Biomedical Scholarly Article Editing and Sharing using Holistic Semantic Uplifting Approach

Asim Abbas, Steve Fonin Mbouadeu, Fazel Keshtkar, Joan DeBello, Syed Ahmad Chan Bukhari*
{abbasa, steve.mbouadeu19, keshtkaf, debelloj, bukharis}@stjohns.edu
Division of CSMS, St. John's University, Queens, NY 11439, USA

Abstract

Efficient practices to provide access to biomedical publications facilitate the timely transfer of information from the scientific research community to peer investigators and other healthcare practitioners. At present, the portable document format (PDF) is one of the dominating formats to share scientific knowledge offline. Additionally, some HTML-based formats have been introduced to share scientific content online. Online Search engines, e.g., GoogleScholar, require machine-interpretable metadata to correctly index items in a context-aware manner for accurate biomedical literature searches. We have developed a lightweight technical infrastructure (goSemantically) and miniaturized that as Google Docs add-ons that helps authors to add machine-interpretable metadata at the content and structural levels while authoring biomedical content. The infrastructure uses the NCBO Bioportal resources to annotate the biomedical content with appropriate semantic vocabularies. It further utilizes the Schema.org meta tags and provides an intuitive interface for users to associate the semantics tags at the document level. Additionally, our infrastructure supports users in exporting their content in various online interoperable formats preserving the embedded semantics. As a result, the biomedical metadata content would easily be indexed by search engines, making them more favorable for semantic intelligence searches.

Keywords: Structured data, Biomedical Semantics, Structured data publishing, FAIR Biomedical Data, Biomedical Content Authoring

Introduction and Background Work

Efficient practices for accessing biomedical publications are crucial to the timely transfer of information from the scientific research community to peer investigators and other healthcare practitioners. The portable document format (PDF) is one of the dominant formats for offline sharing of scientific knowledge [Brady 2015]. Some HTML-based structures have been introduced recently for online scientific publications [Spinaci 2017].

In the following, we document the available tools that were developed recently to edit scientific articles leveraging semantic technologies. Dokieli [Capadisli 2017] is a decentralized browser-based authoring and annotation tool. It is developed for creating HTML+RDF annotations notifications in scientific articles. A dokieli's generated article appears as an HTML page where users can include contextual information in Turtle, JSON-LD, or TriG formats that are commonly used in nano-publication, see Table.1. Another framework similar to dokieli called Research Articles in Simplified HTML(RASH) [Peroni 2017] introduced writing HTML-based scholarly articles. The evaluation study revealed certain challenges with RASH adoption in particular and HTML adoption in general for non-technical users. Additional functionalities are required to make it widely acceptable, e.g., enabling additional conversions from/to existing formats such as OpenXML, Table. 1. Authorea is a tool that allows scholars to write, cite, collaborate, host data, and publish their work online. It generates output in four distinct formats such as PDF, LaTeX, DOCX, and a zipped bundle with many HTML pages[1], Table.1. The Scholar-Markdown is a framework for producing academic articles in a lightweight HTML syntax that is automatically translated into HTML+RDFa and printed to PDF into a standard scientific template via browser Table.1. It gives syntactic sugar to cite articles, write math equations, and more [Lin 2015]. FidusWriter[2] is another Web-based program that uses a wordprocessor-like interface to create HTML academic publications. While the exact format is not mentioned, it does enable the translation of HTML documents created within the application into two alternative formats: EPUB and LaTeX.(alongside HTML) Table.1.

The unavailability of the technological infrastructure required to add machine-interpretable metadata into growing publications makes them inaccessible to literature search engines like Google Scholar. Search engines require metadata to correctly index items in a context-aware manner for accurate biomedical literature searches. Incorporating machine-interpretable semantic metadata at the pre-publication stage (while writing) of biomedical content and preserving them during online publishing is always desirable and will be a great value addition to the broader semantic web vision [Mbouadeu 2022]. However, these complex processes require deep technical and/or domain knowledge. Biomedical annotators use biomedical ontologies available at public

[1] www.authora.com
[2] www.fiduswriter.org

Table 1: A comparison of goSemantically with existing HTML-oriented formats for scholarly papers according to five distinct features.

Format	Syntax (HTML)	Semantic Annotation (RDF,JSON-LD,RDFa XML,Turtle)	WYSIWYG editor	Conversion From (DOCX,PDF, ODT,LaTeX)	Conversion To (DOCX,PDF,EPUB ODT,LaTeX,HTML)
dokieli	✓	✓	✓	x	✓
RASH)	✓	✓	✓	✓	✓
Authorea	✓	x	✓	✓	✓
ScholarMarkdown	✓	x	✓	x	✓
Fiduswriter	✓	x	✓	x	✓
goSemantically	✓	✓	✓	✓	✓

repositories such as BioPortal [Whetzel 2011] and UMLS [Abbas 2019] to transform the unstructured content into a structured format by associating the appropriate ontology concepts with content. We have developed a lightweight add-ons for Google Docs called "goSemantically" employing a holistic approach that helps researchers add semantic metadata at the content and structural levels with no prior technical knowledge by utilizing biomedical ontologies and schema.org metadata tags. Furthermore, the proposed infrastructure aid the users to download semantically enriched content in distinct interoperable formats for decentralized hosting and sharing. The embedded semantic metadata makes the uplifted biomedical content favorable for search engines to index them for semantic search intelligently.

Proposed Methodology

The proposed "goSemantically" lightweight GoogleDocs add-ons assists biomedical practitioners and researchers to uplift unstructured biomedical content by automatically suggesting the appropriate ontology vocabularies for semantic annotation at content and structural levels. Subsequently, an array of Schema.org compliant web publishing options are provided to the users to export their semantically enhanced content in various interoperable formats for hosting and sharing in a decentralized fashion. Uplifting syntactic content to semantic content will enhance the FAIRness of published research. Thus, it will help in the timely sharing and reusing scientific knowledge for new discoveries. The following section explains the content-level and structural-level semantic enrichment Figure. 1.

The Content-level Semantic Enrichment

A biomedical annotator is an essential component of the content-level semantic enrichment process [Abbas 2021]. These annotators use publicly available biomedical ontologies, such as Bioportal [Whetzel 2011] and UMLS [Abbas 2019]. However, the semantic annotation and enhancement process can not be easily automated and often requires expert curators. Additionally, there is a lack of an easy-to-use system that facilitates the semantic enrichment process. Therefore, we utilized an NCBO Bioportal [Jonquet 2009] web-service resources in the proposed application that processes the raw textual data and tags them with relevant biomedical ontology concepts.

The semantic enrichment process is comprised of (1) ontology identification and (2) semantic information extraction. We believe this level of information plays a vital role in the semantic enrichment process at the content level to enhance its discoverability and interactivity for both humans and machines.

1: An ontology is a collection of concepts and their semantic relationships among them. The biomedical text provided to "goSemantically" is mapped to an ontology repository, and individual terminologies are matched to ontologies based on the context. The "goSemantically" automates this process from the user's perspective, removing technical knowledge as a barrier toward operating the add-on effectively. We leveraged the Bioportal Ontology web service a repository containing over 729 ontologies in the biomedical domain, to identify the relevant ontologies for the given content [Jonquet 2009]. Ontology metadata can be accessed through their REST API and can be queried with various parameters, such as indicating exact matches of a term and providing a restricted ontology list. We believe that ontologies can improve the semantic enrichment process because it often considered as an acceptable source of semantics and interoperability in all artificial intelligent systems[Abbas 2021].

2: Semantic information extraction is the process of extracting insight and detailed information about biomedical terminology. This information assists in identifying the contextual similarity and relationship between concepts or biomedical terms. A concept can be categorized and classified into specific names or entities that help in concept annotation or labeling. From the semantic information, we extracted ontology ID, ontology name, and its semantic definition and display in the annotation panel of "goSemantically" for user interpretation and understanding.

The Structural-level Semantic Enrichment

We utilized the schema.org markup types arranged in a hierarchy in the proposed application. The commonly used types of schema.org are Creative Works, Event, Health and medical types, Organization and Person, etc. We utilized Health and Medical types tags in the proposed application, which is useful for publishers wishing to mark up health and medical content on the web. The Health and Medical types have further subtypes such as MedicalCondition, Drug, MedicalGuideline, MedicalWebPage, and MedicalScholarlyArticle. As We aim to contribute to the semantic uplifting of

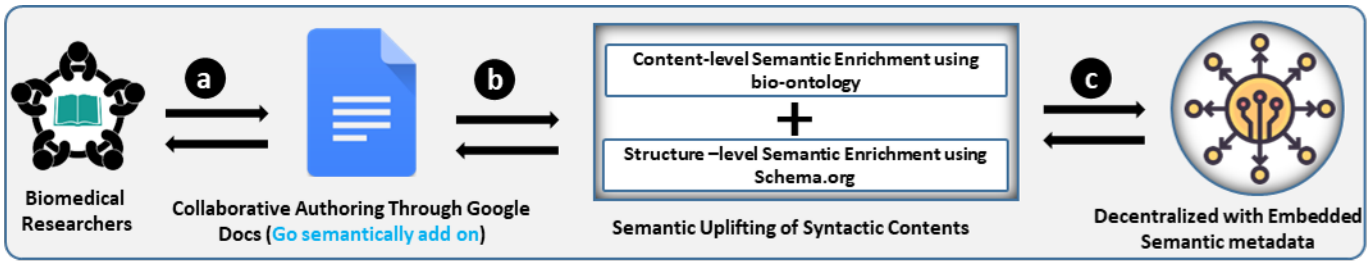

Figure 1: A workflow of the proposal Holistic Semantic Uplifting Scheme

the medical scholarly article by enhancing its discoverability, interactivity, openness, and (re) usability for both humans and machines, we chose the MedicalScholarlyArticle schema type further. The MedicalScholarlyArticle Schema type has different properties that give a detailed description of the medical content, such as "publicationType", "abstract", "publisher", etc.
In the "goSemantically" at the right side of the sidebar, contain buttons "Structural" as shown in the Figure.2(c) support for structural level semantic uplifting process. When a user click on "Structural" button the MedicalScholorlyArticle properties list will be appear in the drop-down menu, where user is allow to choose suitable properties for article based on his experience and knowledge. Subsequently a text field shown for each chosen property from drop-down list. Each text field contains a property placeholder that indicates the user to input relevant content Figure.2(c).User is not allowed to leave any text field empty, or user can go back and uncheck the property checkbox unwanted.After successful filling of the selected properties text field, finally, the biomedical contents with MedicalScholorlyArticle properties are generated in JSON-LD format, the language code for HTML pages or search engines. These properties will be embedded directly into the article <head> tag in JSON-LD format, the language code for HTML pages or search engines for searching and indexing. The notation used for "@context" and "@type" attributes is to specify the vocabulary in schema.org Figure.2(e,f).

Demo and Decentralized Publishing

A run-through of "goSemantically" utilizing a biomedical research paper will be part of the demonstration session. The conference attendees will be given an overview of the user interface and controls. Attendee volunteers can simultaneously use the "goSemantically" browser add-on and experience the application in action by following the instructions below. The "goSemantically" interface consists of two buttons, (a) Content Level and (b) Structural Level, as shown in the Figure.2(b,c).

Content Level

1. Copy and paste any biomedical text from Pubmed.org to the Google Doc editor. We will demonstrate this with a cancer disease article [Hausman 2019] Figure.2(a).
2. Click on Annotate after choosing "Content Level" button Figure.2(c).
3. After the annotations are retrieved, users can edit annotations by clicking on each context in the Google Doc editor. For example, by clicking on the "Not satisfied with this annotation" button, the user can select an appropriate ontology from a list. The "Removing Annotations" button removes the suggested annotation for the context.
4. When the user clicks on annotated term, its preflabel, recommended ontology, definition, and a tree browser of ontologies is appeared in the panel. The completion of annotation retrieval is indicated by the loading icon terminating and the highlights appearing on the content Figure.2(b).

Structural Level

1. Click on the "Structural Level" button in the Figure.2(d). A drop-down list of MedicalScholarlyArticle schema types will appear.The user is allowed to select the appropriate properties from the drop-down list.
2. A text field appears contain a MedicalScholorlyArticle type property placeholder that indicates the user to input relevant content.
3. Finally, the user clicks on the "Generate JSON" button to directly couple and embed the MedicalEntity and MedicalScholarlyArticle meta-tags in the heading of article in JSON-LD format.
4. A download button is enabled for the user to download the content and publish it in plain HTML and JSON-LD format, as shown in the Figure. 2(e,f). Subsequently, the content is ready to be hosted online.

Decentralized Publication

In a decentralized environment, applications run independently of a centralized interconnected system. While the web was supposed to be a decentralized environment, individual authors currently lack the ability to author and publish documents in a genuinely decentralized manner and engage in social interactions with other people's documents. We demonstrated the "goSemantically" browser-based application, which provides a decentralized environment for users to host and publish semantically enhanced data in HTML and JSON-LD formats. The schema.org website provides the semantic markup tag. Extending the schema.org MedicalEntity and MedicalScholorlyArticle "goSemantically" al-

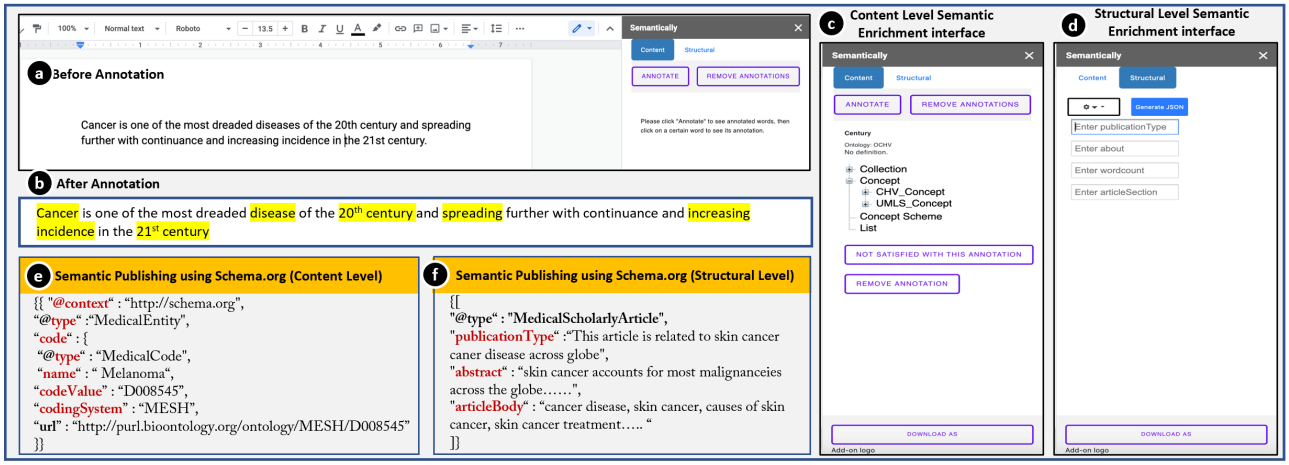

Figure 2: Demonstration Scenario and User Interface of the goSemantically Google add-on

lows the user to download and publish the content as discussed in section (Content-level Semantic Enrichment). The MedicalScholorlyArticle provides a document-level structural, semantic markup such as publicationType, abstract, articlebody, etc. In contrast, MedicalEntity provides content-level structural-semantic markups such as entitytype, its label as shown in Figure.2(e,f). Finally, we couple up the content level and structural level semantic markup by extending Schema.org to publish the biomedical contents online. The "goSemantically" enables users to publish biomedical content automatically by using proposed semantic markups. The published web pages will preserve both the structural (e.g., publication type) and content-level semantics of the biomedical data, as shown in Figure2.

Conclusion

Incorporating machine-interpretable semantic annotations at the pre-publication stage of biomedical content is undoubtedly desired and will add a lot to the larger semantic web vision.The "goSemantically" holistic approach uplifts the syntactic biomedical text by automatically annotating it with the appropriate ontology vocabulary at the sentence level. Additionally, users can assign structural-level schema.org metadata intuitively. The proposed infrastructure allows users to download semantically enhanced information in various web-first compatible formats for hosting and sharing on a decentralized basis. Finally, the incorporated semantic information makes it easier for search engines to index the elevated biomedical contents for intelligent semantic search.

Acknowledgments

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

References

Brady, E., et.al 2015. Creating accessible PDFs for conference proceedings. In Proceedings of the 12th Web for All Conference (W4A 2015): 34-37. ACM.

Spinaci, G.et.al, 2017, August. The RASH JavaScript Editor (RAJE) A Wordprocessor for Writing Web-first Scholarly Articles. In Proceedings of the 2017 ACM Symposium on Document Engineering (pp. 85-94).

Capadisli, S. et.al, 2017, June. Decentralised authoring, annotations and notifications for a read-write web with dokieli. In International Conference on Web Engineering (pp. 469-481). Springer, Cham.

Peroni, S.,et,al., 2017. Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles. PeerJ Computer Science, 3, p.e132.

Lin TTY, Beales G. 2015. ScholarlyMarkdown Syntax Guide. Guide, 31 January 2015.

Mbouadeu, S. F.et.al,2022, January. Towards Structured Biomedical Content Authoring and Publishing. In 2022 IEEE 16th International Conference on Semantic Computing (ICSC) (pp. 175-176). IEEE.

Whetzel, P.L. et.al, 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research, 39, pp.W541-W545.

Abbas, A. et.al, 2021. Clinical Concept Extraction with Lexical Semantics to Support Automatic Annotation. International Journal of Environmental Research and Public Health, 18(20), p.10564.

Abbas, A. et.al, 2019. Meaningful Information Extraction from Unstructured Clinical Documents. Proceedings of the Asia-Pacific Advanced Network, 48, pp.42-47.

Jonquet, C.et.al, 2009, October. NCBO annotator: semantic annotation of biomedical data. In International Semantic Web Conference, Poster and Demo session (Vol. 110). Washington DC, USA.

Hausman, D.M., 2019. What Is Cancer?. Perspectives in biology and medicine, 62(4), pp.778-784.