Developing a Wikimedia-related research structure in a developing country

**Houcemeddine Turki, Montasser Akermi, Amina Amara, Mohamed Ali Hadj Taieb,
Khalil Chebil, Daniel Mietchen, Mohamed Ben Aouicha**

Data Engineering and Semantics Research Unit, University of Sfax, Tunisia
Ronin Institute for Independent Scholarship, Montclair, New Jersey, USA

Abstract

We propose to set up socially-engaged research structures in universities to facilitate collaborations between Wikimedia researchers, industry, government agencies, and civil society actors for real-world applications of Wikimedia projects. The Data Engineering and Semantics Research Unit serves as an example and is accredited, with a scalable data center and outreach efforts to promote diverse interactions with Wikimedia projects.

Keywords: Triple Helix, Wikimedia Research, Data Center, Research Structures, Community Management

Introduction

Scholarly research has evolved from a focus solely on scientific advancement to include solving real-world problems and creating tangible benefits for society. This has resulted in collaborations between universities, industry, and governments known as the Triple Helix model (Etzkowitz et al., 2000). Civil society actors, such as NGOs, community groups, and social entrepreneurs, have also become important players in shaping research agendas and driving innovation (Leydesdorff et al., 2011). The Wikimedia Foundation, as a civil society actor, contributes to research initiatives by maintaining and analyzing its projects for open data and knowledge sharing, launching grant programs, advancing research, and promoting transparency and openness (Nielsen, 2019). However, Wikimedia has faced challenges in expanding its collaborations beyond direct partnerships with scientists to include a more diverse range of stakeholders. One obstacle is the perception of Wikipedia as an unreliable source of information, despite significant improvements in its accuracy and reliability (Okoli et al., 2014). Another challenge is the decentralized and volunteer-driven nature of Wikimedia projects, which complicates long-term partnerships with external stakeholders (Hladchenko et al., 2019). To address these limitations, we propose the establishment of socially-engaged research structures within universities that can facilitate collaborations between Wikimedia researchers, industry partners, government agencies, and civil society actors. These structures can help to develop real-world applications of Wikimedia projects and ensure that the research is relevant and beneficial to society as a whole. In this context, we showcase the Data Engineering and Semantics Research Unit established at the University of Sfax, Tunisia (DES-RU) as a Wikimedia-related research structure following the same guiding principles as socially engaged research structures and AI startups collaborating with African communities (Akiki et al., 2022; Currin et al., 2023).

Seeking Accreditation

In Tunisia, there are two types of research structures: research laboratories and research units, with the former having more responsibilities and more permanent staff. DES-RU was created in 2021. It is founded and directed by Dr. Mohamed Ben Aouicha. This unit is linked to the Faculty of Sciences of Sfax, the University of Sfax, and the Ministry of Higher Education and Scientific Research of Tunisia. The evaluation of DES-RU is conducted by the National Committee for the Evaluation of Scientific Research Activities (NCESRA). It is composed of research professors from different universities as well as specialties. Its role is the evaluation of research structures. In this context, this committee gave a favorable opinion of DES-RU after the evaluation of the project proposed by this unit over four years. Tunisia's Minister of Higher Education and Scientific Research has approved this assessment. This evaluation process lasted almost two years. Accreditation is vital for research structures like DES-RU as it provides credibility and recognition, and demonstrates that they meet the required standards and criteria. Accreditation helps attract funding, partnerships, collaboration, and top talent. Many funding agencies require accreditation to ensure their investments go to credible research institutions. Accredited research structures also provide a more favorable environment for research and learning, attracting talented individuals committed to producing high-quality research (Camara et al., 2010).

Data Center Building

A data center houses powerful servers and equipment to manage, process, and distribute large amounts of data. Our research unit uses it for Wikimedia projects, and it is designed for reliability, scalability, and security, with redundant power and cooling systems, raised floors, and security measures (Jin et al., 2020). The data center is designed with scalability in mind. This involves expanding the facility's power and

cooling system to accommodate additional equipment without interfering with running operations (Jin et al., 2020). This enables flexible maintenance, software installation, big data processing, and running large deep learning models. Running multiple virtual machines on a single server maximizes hardware use, making it more affordable and easier to maintain (Jin et al., 2020). Adding GPUs can improve data center performance for compute-intensive tasks like data analytics and deep learning. Open-source frameworks and tools are favored in the data center for their flexibility, reliability, cost-effectiveness, and strong user communities. Popular tools include the *Linux* operating system, *Proxmox* virtual environment, *Hadoop* ecosystem, *Nginx* web server, *Kubernetes* container orchestration system, *Ansible* automation tool, *pfSense* firewall, and *Zabbix* monitoring solution (Hausenblas, 2018).

Building Capacities and Community Outreach

DES-RU is dedicated to encouraging its members to use Wikimedia projects and contribute to them. This is achieved by providing training that covers both the backend and frontend perspectives of these projects, including an understanding of their purpose and governance. Members are taught to use tools such as the MediaWiki API and to develop user scripts, apps, and bots for the Wikimedia projects (Table 1). They also attend tutorials from the Wikimedia Technical Community and local conferences to learn about the technical side of the Wikimedia projects and develop their skills. This prepares members to create and manage large-scale Wikimedia research projects. The research unit organizes workshops on topics like semantic technologies, database management, data science, web development, and knowledge

engineering to enhance members' theoretical foundations, enabling them to contribute effectively to Wikimedia R&D. The unit collaborates with Wikimedia affiliates to share findings (Table 2), apply for research funding (Table 3), and gain insights into the needs of the Wikimedia community. Active Wikimedians from Tunisia and abroad play a critical role in conducting research, enabling the unit to understand and enrich the Wikimedia ecosystem (Figure 1). The unit participates in Wikimedia conferences to interact with the community, share findings, and learn about new developments in the field.

Conclusion

This work advocates for the creation of socially-engaged research structures in universities to facilitate collaborations between Wikimedia researchers, government agencies, industry partners, and civil society actors. This expands the impact of Wikimedia projects by engaging a wider range of stakeholders. DES-RU is presented as an example of a socially-engaged research unit that uses a scalable data center, open-source frameworks, and tools, provides Wikimedia project training to its members and contributes systematically to Wikimedia projects.

Acknowledgments

References

Etzkowitz, H., and Leydesdorff, L. 2000. The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research policy*, 29(2): 109-123.

Leydesdorff, L., and Deakin, M. 2011. The triple-helix model of smart cities: A neo-evolutionary perspective. *Journal of urban technology*, 18(2): 53-63.

Nielsen, F. Å. 2019. *Wikipedia research and tools: Review and comments*. Technical University of Denmark.

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., and Lanamäki, A. 2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12): 2381-2403.

Akiki, C., Pistilli, G., Mieskes, M., Gallé, M., Wolf, T., Ilic, S., and Jernite, Y. 2022. BigScience: A Case Study in the Social Construction of a Multilingual Large Language Model. In *Workshop on Broadening Research Collaborations 2022*.

Currin, C. B., Asiedu, M. N., Fourie, C., Rosman B., Turki, H., Siam M., Tonja, A. L., Abbott, J., Ajala, M., Azime, Adedayo, S. A., Emezue, C. C., and Machangara, D., 2023. A Framework for Grassroots Research Collaboration in Machine Learning and Global Health. In *2023 ICLR First Workshop on Machine Learning and Global Health*.

Hladchenko, M., and Pinheiro, R. 2019. Implementing the triple helix model: means-ends decoupling at the state level?. *Minerva*, 57(1): 1-22.

Jin, C., Bai, X., Yang, C., Mao, W., and Xu, X. 2020. A review of power consumption models of servers in data centers. *applied energy*, 265: 114806.

Hausenblas, M. 2018. *Container Networking*. O'Reilly Media, Incorporated.

Camara, A., and Toure, K. 2010. African universities strategize and struggle to research and make research matter. *Research Global*, (23): 1-4
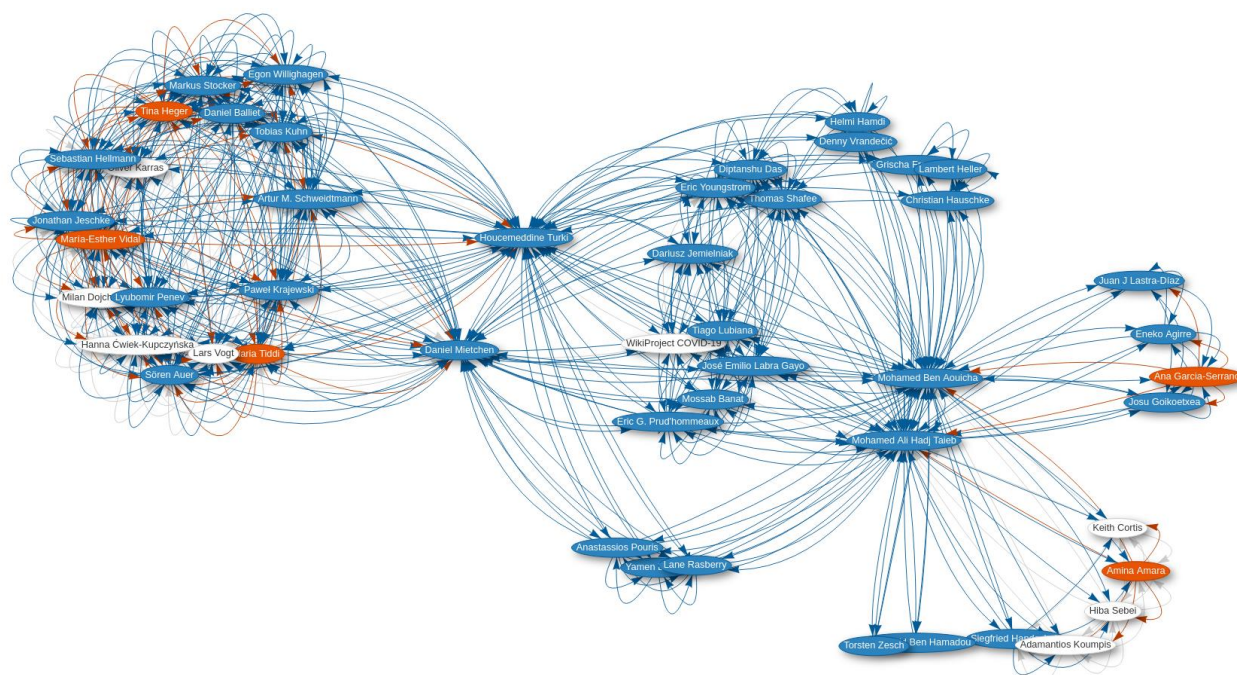
Figure 1: Partial collaboration network of members of the Data Engineering and Semantics Research Unit with the Open Science community and active Wikimedians as of March 2023 (Source: https://w.wiki/6V73)

Table 1: Wikimedia Development Projects by Data Engineering and Semantics Research Unit

Name	Description	URL
Structured Categories	A user script that adds a structured description to a Wikipedia category based on the semantic information about its members in Wikidata	https://w.wiki/4ACt
Sawtpedia	A web service that generates a QRCode related to a topic that once scanned will fetch the Wikidata item for that topic and then open the audio file for the Wikipedia article about the monument in the mobile device's language if available in Wikimedia Commons or generates audio from the lead of the Wikipedia article in the user language.	https://sawtpedia.toolforge.org/
RefB	A Python bot that identifies unsupported biomedical relations on Wikidata using a SPARQL query and finds references to them in PubMed by mining the associations of the MeSH Keywords using the BioPython Python Library.	https://github.com/Data-Engineering-and-Semantics/refb

Table 2: Main scholarly publications by Data Engineering and Semantics Research Unit

Year	Title	Source	DOI
2022	Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata	Semantic Web	10.3233/SW-210444
2022	Using logical constraints to validate statistical information about disease outbreaks in collaborative knowledge graphs: the case of COVID-19 epidemiology in Wikidata	PeerJ Computer Science	10.7717/peerj-cs.1085

Table 3: Grants received by Data Engineering and Semantics Research Unit

Funder	Title	URL
Wikimedia Foundation	Adapting Wikidata to support clinical practice using Data Science, Semantic Web and Machine Learning	https://w.wiki/6bkM