A Sociotechnical Framework for Semantic Biomedical Content Authoring and Publishing

Steve Fonin Mbouadeu[1], Asim Abbas[1], Fazel Keshtkar[1], Iram Wajahat[2], Syed Ahmad Chan Bukhari*[1]
{steve.mbouadeu19, abbasa, keshtkaf, bukharis}@stjohns.edu
{iramwajahat.pharmd}@gmail.com
[1]Division of Computer Science, Mathematics and Science
St. John's University, Queens, NY, USA
[2]Allied Institute of Medical Sciences, Gujranwala, 52250, Pakistan

Abstract

Due to the ubiquity of unstructured biomedical data, significant obstacles still remain in achieving accurate and fast access to online biomedical content. In lieu of the growing volume of biomedical content on the web, embedding semantic annotations has become key to enhancing search engine context-aware indexing, thereby improving search speeds and retrieval accuracy. We introduce *Semantically*: a socio-technical framework for semantic biomedical content authoring and publishing. Identifying the appropriate semantic vocabulary for biomedical content annotation is a time-consuming and technically challenging process. *Semantically* automates this search by recommending highly accurate annotations from a wide range of biomedical ontologies. Furthermore, the framework is integrated with a knowledge-sharing system which allows biomedical authors to collaborate on identifying precise annotations during the content authoring process. Similarly, preserving content-level semantics during and after publishing to foster semantic search remains a research challenge. *Semantically* addresses this barrier by extending Schema.org, a community-agreed and research engine endorsed guideline for publishing structured content on the web. https://gosemantically.com/

Keywords: Structured data, Biomedical semantics, Automated semantic annotation, Biomedical content authoring, Peer-to-Peer, Structured data publishing, FAIR biomedical data

Background

Semantic Content Authoring (SCA) is the process of embedding semantic annotations with content to improve its machine interoperability and overall FAIRness (Findability, Accessibility, Interoperability, and Reusability) (Khalili and Auer 2013). The process of embedding semantic annotations is mainly composed of 1) finding the reference ontologies, 2) availability of annotator system, and 3) user-friendly interface(s). The reference ontologies, also known as a knowledge base, act as a repository of semantic vocabularies from which the system draws from. More comprehensive and robust taxonomy repositories such as Unified Medical Language System (UMLS) and NCBO BioPortal ontologies yield more accurate annotations (Aronson and Lang 2010). The primary responsibility of an annotation system or annotator is to map entities found in the content to specific ontology concepts (Aronson and Lang 2010). The role of the user interface is to mainly provide a point of control for the various semantic options in the system to abstract the underlying mechanisms.

Introduction

Due to the rapidly expanding research in the biomedical sciences, a massive volume of unstructured biomedical literature has been made available over recent years. There are over 1700 large-scale biological databases and over 30 million citations for biomedical literature on PubMed alone (Jonquet et al. 2009). The rapid growth in the biomedical domain has introduced an access-level challenge for researchers and practitioners (Shah et al. 2009). Despite there being vital information freely available on the web, the lack of machine-interoperable metadata (semantic annotations) renders it ambiguous for search engines to perform information retrieval and knowledge extraction. The ability to disseminate routinely composed and generated medical data into a form of interoperable knowledge is necessary to achieve optimal search accuracy and speed (Bukhari 2017) (Bukhari et al. 2018). The incorporation of machine-interoperable semantic annotations before publication, i.e., while authoring content and maintaining it during its dissemination and publishing, is required to achieve FAIRness in the biomedical domain (Shah et al. 2009). However, exporting and sharing ontological enriched documents in a manner that preserves embedded semantics is complicated and requires advanced technical skills and subject expertise. A cutting-edge, freely accessible biomedical semantic content authoring and publishing framework would change the landscape. To that aim, we introduce *Semantically*, a web framework designed for biomedical researchers and content creators of all experience levels to compose and export semantic biomedical content. Additionally, the second goal of this study is to present the design and development of a publicly available complimentary system that allows *Semantically* users with varying levels of competence in the biomedical domain to collaborate on authoring and to publishing biomedical semantic content. Such a knowledge-sharing approach is ideal for achieving more accurate annotations because it pulls from the experience of a group of domain experts to identify annotation candidates while leaving the au-

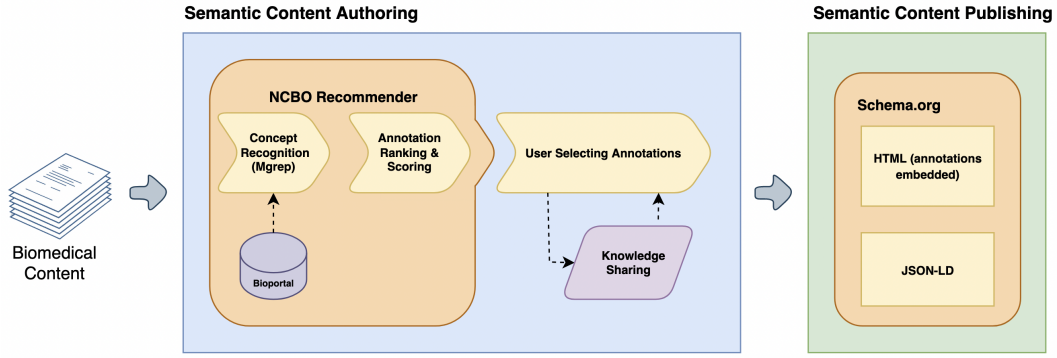

Figure 1: A diagram of the biomedical content authoring and publishing workflow on Semantically

thor with the final decision on which is best for their content (d'Aquin et al. 2008).

The rest of the paper is organized as follows. First, the related work section evaluates other available systems and explains where our proposed approach differs from them. Next, the methodology section covers implementation details of *Semantically*, including the knowledge sharing system. Then, we run through some use cases and demonstrate how to utilize the platform effectively. Lastly, we conclude and discuss future work.

Related Work

The current standard for biomedical semantic content authoring (SCA) is through word processors, which do not produce semantic annotations compatible with the web. Further, the majority of these word processors fail to adhere to Schema.org, a semantic annotation standard for the web endorsed by many search engines, including Google (Seringhaus and Gerstein 2008). This disconnect has brought about the rise of semantic web tagging schemes such as Merkle's Schema Markup Generator, Schema Builder from Schema.dev, and Google's Structured Data Markup Helper. Semantic tagging utilities come in forms ranging from full-fledged web applications to browser extensions, but they all function to retroactively produce and apply semantic annotations to content after publication. Technicians usually perform these tasks intending to boost the content's Search Engine Optimization (SEO), often with little knowledge or involvement in creating that content. A tremendous blocker on the progress of semantic authoring and publishing research is that researchers failed to recognize the importance of involving the original content creator: the author (Mbouadeu et al. 2022). Instead, they concentrated on technological sophistication, limiting system interactions to technical individuals. Often, only the author(s) understand why a certain phrase was employed to convey an idea. Third-party technicians are not privy to such insider information, making them poorly equipped to oversee the SCA process (Abbas et al. 2021). However, most authors lack technical and/or subject (domain) expertise, leaving a steep learning curve to acquire the requisite skills. For them to undertake the process unassisted would be highly time-consuming and distracting.

Proposed Methodology

Our proposed methodology aims to assist biomedical practitioners and researchers uplift their unstructured biomedical content quickly, accurately, and intuitively. The goal is to automatically suggest the appropriate ontology vocabularies for annotation while balancing the accuracy and speed of the entire SCA process. Meanwhile, several Schema.org compliant web publishing options are provided, such as JSON-LD and HTML (with embedded meta tags), which can maintain the content's semantic integrity. The following sections explain our system's semantic content authoring and publishing workflows (Figure 1).

Semantic Content Authoring

The *Semantically* framework was designed for users ranging from bench scientists and medical doctors to writers simply involved in medical journalism. The process starts when the user creates a new document on the platform. They can choose to either start from scratch or import pre-existing content. Following creating a document, the user is redirected to a text editor to view and edit their content before annotating it. To help users make the most out of the platform, *Semantically* provides two annotation modes to get started with the process: *all* and *select ontologies*. The *all* option instructs the annotation system to consider all Bioportal ontologies, whereas *select ontologies* allows users to specify which taxonomies to reference.

Semantically utilizes Bioportal's Recommender API (Jonquet et al. 2009) to perform concept recognition and information extraction. The primary reason we chose Bioportal's annotator rather than alternatives such as Metamap (Aronson and Lang 2010) is that it uses Mgrep for concept recognition which has the advantage of being faster and language agnostic. Metamap utilizes the UMLS corpus and has a more thorough recognition algorithm, making it slower. Once the user selects an annotation mode and completes the ontology configurations, if applicable, the content is passed through Bioportal's Recommender API.

Semantically now has a map of concepts mentioned in the content to ontologies they were matched with. These ontologies are then ranked based on frequency and context analysis on their corresponding terminology (Abbas et al. 2019). An-

other process the framework performs on the matches has to do with Bioportal's annotator considering any entity in the content, both single and complex, for annotations instead of using the longest match. For example, even if the entity *breast cancer* is recognized, the sub-entities *breast* and *cancer* could also be recognized and mapped individually. Our system obfuscates sub-entities and prioritizes overarching entities. Afterward, the annotations are viewable to the user to alter or delete (Figure 2).

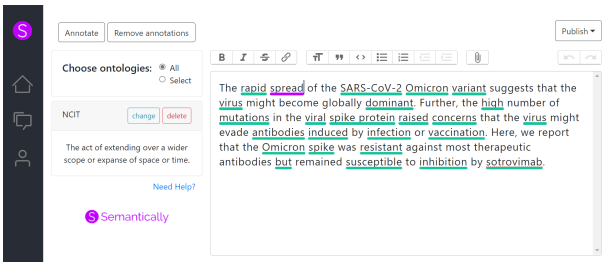

Figure 2: Semantically annotation editing

The Socio-technical Approach

Although only the original author can truly make the right semantic considerations, there often still exists a gap between that knowledge and familiarity with the semantic syntax to represent it, putting these users at a severe disadvantage (Hinze et al. 2019). To address this gap, *Semantically* includes an integrated socio-technical (Semantically Knowledge Café) system for authors within similar domains or expertise to collaborate on semantic annotation. Such community-derived annotation decision-making has been proven to result in more accurate semantic annotations (Hinze et al. 2019). While selecting annotations, authors also have the option of asking for help from the editor view (Figure 2). Once engaged, the user can create a new post and ask a question regarding annotating the selected terminology. Replies are ranked and ordered based on community votes. When a reply is submitted, it becomes viewable to the post's author through the corresponding document's editor panel (Figure 2). From there, the user can either accept or reject the suggestion. Accepting a suggestion changes the ontology of the selected annotation to the one referenced in the response; subsequently, the response is marked as the final answer, and the post is closed, preventing further engagement. Users can utilize this function simultaneously with multiple annotations across documents to quicken the SCA process.

Semantic Content Publishing

Following the SCA process *Semantically* offers, authors can export their content from the framework through Schema.org compliant formats (Figure 3). We utilize Schema.org provided high-level structural tags such as Medical Scholarly Article and Medical Entity and then couple them up with the content-level semantic recorded during the SCA process to generate the semantic metadata embedded in the content.

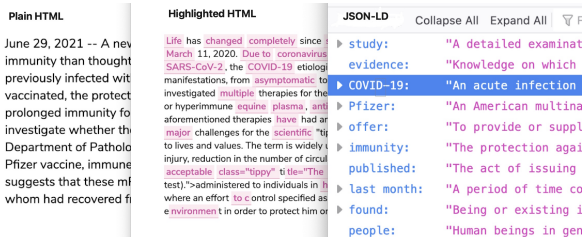

Figure 3: Semantically export formats

HTML Format: The HTML export format is ideal for user's intending to post their content on the web. Semantic annotations are embedded within the document in a format that is easily accessible and interoperable by search engines. **JSON-LD:** *Semantically* provides a JSON-LD export format based on the Schema.org entity structure. JSON-LD is a standardized format for semantic annotations that allows authors to convert it to other preferred structures such as RDF and OWL.

Use Cases and Results

The *Semantically* framework is the ideal platform to semantically enrich and improve the FAIRness of biomedical content ranging from free-written clinical notes to more polished blogs or articles. We will cover two examples of how the framework can be effectively utilized to make biomedical content interoperable from user perspectives with different levels of familiarity with semantic annotations.

Medical Blogging & Casual Writers

We will be annotating a blog post from Medium.com's health section for this demonstration. While annotating this post, we will be behaving like a user who has little to no experience interacting with annotation systems.

To begin, we create a new document on the platform's dashboard. With the content copied into the platform, we can then proceed to annotate using the default settings. Once the system has retrieved annotation candidates for the content, we can begin viewing them and selecting which ones best suit the annotated terminologies. Once that is done, we are ready to publish.

Publishing semantically enriched content is simple on *Semantically*. From the framework's three options, one of the HTML formats would be perfect for this particular scenario. Given that "Medium" allows for importing articles through HTML and links, this option would allow exporting the semantic enriched content from the framework to a blogging platform like "Medium" to happen in just a few clicks.

Biomedical Research & Professional Writers

We will be annotating a biomedical research paper from *Pubmed* for our second demonstration. While annotating

this study, we will be undertaking the perspective of a researcher with experience with semantic annotations and taxonomies used in their domain.

The general setup is the same as the previous run-through, where we create a new document and copy over the content. However, before retrieving annotation candidates, we can make a few changes to the configurations to allow specific ontologies to be matched against our content. Dictionaries such as the Microbial Phenotype Ontology (MPO) and the Infectious Diseases and Antimicrobial Resistance Ontology (IDAR) are a few choices a researcher might consider using utilizing. With one or more dictionaries selected, *Semantically* will retrieve annotation candidates in the content present in those dictionaries exclusively. The researcher can then choose the concepts that best align with the intended definitions they intended.

A blocker the researcher might encounter might be when annotating the terminology *fluoroquinolone*, an antibiotic used to treat bacterial infections. They might not be very familiar with the particular antibiotic and may need some assistance when choosing an annotation for it (d'Aquin et al. 2008); there lies the purpose of *Semantically Café*. After posting a question, such as asking which ontology to use, domain experts can assist by suggesting dictionaries that best align with the terminology from their experience. When one or more replies have been posted, the researcher will review them directly from the document's editor view (Figure 2) and choose which they think is most suitable, using the responses' reasoning and rating as reference.

After annotating the paper, the research could export the annotations. The best format for this example would be to choose the JSON-LD metadata option, which can be submitted alongside the study to a journal (Bukhari, Krauthammer, and Baker 2014).

Conclusion and Future Work

In the future, we aim to minimize repetitive questions. Although all posts are currently made public, there is no easy way for users to query the posts to see if what they need help with has already been addressed previously.

At the pre-publication stage, context-aware authoring and sharing is the least explored aspect of the semantic content life-cycle (d'Aquin et al. 2008). 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 needed. Our future aim is to continue developing the socio-technical infrastructure even further to improve turnaround time and the overall speed of the SCA process on *Semantically*. We look forward to taking further steps towards equipping all biomedical stakeholders with the tools and resources to uplift their content semantically.

Acknowledgments

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

References

[Abbas et al. 2019] Abbas, A.; Afzal, M.; Hussain, J.; and Lee, S. 2019. Meaningful information extraction from unstructured clinical documents. *Proceedings of the Asia-Pacific Advanced Network* 48:42–47.

[Abbas et al. 2021] Abbas, A.; Afzal, M.; Hussain, J.; Ali, T.; Bilal, H. S. M.; Lee, S.; and Jeon, S. 2021. Clinical concept extraction with lexical semantics to support automatic annotation. *International Journal of Environmental Research and Public Health* 18(20).

[Aronson and Lang 2010] Aronson, A. R., and Lang, F.-M. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17(3):229–236.

[Bukhari et al. 2018] Bukhari, S. A. C.; Martínez-Romero, M.; O'Connor, M. J.; Egyedi, A. L.; Willrett, D.; Graybeal, J.; Musen, M. A.; Cheung, K.-H.; and Kleinstein, S. H. 2018. Cedar ondemand: a browser extension to generate ontology-based scientific metadata. *BMC bioinformatics* 19(1):1–6.

[Bukhari, Krauthammer, and Baker 2014] Bukhari, A. C.; Krauthammer, M.; and Baker, C. J. O. 2014. Sebi: An architecture for biomedical image discovery, interoperability and reusability based on semantic.

[Bukhari 2017] Bukhari, S. A. C. 2017. *Semantic enrichment and similarity approximation for biomedical sequence images*. Ph.D. Dissertation, University of New Brunswick (Canada).

[d'Aquin et al. 2008] d'Aquin, M.; Motta, E.; Dzbor, M.; Gridinoc, L.; Heath, T.; and Sabou, M. 2008. Collaborative semantic authoring. *IEEE Intelligent Systems* 23(3):80–83.

[Hinze et al. 2019] Hinze, A.; Heese, R.; Schlegel, A.; and Paschke, A. 2019. Manual semantic annotations: User evaluation of interface and interaction designs. *Journal of Web Semantics* 58:100516.

[Jonquet et al. 2009] Jonquet, C.; Shah, N.; Youn, C.; Callendar, C.; Storey, M.-A.; and Musen, M. 2009. Ncbo annotator: semantic annotation of biomedical data. In *International Semantic Web Conference, Poster and Demo session*, volume 110. Washington DC, USA.

[Khalili and Auer 2013] Khalili, A., and Auer, S. 2013. User interfaces for semantic authoring of textual content: A systematic literature review. *Journal of Web Semantics* 22:1–18.

[Mbouadeu et al. 2022] Mbouadeu, S. F.; Abbas, A.; Ahmed, F.; Keshtkar, F.; De Bello, J.; and Bukhari, S. A. C. 2022. Towards structured biomedical content authoring and publishing. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, 175–176.

[Seringhaus and Gerstein 2008] Seringhaus, M., and Gerstein, M. 2008. Manually structured digital abstracts: a scaffold for automatic text mining.

[Shah et al. 2009] Shah, N. H.; Bhatia, N.; Jonquet, C.; Rubin, D.; Chiang, A. P.; and Musen, M. A. 2009. Comparison of concept recognizers for building the open biomedical annotator. In *BMC bioinformatics*, volume 10, 1–9. Springer.