Towards Structured Biomedical Content Authoring and Publishing

Conference Paper · January 2022

DOI: 10.1109/ICSC52841.2022.00036

6 authors, including:

Steve Mbouadeu
St. John's University
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Asim Abbas
St. John's University
15 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Fazel Keshtkar
St. John's University
48 PUBLICATIONS **376** CITATIONS
SEE PROFILE

Syed Ahmad Chan Bukhari
St. John's University
92 PUBLICATIONS **1,293** CITATIONS
SEE PROFILE

Towards Structured Biomedical Content Authoring and Publishing

Steve Fonin Mbouadeu
St. John's University, USA
steve.mbouadeu19@stjohns.edu

Asim Abbas
St. John's University, USA
abbasa@stjohns.edu

Faizan Ahmed
St. John's University, USA
faizan.ahmed18@my.stjohns.edu

Fazel Keshtkar
St. John's University, USA
keshtkaf@stjohns.edu

Joan DeBello
St. John's University, USA
debelloj@stjohns.edu

Syed Ahmad Chan Bukhari*
St. John's University, USA
bukahris@stjohns.edu

2022 IEEE 16th International Conference on Semantic Computing (ICSC) | 978-1-6654-3418-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSC52841.2022.00036

***Abstract*—Significant barriers exist in achieving fast and accurate access to online biomedical content because of the proliferation of unstructured biomedical information. Accompanying semantic annotations with growing biomedical content is critical to enhancing search engines' context-aware indexing, improving search speeds and retrieval accuracy. We have developed "Semantically": a biomedical structured content authoring and publishing framework to enhance biomedical content FAIRness (Findability, Accessibility, Interoperability, and Reusability). Finding the appropriate semantic vocabulary to annotate biomedical content is time-consuming and technically challenging. "Semantically" automates and streamlines this process for users by recommending highly accurate annotations from an array of biomedical ontologies. Similarly, preserving content-level semantics at the content publishing stage to foster semantic search remains a critical research challenge. "Semantically" addresses this obstacle by extending schema.org, a community-agreed and research engine endorsed guideline for publishing structured content on the web. In future works, we aim to improve the biomedical content annotation process through a socio-technical approach by enabling a collaborative annotation scheme. The demo of the system is accessible at: https://gosemantically.com/**

***Index Terms*—Structured data, Biomedical Semantics, Structured data publishing, FAIR Biomedical Data, Biomedical Content Authoring**

I. Introduction

A large number of unstructured biomedical content has been produced over recent years from growing scientific research [1]. However, the lack of machine-interpretable metadata associated with online biomedical content makes it inaccessible through commonly used search engines, e.g., Google. Search engines rely on embedded metadata to efficiently index content to perform fast and accurate queries and to support secondary activities such as meta-analysis and automated integration [2]. Therefore, incorporating interoperable semantic annotations and maintaining them during their dissemination and publishing is critical to achieving FAIRness in the biomedical domain [3]. Several biomedical semantic annotators such as NOBLE Coder [4], NCBO Annotator [1], and Open Biomedical Annotator [5] have been introduced by researchers to incorporate biomedical semantics into biomedical content automatically and semi-automatically. However, most of the available biomedical annotators failed to balance between speed and accuracy. The challenge is to select the correct ontological vocabularies from the several available and then to associate them with unstructured biomedical contents in the shortest time possible [6]. This paper introduces Semantically, a web framework designed for biomedical researchers and content creators of all experience levels for the authoring and publishing of semantic biomedical content. Semantically facilitates the process for users by selecting the right semantic vocabularies, finding and associating them to the unstructured biomedical contents. The embedded machine-interpretable metadata is preserved while exporting the content for the web.

This work is supported by the National Science Foundation grant ID: 2101350

II. Proposed Methodology

The Semantically framework was designed for users ranging from bench scientists and medical doctors to casual users simply get involved in medical journalism. Initially, users have the option to either import pre-existing content or start typing directly in the Semantically text editor (Fig.1). Afterward, they are given a few annotation options to select from depending on their level of experience and familiarity with certain ontologies. Users without a technical background may easily navigate a simplified interface while more sophisticated users may utilize advanced options to take further control of the semantic annotation process. The Semantically editor performs the semantic breakdown process as it annotates textual content. The process is comprised of 1) Ontology identification which is a collection of concepts and semantic relationships among concepts, 2) Biomedical concept recognition utilizing BioPortal's Annotator API [1] and 3) semantic information extraction that presents contextual similarity and relationship among concepts. (Fig.1(c)). The output of this stage is ontology based structured content. Terminologies for which Semantically was successfully able to match to one or more ontologies are underlined (Fig.2) following the completion. Users can then click on the terms to view their attached semantic description. If necessary, users may further optimize the semantic description

978-1-6654-3418-8/22/$31.00 ©2022 IEEE
DOI 10.1109/ICSC52841.2022.00036

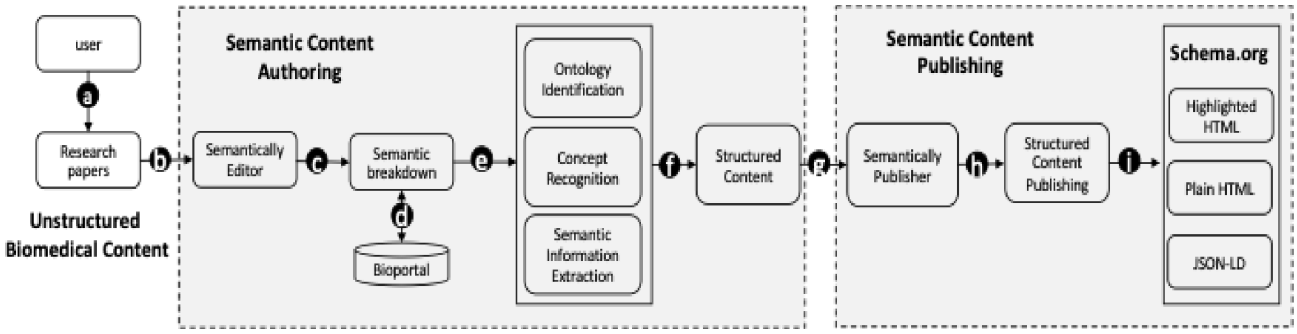

Fig. 1: Unstructured biomedical content transformation into structured content methodology workflow

to the one they prefer, changing the linked ontology or concept in the process (Fig.2).

The schema.org [7] is a collaborative community-defined set of vocabularies designed to create structured web content from a multitude of domains. It's endorsed by Google, Yahoo, Yandex, and Bing with the goal to improve the indexing of web pages. We utilize the schema.org provided high-level structural tags such as MedicalScholarlyArticle and MedicalEntity and couple them up with the content-level semantic that generated during the content authoring process [7]. Following the annotation of the content, our system offers a variety of publication format options, all of which employ the schema.org [7] framework such as JSON-LD, metadata embedded plain HTML, and highlighted HTML Formats (Fig.1).

Semantically is a web-based application that uses React JS and LAMP (Linux, Apache, MySQL, and PHP) architecture. The system communicates with two external entities, as shown in (Fig.1): Schema.org [7] and BioPortal [1]. Semantically and BioPortal are connected reciprocally and uses Bioportal as the knowledge base for a large collection of ontologies to suggest appropriate annotations to the author. In response, Semantically keeps track of author annotation choices and shares them with BioPortal to enhance its recommendation algorithms. It also expands the Schema.org standard for structured online content in a similar way by offering structural level information. By enhancing findability and interoperability, these additional features contribute to the FAIR [3] data initiative.

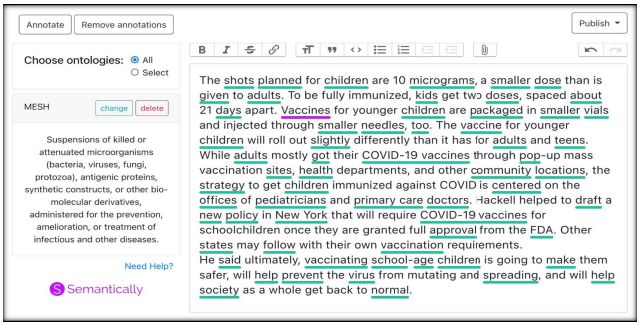

Fig. 2: A screenshot of Semantically exhibiting the semantic content authoring process of a Biomedical Article.

III. Conclusion

Context-aware authoring and sharing at the pre-publication stage is the least explored aspect of the semantic content life-cycle. This research advances state-of-the-art biomedical semantic research and systems, enabling various biomedical users to author and publish context-aware content with no prior technical skills. Our future aim is to extend the current Semantically infrastructure to address the advanced annotation challenges by introducing a social-technical model. Furthermore, our system remains in a prototype state where key features for convenience such as importing web documents through URL, direct publishing from semantically environment to blogs are still missing. We look forward to adding such features in the future.

Acknowledgment

This work is supported by the National Science Foundation grant ID: 2101350.

References

[1] C. Jonquet and N. Shah and C. Youn, C. : "NCBO annotator: semantic annotation of biomedical data," in International Semantic Web Conference, Poster and Demo session, 2009, vol. 110, [Online].

[2] Mbouadeu, S.F. and Keshtkar, F. and Bukhari, S.A.C.:Semantically: A Framework for Structured Biomedical Content Authoring and Publishing.

[3] Wilkinson, M.D. and Dumontier, M. and Aalbersberg:The FAIR guiding principles for scientific data management and stewardship. Sci Data 3, 160018 (Mar 2016).

[4] E. Tseytlin, K. Mitchell, E. Legowski, J. Corrigan, G. Chavan, and R. S. Jacobson, "NOBLE – Flexible concept recognition for large-scale biomedical natural language processing," BMC Bioinformatics, vol. 17, no. 1. 2016, doi: 10.1186/s12859-015-0871-y.

[5] Jonquet, Clement, Nigam H. Shah, and Mark A. Musen. "The open biomedical annotator." Summit on translational bioinformatics 2009 (2009): 56.[online].

[6] J. Cuzzola, J. Jovanović, and E. Bagheri, "RysannMD: A biomedical semantic annotator balancing speed and accuracy," J. Biomed. Inform., vol. 71, pp. 91–109, Jul. 2017.

[7] R. V. Guha, D. Brickley, and S. MacBeth, "Schema.org: Evolution of Structured Data on the Web," Queue, vol. 13, no. 9. pp. 10–37, 2015, doi: 10.1145/2857274.2857276.