, J., Bagheri, E.: Semantic annotation in biomedicine: the current landscape. J. Biomed. Semantics **8**(1), 44 (Sep 2017)
5. Shah, N.H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A.P., Musen, M.A.: Comparison of concept recognizers for building the open biomedical annotator (2009)
6. Solbrig, H.R., Solbrig, H.R., Prud'hommeaux, E., Booth, D., Endle, C.M., Stone, D.J., Jiang, G.: Exploring JSON-LD as an executable definition of FHIR RDF to enable. AMIA Annu. Symp. Proc. **2020**, 1140–1149 (2020)
7. Vallejos, J., Zoni, M.G.: Ivermectin to prevent hospitalizations in patients with COVID-19 (IVERCOR-COVID19): a structured summary of a study protocol for a randomized controlled trial. Trials **21**(1), 965 (Nov 2020)
8. Vassiliadis, P.: Schema evolution and gravitation to rigidity: A tale of calmness in the lives of structured data (2017)
9. Warren, P., Davies, J.: The semantic web– from vision to reality