Adapting Wikidata to support clinical practice using Data Science, Semantic Web, and Machine Learning

Data Engineering and Semantics Research Unit, University of Sfax, Tunisia

Description

Nowadays, semantic resources have been proven as efficient to drive computer applications in a variety of domains, particularly healthcare[1]. Semantic resources provide detailed knowledge about various aspects of medicine including diseases, drugs, genes, and proteins[2] and they can consequently be used to retrieve, process, and represent clinical information including electronic health records[3] and scholarly publications[4]. These databases enable biomedical relation extraction[4], biomedical relation classification[5], biomedical data validation[6], biomedical data augmentation[7], and biomedical decision support[8]. However, the implementation of such biomedical resources in the Global South, particularly in Africa, is still limited due to the lack of consistent funding and human capacities[9]. Here, open knowledge graphs, particularly Wikidata, can be valuable to reduce the financial and technical burden of developing digital health in developing countries[2]. As a free and collaborative large-scale multilingual knowledge graph, Wikidata became a confirmed database that can represent multiple kinds of clinical information, particularly in the context of COVID-19[10]. Its representation in the Resource Description Framework (RDF) format enables the flexible enrichment of biomedical information using computer programs and crowdsourcing, the intrinsic and extrinsic validation of clinical knowledge, and the extraction of features from the medical data for decision making and human and machine learning[10]. Yet, Wikidata still lacks a full representation of several facets of biomedical informatics [2] and its data suffers from critical inconsistencies[11]. For instance, Wikidata items about genes, proteins, and drugs have an average of 10+ statements per item while anatomical structures have only an average of fewer than 4.6 statements per item[2]. Furthermore, more than 90% of the Wikidata statements about human genes and proteins are supported by references whereas only less than 50% of the statements about the anatomical structures are assigned references[2]. Moreover, the linguistic representation of biomedical entities in Wikidata is dominated by German and English when other natural languages are partially or rarely covered[10].

In this research project, we propose to:

- *Turn Wikidata into a large-scale biomedical semantic resource covering most of the aspects of the clinical practice in a significant way*: This is allowed thanks to the development of bots and tools to mass import clinical information from external resources already aligned with Wikidata and to the creation of machine learning algorithms to extract clinical information from the full texts and bibliographic information of scholarly publications indexed in PubMed, a large-scale bibliographic database hosted by National Center for Biotechnology Information and National Institutes of Health. This implies the enrichment of the facets of biomedical knowledge represented in Wikidata and the support of new kinds of clinical information that were not covered by Wikidata during the last few years.
- *Validate the biomedical information freely available in Wikidata*: This is enabled thanks to comparison to external resources through the use of semantic alignments between Wikidata items and external biomedical resources, and to intrinsic validation through the use of SPARQL for identifying mismatches between statements and the use of shape based methods such as ShEx and SHACL as well as property constraints for verifying the accuracy of the formatting and data modeling of the clinical knowledge in Wikidata. These

methods are coupled with the development of *Wikidata Game*-like human validation tools of medical information in Wikidata.

- ➢ *Promote the biomedical use of Wikidata in the Global South*: This is permitted thanks to online capacity-building events for the biomedical community in Africa about Wikidata and its tools, to the publication of surveys and position papers on biomedical applications of Wikidata in recognized research journals. The integration of Wikidata into Fast Healthcare Interoperability Resources, a semantic system for driving Electronic Health Records, is also envisioned to enable the use of human-validated subsets of Wikidata information for clinical reasoning in the Global South.

This project only uses public data and text, and never touches private, restricted, or personal data or health information. The development of these three tasks will allow not only the significant amelioration of Wikidata as a secondary knowledge base for medical information but also the development of a framework for the curation of other famous medical knowledge graphs such as Disease Ontology. The reproducibility of the Project will allow the development of solutions for the enrichment of Wikimedia Projects with knowledge about other research areas such as social science and computer science from open resources, particularly knowledge graphs and bibliographic databases.

To achieve these three tasks, we envision doing the following activities :

- ➢ **S1:** *Turn Wikidata into a large-scale biomedical semantic resource covering most of the aspects of the clinical practice in a significant way*
  - S1.A1: Enriching Wikidata with biomedical knowledge available in external resources: On the one hand, we will develop bots that use semantic alignments (e.g., Disease Ontology ID) between Wikidata items and their equivalents in external resources to extract semantic relations between Wikidata items from these resources and mass import them to the Wikidata knowledge graph. On the other hand, we will apply machine-learning models to bibliographic metadata available in PubMed, mainly the MeSH keywords, to extract biomedical relations between Wikidata items. We already developed *MeSH2Matrix* as an approach for the MeSH keyword-based classification of Wikidata relations[12] and we will be interested during this project to upgrade it using semantic similarity measures and deep learning to turn it into a method for biomedical relation extraction based on PubMed bibliographic metadata.
  - S1.A2: Adding support for new types of biomedical knowledge in Wikidata: We will try to add new properties and classes to Wikidata. We have already proposed new Wikidata properties (e.g., risk factor and medical indication) and we have also added support for new Wikidata classes from scratch (e.g., COVID-19 app). We can reproduce this experience when needed. An example is the inclusion of knowledge about clinical trials in the Wikidata knowledge graph within the framework of WikiProject Clinical Trials[13].
- ➢ **S2:** *Validate the biomedical information freely available in Wikidata*
  - S2.A1: Developing bots and tools for the cross-validation of Wikidata biomedical information from external resources: This work will be a development of the efforts of Wikimedia Deutschland in this context, particularly the Reference Island Project that assigns a reference to a Wikidata statement when it exists in an external knowledge resource. The bots will verify the availability of Wikidata statements in external resources (e.g., Disease Ontology) based on semantic alignments (e.g., Disease Ontology ID) in Wikidata. Toolforge tools will be based on the analysis of bibliographic metadata of scholarly publications coupled with human validation to decide whether a biomedical relation in Wikidata is accurate or not. RefB funded by WikiCred Grant Initiative is an example of preliminary work regarding the use of PubMed data mining for validating and adding reference support to Wikidata biomedical relations.

http://www.deslab.org/ o Route de la Soukra km 4, Sfax 3038, Tunisia

- S2.A2: Developing SPARQL-based approaches for the intrinsic validation of Wikidata clinical knowledge: Despite the usefulness of Shape-based methods (ShEx and SHACL), they would not allow the verification of Wikidata statements through the comparison of their values. Here, SPARQL can be used to identify mismatches between the values of two statements. In the context of the COVID-19 pandemic, we developed a SPARQL-based method for the validation of the epidemiological data about the disease[11]. We look forward to developing our work to cover other biomedical use cases of SPARQL-based validation that cannot be fulfilled by Shape-based methods.
- S2.A3: Developing EntitySchemas in ShEx for validating the shape and representation of medical concepts in Wikidata: We will build upon the efforts of WikiProject COVID-19 to develop data models for supporting biomedical entities related to the ongoing pandemic. We will reuse the data modeling output of WikiProject COVID-19 and extend it to cover other aspects of clinical practice.

➢ **S3:** *Promote the biomedical use of Wikidata in the Global South*

- S3.A1: Organize office hours to demonstrate Wikidata and its medical outputs: We will build upon the success of our previous presentations in Wikimedia Conferences (e.g., Wikimania 2019 and WikiArabia 2021) on the matter and customize our materials to let them more adapted to the healthcare industry and the computer science community in Africa. We will also deal with the technical side of reusing Wikidata in intelligent systems for the clinical practice that was not evocated in our previous presentations and that was used in our research on the topic (e.g., Wikibase Integrator, MediaWiki API, and SPARQL). We will also show examples of several clinical applications where Wikidata can be very useful based on what we have presented in Wikimedia conferences and peer-reviewed research venues.
- S3.A2: Publishing scholarly publications about Wikidata-driven biomedical applications: Develop a roadmap for integrating Wikidata with Fast Healthcare Interoperability Resources to enable the reuse of Wikidata in FHIR-based systems, as well as research works about how to practically use Wikidata in the clinical context and publish them in indexed scholarly journals.

This project goes in line with three points of the 2030 Wikimedia Strategic Direction: “Coordinate Across Stakeholders”, “Increase the Sustainability of Our Movement” and “Identify Topics for Impact”. Developing a framework to update, enrich and validate Biomedical Knowledge in Wikidata will allow ensuring better data quality for Wikidata in the healthcare context. Such quality for a freely available resource will increase the trustworthiness of Wikidata as a reference for physicians, pharmacists, and other medical professionals. This will allow for better patient management and health education in the Global South. This will solve representation gaps related to medical content for content, contributors, and readers as defined in the knowledge gaps taxonomy. Extending the use of Wikidata for clinical practice will allow the creation of knowledge-based medical systems at a low cost. This will allow the achievement of three UN Sustainable Development Goals: “Good Health and Well-Being” (SDG 3), “Quality Education” (SDG 4), and “Sustainable Cities and Communities” (SDG 11). From the perspective of the Wikimedia Movement, the Project will be referential for Wikimedia affiliates and communities from Africa, particularly Wikimedia Tunisia and African Wikimedia Developers Project, if they would like to continue working on the medical output of Wikidata and create projects about biomedical applications of Wikidata or if they would like to formulate a research project and apply for the next editions of Wikimedia Research Fund.

To measure the success of our research project, several objective metrics can be used to evaluate the reach and productivity of our upcoming work:

- Number of scholarly publications in Scimago Q1 computer science and medical research

journals: 3+

http://www.deslab.org/ o Route de la Soukra km 4, Sfax 3038, Tunisia

- Number of proceedings papers in the main tracks of CORE A or A* scholarly conferences: 1+
- Number of proceedings papers in the Workshops of CORE A or A* scholarly conferences: 2+
- Number of office hours: 6+
- Number of presentations in Wikimedia conferences: 3+
- Number of attendees to office hours: 30+ per session

For the dissemination of this Project, we envision publishing most of our research results in recognized scholarly journals as Open Access publications. We look forward to presenting our efforts in Wikimedia Research-related venues such as Wiki Workshop, Wikidata Workshop, and Wiki-M3L as well as in premier scholarly conferences for Knowledge Engineering and Machine Learning (CORE A*), particularly SIGIR and WWW. We will publish our source codes on GitHub under the MIT License for reproducibility purposes. We will participate in Wikimedia Conferences (e.g., WikiArabia, WikiIndaba, Wikimania, and WikidataCon) to disseminate the outcomes of our work to the Wikimedia Community. We will organize regular office hours where we demonstrate our tools live on YouTube and Zoom to the information retrieval, semantic web, biomedical informatics, and clinical medicine communities.

Budget

Approximate amount requested in USD.

49,867.21

Budget Description

We request an amount of 49,867.21 USD for this research project.

Month	Expenses
Month 0	17 696.50 USD
Month 6	12 912.09 USD
Month 9	8,837.72 USD
Month 12	10,420.90 USD
Overall	49,867.21 USD

Salaries

25,970.71 USD is requested for salaries for seven individuals. This covers the creation of research papers related to our scholarly efforts, the development of bots and tools for enriching

and validating medical outputs in Wikidata, the networking efforts with the data-driven Biomedical Informatics R&D Industry across the African continent, and the preparation and organization of online events for the dissemination of medical applications of Wikidata.

Development Equipments

We requested to purchase development equipments at the cost of 15,066.50 USD. This will be useful to implement machine-learning algorithms for enriching and validating Wikidata. Validating all the Wikidata medical output by hand is time-consuming and can lack effectiveness. However, the automatic recognition of Wikidata statements to be added, removed, or updated can help us gain time. While we will make use of Wikimedia-run infrastructure like Toolforge when feasible (especially for testing), having dedicated hardware on-site provides better options to reach production-level support for the new workflows, as well as redundancy to workflows based solely on Wikimedia-run infrastructure. To ensure the accuracy of the output provided by algorithms, we will add a validation layer by human experts for the retrieved data before applying them to Wikidata.

Device	Month 0	Month 6	Month 9	Month 12	Overall
GPU NVIDIA A100	9,955.83 USD	-	-	-	9,955.83 USD
SSD SAS Hard Drive 2 TB * 4	1,659.31 USD	-	-	-	1,659.31 USD
2 Laptops	3451.36 USD	-	-	-	4,196.50 USD
Equipments	15,066.50 USD	-	-	-	15,066.50 USD

OA and Software Expenses

We requested 8,830 USD for OA and Software Expenses. These expenses are meant to cover the article processing fees for our upcoming research publications (2-6 Papers) in recognized journals so that they can be published as Open Access Outputs, allowing easy access to our research results by communities in the Global South for educational and reproducibility purposes. As well, they cover registration to indexed scholarly conferences for demonstrating and discussing our work with the scientific community. Grammarly Premium (300 USD) is required to check the grammar of our research publication. The primary institution applying for the Grant is located in an African country where English is not as used as Arabic and French. Grammarly will save time on proofreading enabling the allocation of more time to research efforts. Overleaf Collaborator (180 USD) is required to allow the seven contributors to work on a paper together when using LaTeX. Using Overleaf will allow us to directly work on Templates without worrying about formatting. As for Zoom Pro (150 USD), we need it to organize online events to disseminate our work as well as for managing the online meetings of the consortium to develop the Project.

Service	Month 0	Month 6	Month 9	Month 12	Overall

Zoom Pro	150 USD	-	-	-	150 USD

Grammarly Premium	300 USD	-	-	-	300 USD
Overleaf Collaborator	180 USD	-	-	-	180 USD
Open Access Publishing	-	3,500 USD	-	3,500 USD	7,000 USD
Top-Tier Scholarly Conference Registration	-	600 USD	-	600 USD	1,200 USD
OA and Software Expenses	630 USD	4,100 USD	-	4,100 USD	8,830 USD

Past Contributions

Three applicants are active long-term contributors to Wikimedia Projects, particularly Wikidata. All of them are Wikimedia researchers and open science advocates with scholarly outputs in major research venues and presentations in the main Wikimedia conferences (e.g., Wikimania, WikidataCon) and in open science events (e.g., Mozilla Festival, Creative Commons Summit). They are also members of Wikimedia Medicine for several years:

- Houcemeddine Turki (User Statistics): Wikimedian since 2009, he served as a Programme Committee Member for WikiIndaba Conference (2018, 2019, and 2021) and WikiConvention Francophone (2021). As well, he was a member of the Wiki Indaba Steering Committee (2018-2021), a member of Wikimedia and Libraries User Group Steering Committee (2019-2021), and the Secretary of the Affiliations Committee (2022). He is currently a member of the Affiliations Committee, a member of the Wikimania 2022 Core Organizing Team, and the Vice-Chair of Wikimedia Tunisia. As a contributor to Wikimedia Projects, he has contributed for ten years to French and English Wikipedia before shifting his interest to Wikidata and Wikifunctions. His main areas of interest are reference support, Tunisia-related topics, library and information science, biomedical informatics, and Applied Linguistics. Furthermore, he is a co-founding member of the Data Engineering and Semantics Research Unit, the first research structure in Tunisia to be specialized in Wikimedia Research and Open Science.
- Daniel Mietchen (User Statistics): He is a biophysicist interested in integrating open research and education workflows with the web, particularly through open licensing, open standards, open collaboration, public version histories, and forkability. With research activities spanning from the subcellular to the organismic level, from fossils to developing embryos, from biodiversity informatics to data science more broadly, and how this all fits with sustainable development, he experienced multiple shades of the research cycle and a variety of approaches to collaboration, sharing and reproducibility in research contexts. He has also been contributing to Wikipedia and its sister projects for about two decades and is actively engaged in increasing the interactions between the Wikimedia and research communities, particularly around Wikidata. All of this informs his current activities as a researcher at the Fraunhofer Institute for Biomedical Engineering, at the Leibniz Institute of Freshwater Ecology and Inland Fisheries, and at the Ronin Institute.
- Lane Rasberry (User Statistics): He is a Wikimedian-in-residence at the School of Data Science at the University of Virginia. His professional interests include popular science, access to health information, clinical research, the Open Movement, data science, and Wikimedia projects. Rasberry holds a B.Sc. in Chemistry from the University of Washington in 2006. He is among the first people to propose that Wikipedia articles should be cited to

ensure better scholarly recognition of Wikipedia as a source for added-

value research outputs. He has also worked on promoting the use of Wikipedia for clinical purposes such as physician-patient communication, medical education, and research. Currently, he is working on the use of Wikidata for representing and integrating clinical trials.

Data Engineering and Semantics Unit is the first research structure in Tunisia to be specialized in Wikimedia Research. It organized two Wikimedia-funded events for the dissemination of Wikidata (AICCSA 2017 Wikidata Presentation and SPARQL: Be connected to Wikidata) and developed the first Tunisian Wikidata user script, Toolforge tool, and bot with the collaboration of Wikimedia Tunisia and through the support of WikiCred Grant Initiative. In 2021, Data Engineering and Semantics has established an advanced research collaboration with Sisonkebiotik, an open African community for Biomedical Machine Learning. They are jointly working on a project entitled "Semantic Applications for Biomedical Data Science" that looks for developing applications of open biomedical knowledge graphs to support clinical efforts. The main two co-founders of the Data Engineering and Semantics Research Unit have been interested in publishing real-life applications of the Wikimedia Projects since 2012. Their research outputs have been successfully published in recognized scholarly journals, particularly *Engineering Applications of Artificial Intelligence*, *Knowledge-Based Systems*, *Journal of Biomedical Informatics*, *Applied Intelligence*, and *Neurocomputing*:

- Mohamed Ali Hadj Taieb is a senior researcher at Data Engineering and Semantics Research Unit. He is an assistant professor of computer science at the Faculty of Sciences of Sfax, University of Sfax, Tunisia. He holds a Ph.D. in Computer Science from the University of Sfax, Tunisia in 2014. His main fields of interest are Semantic Technologies, Scientometrics, Biomedical Informatics, Big Data, Social Networks, and Data Science.
- Mohamed Ben Aouicha is the head of the Data Engineering and Semantics Research Unit. He is an associate professor of computer science at the Faculty of Sciences of Sfax, University of Sfax, Tunisia. He holds a Ph.D. in Computer Science from the University of Sfax, Tunisia, and Paul Sabatier University of Toulouse, France in 2009, and a higher doctorate in Computer Systems Engineering from the University of Sfax, Tunisia in 2016. His main fields of interest are Semantic Technologies, Scientometrics, Information Retrieval, Big Data, Social Networks, and Data Science.

Anastassios Pouris (University of Pretoria, South Africa) and Khalil Chebil (Data Engineering and Semantics Research Unit, Tunisia) have had relevant experience of research collaborations with the industry for many years and practically know what are the needs of the clinical industry in terms of data. Consequently, they have the required skills to define the facets of biomedical knowledge that need to be prioritized for consistent revision in Wikidata and to find third-party individuals to review the output of this research project for human validation of clinical information in Wikidata.

References:

1. Callahan, T. J., Tripodi, I. J., Pielke-Lombardo, H., & Hunter, L. E. (2020). Knowledge-based biomedical data science. *Annual review of biomedical data science, 3*, 23-41. PMC:8095730.
2. Turki, H., Shafee, T., Hadj Taieb, M. A., Ben Aouicha, M., Vrandečić, D., Das, D., & Hamdi, H. (2019). Wikidata: A large-scale collaborative ontological medical database. *Journal of biomedical informatics, 99*, 103292. doi:10.1016/j.jbi.2019.103292.
3. Sun, H., Depraetere, K., De Roo, J., Mels, G., De Vloed, B., Twagirumukiza, M., & Colaert, D. (2015). Semantic processing of EHR data for clinical research. *Journal*

of biomedical informatics, 58, 247-259. doi:10.1016/j.jbi.2015.10.009.
4. Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E. M., & Kors, J. A. (2014). Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics, 15*(1), 64:1-64:8. doi:10.1186/1471-2105-15-64.

5. Hong, G., Kim, Y., Choi, Y., & Song, M. (2021). BioPREP: Deep learning-based predicate classification with SemMedDB. *Journal of Biomedical Informatics, 122*, 103888. doi:10.1016/j.jbi.2021.103888.
6. Nicholson, N. C., Giusti, F., Bettio, M., Negrao Carvalho, R., Dimitrova, N., Dyba, T., et al. (2021). An ontology-based approach for developing a harmonised data validation tool for European cancer registration. *Journal of Biomedical Semantics, 12*(1), 1:1-1:15. doi:10.1186/s13326-020-00233-x.
7. Slater, L. T., Bradlow, W., Ball, S., Hoehndorf, R., & Gkoutos, G. V. (2021). Improved characterisation of clinical text through ontology-based vocabulary expansion. *Journal of Biomedical Semantics, 12*(1), 7:1-7:9. doi:10.1186/s13326-021-00241-5.
8. Tehrani, F. T., & Roum, J. H. (2008). Intelligent decision support systems for mechanical ventilation. *Artificial intelligence in medicine, 44*(3), 171- 182. doi:10.1016/j.artmed.2008.07.006.
9. Odekunle, F. F., Odekunle, R. O., & Shankar, S. (2017). Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. *International journal of health sciences, 11*(4), 59. PMC:5654179.
10. Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., et al. (2021). Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata. *Semantic Web, 13*(2), 233-264. doi:10.3233/SW210444.
11. Turki, H., Jemielniak, D., Hadj Taieb, M. A., Labra Gayo, J. E., Ben Aouicha, M., Banat, M., et al. (2022). Using logical constraints to validate statistical information about COVID-19 in collaborative knowledge graphs: the case of Wikidata. *PeerJ Computer Science* (forthcoming). doi:10.5281/zenodo.4008358.
12. Turki, H., Dossou, B. F. P., Emezue, C. C., Hadj Taieb, M. A., Ben Aouicha, M., Ben Hassen, H., & Masmoudi, A. (2022). MeSH2Matrix: Machine learning-driven biomedical relation classification based on the MeSH keywords of PubMed scholarly publications. In *BIR 2022: 12th International Workshop on Bibliometric enhanced Information Retrieval at ECIR 2022* (forthcoming).
13. Rasberry, L., Tibbs, S., Hoos, W., Westermann, A., Keefer, J., Baskauf, S. J., ... & Mietchen, D. (2022). WikiProject Clinical Trials for Wikidata. *medRxiv*. doi:10.1101/2022.04.01.22273328v1.

http://www.deslab.org/ o Route de la Soukra km 4, Sfax 3038, Tunisia