Promoting Transparency and Trust in Biomedical Data: A FAIR Approach to Content Creation and Sharing

Asim Abbas[1], **Sebastian Chalarca**[1], **Iram Wajahat**[1] ,**Fazel Keshtkar**[1], **Syed Ahmad Chan Bukhari**[*1]
[1]Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. John's University, Utopia,Parkway, Queens, 8000, NY, USA
abbasa@stjohns.edu, sebastian.chalarca19@my.stjohns.edu, wajahati@stjohns.edu, keshtkaf@stjohns.edu, bukharis@stjohns.edu

Abstract

Efficient sharing of scientific knowledge is vital for research advancement and reproducibility. However, PDF dominance poses challenges in adhering to FAIR principles (Findability, Accessibility, Interoperability, Reusability). Some web-based frameworks enhance content interoperability using RDF and linked data but often target technical users. To address this, we introduce "Semantically", a user-friendly platform for biomedical researchers that offers a semantic content authoring module, collaboration tools and improving annotation accuracy. We also propose a publishing infrastructure using schema.org to ensure machine-readable and well-organized datasets, enhancing data FAIRness. Combining Semantically's authoring module and schema.org provides a comprehensive solution for enhancing FAIRness and reproducibility of scientific content. "Semantically" is an open-source tool accessible at Github.

Keywords: Reproducibility, Interoperability, FAIRness, Socio-technical approach, Semantic-enrichment

Introduction

Accessing biomedical content efficiently is crucial for rapid information dissemination within the scientific research community and healthcare professionals [Abbas, 2021]. However, challenges in access have emerged alongside advancements in the biomedical domain. To address these challenges, diverse stakeholders have endorsed the FAIR Data Principles (Findability, Accessibility, Interoperability, and Reusability). These principles aim to enhance the reusability of data holdings by providing a concise and measurable guideline. Unlike initiatives focusing solely on human scholars, the FAIR Principles prioritize enabling machines to autonomously discover and utilize biomedical data, facilitating reuse by researchers and practitioners [Wilkinson, 2016].

Various tools, algorithms, pipelines, and methodologies are developed to prepare machine-interpretable data following FAIR principles. Where their adoption in the biomedical domain has shown benefits like reproducibility, however there are also several challenges that must be taken into account such as lack of standardization, scalability and performance [Madduri, 2019]. To overcome these challenges, a state-of-the-art biomedical semantic content authoring and publishing framework is proposed called "Semantically". This framework would streamline FAIR data preparation, enhance discoverability, and promote FAIR data practices, ultimately making biomedical knowledge more accessible and reusable. The "Semantically" framework aims to enhance the accuracy of biomedical literature and databases through expert peer reviews and automated referencing of biomedical ontology repositories, annotation tools, and web-publishing formats. This open-source framework enables collaborative authoring and publishing of semantically rich content, catering to users with varying levels of expertise in the biomedical domain. By promoting data FAIRness and addressing reproducibility challenges, "Semantically" incorporates a module for the automatic publication of biomedical content as structured data on the web, integrating schema[dot]org[Guha, 2016]. Similarly, developing interactive interfaces is essential for maximizing the framework's utility. "Semantically" empowers biomedical researchers to create, share, and collaborate on research data in a standardized and transparent way, thus advancing scientific knowledge and promoting the reproducibility of biomedical research.

The rest of the article is organized as follows: Section 2 covers the proposed framework and its components. We document experiments and results in Section 3, and Section 4 concludes the paper.

The Proposed FAIR Framework

We address a wide range of stakeholders in the biomedical field, such as academia, industry, funding agencies, and scholarly publishers, responsible for managing biomedical content. Various tools and frameworks have been proposed to help stakeholders achieve FAIR data objectives, but they often have limitations. Introducing "Semantically", a user-friendly infrastructure aimed at aiding users in creating and adhering to FAIR biomedical content, ultimately promoting reproducibility in the biomedical domain see Fig. 1 for an overview. Moreover "Semantically" framework accepts textual data from Biomedical scholarly articles, clinical reports, and other biomedical documents, which are crucial sources

Copyright © 2024 by the authors.
This open access article is published under the Creative Commons Attribution-NonCommercial 4.0 International License.

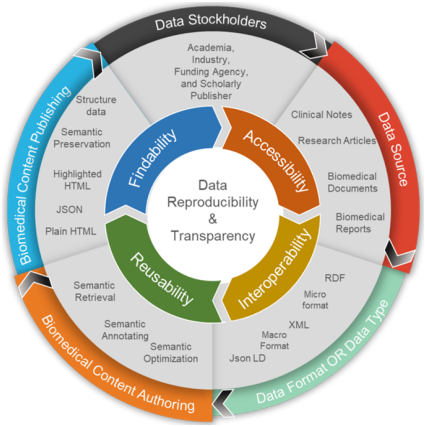

Figure 1: FAIR Biomedical content creation and sharing through "Semantically" Framework process model

of scientific information in the biomedical field. We have introduced data sources like PubMed Central (PMC), MEDLINE and ClinicalTrials.gov providing access to diverse health-related reports and data. These sources employ structured databases, indexing systems, and document repositories to ensure efficient access and retrieval of information. Similarly, various data formats, including RDF, JSON-LD, and XML, play a crucial role in improving the FAIRness of data, enabling better organization, integration, and interoperability of data. Owing to that "Semantically" framework enables users to create and share biomedical content in JSON-LD format, utilizing the schema.org semantic vocabulary. This approach enhances the FAIRness of biomedical data and reusability within the biomedical informatics community. Moreover, the structured representation of data in JSON-LD simplifies interpretation for both humans and machines, fostering data exchange, sharing, and collaborative research efforts.

An interactive biomedical content semantic authoring interface is essential to facilitate end users of any level of expertise in retrieving and optimizing the semantic information of explicit biomedical content in real-time. Regarding real-time content authoring, it is challenging to find precise semantic annotations because a single annotation can exist in multiple biomedical ontologies with varying semantics.In this way, "Semantically" provides an innovative socio-technical and personalized semantic annotation recommendation solution to address this issue and enhance data FAIRness. To study about these approaches, the reader is encouraged to visit these references [Abbas, 2024][Asim, 2023].

Prominent search engines like Google, Yahoo, Bing, and Yandex rely on high-quality semantic data to provide accurate search results. In the similar way, we've developed a method to automatically publish biomedical content in structured formats on the internet, addressing the limitations of current publishing processes. This method aligns with the extension of schema.org and enables semantic searching. Our platform, "Semantically," assists authors in obtaining semantically enriched biomedical content in plain HTML, highlighted HTML, and JSON formats, ensuring the preservation of content-level semantic information for publishing purposes.

Results and Discussion for a FAIR Paradigm

This study focuses on the "Semantically" framework, aiding the biomedical community in preparing FAIR data. The framework transforms unstructured biomedical content into a structured format using a sociotechnical approach to enhance biomedical semantics. To evaluate its efficiency and performance, we utilized the User Experience Questionnaire (UEQ) tool [Schrepp 2015], known for its simplicity and reliability in data analysis. UEQ is widely trusted by organizations as a comprehensive metric for assessing user experience. By employing UEQ, we assessed the user experience of the "Semantically" system effectively. In Fig. 2, we compute the score of "Semantically" users in 6-dimensional scales, where the scale means whose values between -0.8 and +0.8 enact a neutral evaluation, values greater than +0.8 depict a positive assessment and values less than -0.8 represent a negative evaluation of a proposed system. As though *Dependability* in a 6-dimensional scale list score lower than other five means scales because, of *"not secure/secure and unpredictable/predictable"* items. But, overall, the "Semantically" portrays promising user experience in 6-dimensional scales because the values are more significant than 1.6 see Fig. 2. Further, we have comprehensive documentation available online on Github for Semantically users and third parties for system integration.

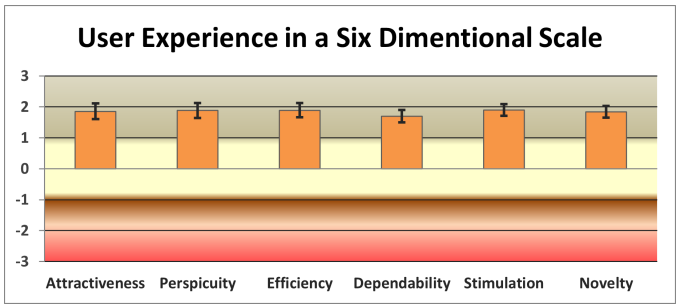

Figure 2: 'Semantically' user experience questionnaire resulting scores on a six-dimensional scale.

Conclusion

In this study we introduces "Semantically," a user-friendly platform designed for biomedical researchers to create and share FAIR-compliant content, promoting reproducibility in science. A key feature is the biomedical semantic content authoring module, which simplifies technical complexities for users while encouraging collaboration with domain experts, resulting in more consistent and relevant annotations. Combining this module with the schema.org data publishing infrastructure ensures machine-readable and interoperable content, addressing FAIRness challenges in biomedical data and advancing scientific accessibility, discoverability, and reproducibility.

References

Journal Article
Abbas, A., Afzal, M., Hussain, J., Ali, T., Bilal, H. S. M., Lee, S., & Jeon, S. (2021). Clinical concept extraction with lexical semantics to support automatic annotation. International Journal of Environmental Research and Public Health, 18(20), 10564.

Journal Article
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

Journal Article
Madduri, R., Chard, K., D'Arcy, M., Jung, S. C., Rodriguez, A., Sulakhe, D., ... & Foster, I. (2019). Reproducible big data science: A case study in continuous FAIRness. PloS one, 14(4), e0213013.

Magazine Article
Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema. org: evolution of structured data on the web. Communications of the ACM, 59(2), 44-51.

Journal Article
Abbas, Asim, et al. "A Socio-Technical Approach to Trustworthy Semantic Biomedical Content Generation and Sharing." Information Sciences (2024): 120441.

Conference Paper
Abbas, Asim, et al. "Personalized Semantic Annotation Recommendations on Biomedical Content Through an Expanded Socio-Technical Approach [Personalized Semantic Annotation Recommendations on Biomedical Content Through an Expanded Socio-Technical Approach]." Personalized Semantic Annotation Recommendations on Biomedical Content Through an Expanded Socio-Technical Approach (2023).

Technical Report
Schrepp, Martin, and Jörg Thomaschewski. "Handbook for the modular extension of the User Experience Questionnaire." Mensch & Computer. 2019.