Informing GLAM Institutions and Humanities Researchers of the Broader Impact of Open Data Sharing via Wikidata

Hanlin Li
University of Texas at Austin

Nicholas Vincent
Simon Fraser University

Abstract

We propose to investigate how GLAM institutions and humanities researchers can help to address knowledge gaps by contributing their local collections, metadata, and records to Wikidata. Many GLAM institutions and projects from humanities scholars (e.g. the Women's Print History Project, a bibliographical database) hold valuable information and historical artifacts that have the potential to fill critical knowledge gaps on Wikidata; however, these contributors often lack visibility into what knowledge gaps their local databases are well positioned to address. This limits their ability to effectively organize their efforts and fulfill their mission of sharing human knowledge and mitigating epistemic injustice. We propose to develop an assessment tool that can highlight what past contributions are particularly valuable for addressing knowledge gaps and providing unique coverage relative to other knowledge technologies (search engines and LLMs). This tool will help GLAM and humanities contributors to better understand the unique value of their various contributions and make informed decisions about their future focus. The successful completion of the project will help GLAM institutions and humanities scholars optimize their efforts to mitigate the most critical gaps in the knowledge ecosystem.

Introduction

GLAM institutions and humanities researchers hold a wealth of information, records, and data essential for the humanities, science, cultural heritage, and the knowledge ecosystem. In recent years, the turn towards "collections as data" has further promoted GLAM institutions to share and aggregate collections and records in digital data formats to facilitate open access [12]. Amidst this movement, Wikidata has emerged as a collaborative platform where GLAM institutions and humanities scholars (e.g. the Women's Print History Project) share their historical information, metadata, and knowledge [5]. Currently, this knowledge base also serves as a source of training data for widely used AI models, in addition to powering search engines, discovery aids, and other Wiki systems such as Wikipedia [1,4,6,10,11,13].

However, librarians, catalogers, archivists, and humanities scholars do not have a way to gauge what knowledge gaps their local databases are well positioned to address, when they seek to share knowledge and mitigate epistemic injustice. In our completed interview study, GLAM contributors reported the lack of insights

into the value of their contributions relative to the knowledge ecosystem as a barrier to effective planning. Will their contributions have the unique potential to address critical knowledge gaps on a given subject? Or is this information already abundantly available and therefore, should be deprioritized? Given their limited resources and staff hours and the scale of their institutional collections and records, they urgently need a way to optimize the time they spend on contributing their localized data to Wikidata and thereby to the broader knowledge ecosystem.

To address this challenge, we propose developing an assessment tool to highlight the knowledge gaps that GLAM and humanities contributors have addressed with their Wikidata contributions to support their reflection and planning. This tool will help GLAM and humanities contributors see how their past contributions extend the coverage of knowledge systems, particularly search engines and large language models. Based on the information the tool provides, GLAM and humanities contributors can make more informed decisions to optimize their time and efforts and to subsequently maximize their impact on the knowledge ecosystem.

Date: July 15, 2025 - July 14,2026.

Related work

Our proposed study will build on prior efforts that aim to connect GLAM institutions and humanities scholars with peer production projects like Wikidata, such as the "Wikipedia Loves Libraries" and "Wikimedian in Residence" projects and various efforts from the Wikimedia Foundation [3,5].

The design of the assessment tool is motivated by findings from Li's completed interview study with 15 GLAM contributions [1] and Vincent's ongoing efforts in the Women's Print History Project [2]. Participants have reported the challenge of knowing what to prioritize when contributing to Wikidata and are concerned about "going down a rabbit hole" unnecessarily and "not knowing when to stop". We aim for our assessment tool to provide better support for contributors to plan and identify valuable areas of contribution. Additionally, we have collected specific feedback on the information and metrics GLAM and humanities contributors would like to see when prioritizing their efforts, including subject areas, sources, and number of existing contributions.

How the tool assesses the value of Wikidata contributions will be based on existing studies of knowledge systems' coverage. There is a growing body of research on benchmarking large language models' knowledge in either specific disciplines such as history [7] or general topic areas [14]. We will draw inspiration from such prior work and develop prompts to test large language models' knowledge on subjects on which GLAM and humanities contributors have made extensive contributions. We will also send queries to search engines to assess whether the information from Wikidata contributions is already present in these systems, building on Vincent's prior work in a similar vein [15].

Methods

We have already collected contribution histories from over 600 information professionals who participated in the PCC Wikidata Pilot Project–an initiative bringing together GLAM

[1] This study has recently undergone major revision at ACM CSCW (2025), a premier academic publication venue for social computing and peer production research.
[2] https://womensprinthistoryproject.com/

institutions across the globe to work together to contribute to Wikidata[3].

Using this dataset, we will first provide an assessment of the unique value of the knowledge shared by GLAM contributions, relative to various snapshots of the "coverage" of the broader knowledge ecosystem – i.e. search engines and various versions of prominent large language models (GPT-4o, GPT-3.5, Llama-3.1-70B). We will construct queries around the knowledge shared by GLAM and humanities contributors, such as "the publication date of The Coinage of the Western Seleucid Mints", and send these queries to search engines and large language models. Moreover, our benchmark process will map GLAM and humanities contributions to various topical and format domains [16] to provide a better understanding of what topics or formats benefit the most from GLAM and humanities contributors.

Informed by our analysis results and our need-finding interview studies, we will then visualize our results with an interactive tool. The tool will inform contributors what domains of past contributions are particularly valuable in addressing knowledge gaps so they can make more informed decisions about what to contribute next. The tool will provide a display of how their past contributions are situated in the broader knowledge ecosystem, with respect to search engines and large language models. For example, contributors will be able to see whether search engines or large language models have accurate, consistent information about the rare book author they have added to Wikidata. The tool will allow contributors to explore the value of their own contributions as well as that of their institutions or project as a whole to support collective reflection and planning.

We will conduct an initial evaluation of the tool with GLAM and humanities contributors in a workshop format. We will propose a workshop at the Texas Conference on Digital Libraries or adjacent conferences to introduce Wikidata and our tool to potential and existing GLAM contributors. We will ask participants to explore the tool and gather responses on whether and how the assessment tool helps them better understand their contributions and prioritize their future efforts. We will work with our institute's IRB analyst to secure approval for human subjects research.

Expected output

The expected outcomes of the project will be:

- An assessment tool for GLAM and humanities contributors to identify what types of knowledge gaps they are well positioned to address. The assessment tool may also be used by other Wikidata contributors to see the unique value they contribute to the knowledge ecosystem.
- A research report for the Wiki Workshop, intended for Wikimedia researchers and interested community members. This report will summarize our process and findings from our initial evaluation.
- A workshop on training potential and existing GLAM contributors to leverage Wikidata and our assessment tool for broader impacts. All training materials will be published on our event page.
- A scientific publication submitted to a leading HCI conference (tentatively SIGCHI), intended for academic researchers and Wikimedia researchers.

[3]https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot

This open access publication will pave the way for researchers to investigate the impact of Wikidata on other knowledge systems and to build tools to make the impact visible to Wikidata contributors.

- A long-term collaborative grant proposal for Li and Vincent that focuses on identifying new ways of motivating GLAM and humanities contributors to share their valuable, localized data and records with Wikidata, such as building a recommender system based on both their past contributions [2] and the value of those contributions.

Risks

The assessment tool may be appealing to actors who want to identify specific areas to vandalize, such as introducing inaccurate information about lesser-known authors to Wikidata. To mitigate this risk, the tool will not include assessment of certain types of contributions that are frequent targets of vandalism such as the gender property. We also plan to release the tool first to GLAM and humanities contributors and collect their feedback on how to maintain the tool to prevent misuse.

Community impact plan

This project aims to make an impact on GLAM and humanities communities. In particular, our workshop described in Methods will serve both as an evaluation opportunity and an outreach activity. To further our impact, we will also share the tool with the LD4 community [4], the University of Texas Linked Data Learning group, and the Women's Print History Project, all of which Li or Vincent is part. The tool and the training materials we develop for the workshop will also be reused in Li's class on data management that trains the future generation of GLAM professionals.

During the development of the tool, we will actively seek input from GLAM and humanities contributors, including those who participated in our interview study and expressed interests in the tool. Additionally, we will provide regular research updates on our project page and monitor community feedback.

While the primary research activities will take place at UT, Vincent from Simon Fraser University will engage similarly with Canadian libraries and humanities projects throughout the project, and in particular, collaborators from the Women's Print History Project.

Evaluation

The success of the project will be measured by the adoption of the assessment tool by GLAM and humanities contributions in their practices. Impact on academic research will be measured by scholarly engagement with our work, such as citations and future studies directly informed by our findings.

Budget

Our budget includes primarily personnel support, conference travels for Vincent and a doctoral student researcher (Li will apply for travel support from her home institution), and compute credits. The personnel portion includes 0.75 months of support for Li and 4.5 months of stipend and tuition support for a doctoral student researcher. See our budget sheet for details.
Budget Sheet

References

1. Tarfah Alrashed, Lea Verou, and David

[4] https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group

Karger. 2022. Wikxhibit: Using HTML and Wikidata to Author Applications that Link Data Across the Web. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–15. https://doi.org/10.1145/3526113.3545706
2. Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces* (IUI '07), 32–41. https://doi.org/10.1145/1216295.1216309
3. Alex Stinson Darnell Sandra Fauconnier, Susanna Ånäs, Liam Wyatt, Jane. 2016. Why you should be paying attention to Wikidata and GLAM. *Wikimedia Foundation*. Retrieved April 16, 2025 from https://wikimediafoundation.org/news/2016/08/23/wikidata-glam/
4. Erenrich. 2024. Wikidata & AI, together again. *Wikimedia Tech News*. Retrieved September 20, 2024 from https://tech-news.wikimedia.de/en/2024/02/19/wikidata-ai-together-again/
5. Jan Dittrich e.V Wikimedia Deutschland. 2019. *English: We talked to 16 users who worked at different cultural ("GLAM") institutions to find out about "How and why do people in cultural institutions use Wikidata?" and thus learn more about participants' motivations, activities and problems. We did the research from June 2019 - September 2019.* Retrieved October 24, 2024 from https://commons.wikimedia.org/wiki/File:Research_Report_%E2%80%93_Use_of_Wikidata_in_GLAM_institutions_(2019-11).pdf
6. Shani Evenstein Sigalov and Rafi Nachmias. 2023. Investigating the potential of the semantic web for education: Exploring Wikidata as a learning platform. *Education and Information Technologies* 28, 10: 12565–12614. https://doi.org/10.1007/s10639-023-11664-1
7. Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and R. M. del Rio-Chanona. 2024. Large Language Models' Expert-level Global History Knowledge Benchmark (HiST-LLM). *Advances in Neural Information Processing Systems* 37: 32336–32369. Retrieved April 16, 2025 from https://proceedings.neurips.cc/paper_files/paper/2024/hash/38cc5cba8e513547b96bc326e25610dc-Abstract-Datasets_and_Benchmarks_Track.html
8. Mo Houtti, Isaac Johnson, Joel Cepeda, Soumya Khandelwal, Aviral Bhatnagar, and Loren Terveen. 2022. "We Need a Woman in Music": Exploring Wikipedia's Values on Article Priority. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2: 1–28. https://doi.org/10.1145/3555156
9. Mo Houtti, Isaac Johnson, Morten Warncke-Wang, and Loren Terveen. 2024. Leveraging Recommender Systems to Reduce Content Gaps on Peer Production Platforms. https://doi.org/10.48550/arXiv.2307.08669
10. Shicheng Liu, Sina J. Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, and Monica S. Lam. 2024. SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions. https://doi.org/10.48550/arXiv.2407.11417
11. Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured Knowledge Infusion for Large Language Models. Retrieved May 8, 2024 from http://arxiv.org/abs/2205.08184
12. Thomas G. Padilla. 2018. Collections as data: Implications for enclosure. *College & Research Libraries News* 79, 6: 296. https://doi.org/10.5860/crln.79.6.296
13. Merrilee Proffitt. 2021. Leveraging Wikipedia: Connecting Communities of Knowledge. *OCLC*. Retrieved September 18, 2024 from https://www.oclc.org/research/publications/2018/leveraging-wikipedia.html
14. Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A Complex, Natural, and Multilingual Dataset for End-to-End Question Answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1604–1619. Retrieved April 16, 2025 from https://aclanthology.org/2022.coling-1.138/
15. Nicholas Vincent, Isaac Johnson, Patrick

Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. *Proceedings of the International AAAI Conference on Web and Social Media* 13: 505–516. Retrieved September 16, 2019 from https://www.aaai.org/ojs/index.php/ICWSM/article/view/3248

16. Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the Web: Constructing Domains Enhances Pre-Training Data Curation. https://doi.org/10.48550/arXiv.2502.10341