COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

MatSci-YAMZ: Integrating Human and Artificial Intelligence Vocabulary

PRACTICE PAPERS

]u[ubiquity press

SCOTT MCCLELLAN
ADDISON IRELAND
COLTON GERBER
JOEL PEPPER
CHRISTOPHER B. RAUCH
MAT KELLY
JOHN KUNZE
JANE GREENBERG

*Author affiliations can be found in the back matter of this article

ABSTRACT

This paper presents MatSci-YAMZ, a novel implementation of Yet Another Metadata Zoo (YAMZ), a crowdsourced vocabulary application. MatSci-YAMZ is designed to assist vocabulary development for materials science by configuring a large language model (LLM) and provenance tracking into the YAMZ workflow. Implementation of MatSci-YAMZ occurred in three phases: requirements and model selection, development, and fine-tuning. The application utilizes the Gemma3:27b LLM to formulate definitions based upon examples provided with human definitions. Provenance tracking captures the evolution of LLM-generated definition evolution, which occurs in response to user comments. MatSci-YAMZ can improve shared semantics among communities of materials researchers whose terminology may contain nuances or idiosyncratic usages. Improved semantic alignment supports more concise data descriptions and use of metadata, which is foundational to supporting the FAIR principles of findability, accessibility, interoperability, and reusability. Improved provenance tracking also provides greater insight into how researchers engage with LLM-generated definitions as well as the changes that occur. MatSci-YAMZ opens several avenues of research into crowdsourced vocabulary development among materials researchers. Current findings highlight the need for additional user testing and study to better understand human-in-the-loop vocabulary development with existing LLMs.

CORRESPONDING AUTHORS:

Scott McClellan

Metadata Research Center, Drexel University, US

sm4522@drexel.edu

Jane Greenberg

Metadata Research Center, Drexel University, US

jg3243@drexel.edu

KEYWORDS:

Metadata; Crowdsourcing; Terminology; LLM; Human in the Loop; Materials Science

TO CITE THIS ARTICLE:

McClellan, S., Ireland, A., Gerber, C., Pepper, J., Rauch, C.B. Kelly, M., Kunze, J. and Greenberg, J. 2026 MatSci-YAMZ: Integrating Human and Artificial Intelligence Vocabulary. *Data Science Journal*, 25: 23, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2026-023

INTRODUCTION

Concise, unambiguous terminology is critical for communication both within and among scientific research communities, particularly in environments where data is being shared. This communication relies upon high quality descriptive metadata to support data sharing. Precise semantic descriptions of digital artifacts are integral to ensuring researchers are describing the same data. However, this is often complicated by discipline-specific terminology, where common terminology can assume idiosyncratic or nuanced meanings. This is especially true in areas where research communities consist of diverse groups. Materials science is a diverse community where precise vocabulary supporting data sharing has gained prominence. Key factors include data sharing policies, national initiatives, such as the Materials Genome Initiative (MGI) (Subcommittee on the Materials Genome Initiative, 2021), and adoption of the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (Wilkinson *et al*., 2016).

Connected with this growth is the development of tools to support vocabulary development (e.g., VocPopuli and the PMD 'ontology playground'). While these tools have helped build infrastructure, there is little evidence of their integration of large language models (LLMs) into their workflows. Contemporary vocabulary tools should take into account the growing reliance of researchers' use of LLMs for a variety of aspects involving language, including, but not limited to, summarizing and describing results. Given their increasing integration into research workflows, understanding how LLMs generate and interpret domain-specific terminology relative to human experts is essential for improving semantic precision and mitigating known limitations, such as hallucinations (Tóth and Abdelzaher, 2024; Wang *et al*., 2025). LLMs present a new opportunity to complement vocabulary tools to help researchers address terminology challenges.

This paper considers this opportunity and reports on a novel approach to configuring an LLM within the workflow of a previously existing application, Yet Another Metadata Zoo (YAMZ), which allows researchers within a community to crowdsource vocabulary development (Greenberg *et al*., 2014). To develop consensus, YAMZ has traditionally leveraged subject matter expertise provided by human researchers, but LLMs offer a possible alternative knowledge base from which to draw. Prior work with YAMZ has studied vocabulary among materials researchers, but MatSci-YAMZ has been developed specifically to study possible points of convergence between definitions produced by human materials science researchers and LLMs (Greenberg *et al*., 2023; McClellan *et al*., 2024).

Following this introduction, this paper describes the three phases of implementing MatSci-YAMZ with an LLM feature. This is followed by a discussion of the LLM and provenance features. The final section presents a conclusion, which identifies next steps and future considerations for advancing MatSci-YAMZ.

IMPLEMENTATION

RELATED WORK

Various tools and infrastructures have been developed to support community-driven vocabulary development and semantic alignment in the materials research domain. Formal ontologies often rely on smaller groups of domain experts working with ontology engineers to construct highly structured knowledge organization systems (KOS). The European Materials Modelling Council's Elementary Multiperspective Materials Ontology (EMMO) provides a high-level framework that can integrate additional subject-specific ontologies into it (Goldbeck *et al*., 2019). Another effort is the PMD core ontology (PMDco), which utilizes an "ontology playground," which supports community-driven sessions whereby materials scientists can suggest new terms, report issues, and learn about ontology development (Bayerlein *et al*., 2024).

Alongside formal ontology efforts, other communal approaches to developing materials science vocabulary have emerged. The National Institute of Standards and Technology (NIST) has focused on building shared materials terminology through both top-down and community contributions to its Materials Resource Registry (NMRR) vocabulary (Medina-Smith *et al*., 2021; Plante *et al*., 2021). Similarly, the Novel Materials Database (NOMAD) takes a more communal approach to metadata and vocabulary developments for its data dictionary, soliciting terms from researchers and standardizing them for use in the NOMAD repository (Draxl and Scheffler, 2018). Another recent collaborative vocabulary development environment is MetaCook, specifically the module VocPopuli, which supports user-generated terminology and definitions through a standalone software environment that facilitates the development of ontologies from them (Bagov *et al.*, 2022).

MatSci-YAMZ does not claim direct technical interoperability with EMMO, PMDco, NMRR, or NOMAD. Similar to these other systems, MatSci-YAMZ supports community-driven terminology development and the overriding aim of generating shared, consistent, and reusable vocabulary. Crowdsourced definitions produced in MatSci-YAMZ are stored in a PostgreSQL database with provenance records, and a future aim is to assign Archival Resource Key (ARK) persistent identifiers. The ARKs could potentially be mapped to equivalent concepts in EMMO, PMDco, or the NMRR, depending on the scope of these communities. The MatSci-YAMZ community validation process (human commenting, voting, and LLM-assisted refinement) is similar to the community curation models supporting both the PMDco's ontology playground and NOMAD's data dictionary. These similar frameworks may potentially provide future infrastructure for term alignment. Technical integration with these other initiatives would require significant commitments and sustainability plans.

Overall, while these developments have advanced approaches to collaborative vocabulary development and consensus building, integration of LLMs into existing tools and their workflows is unknown within the scope of their efforts to develop and curate terminology. There is therefore a need to explore AI in this context given the growing sophistication and integration of LLMs into many data-driven workflows. Researchers have shown interest in AI approaches to describing research results, and LLMs may enhance both the efficiency and quality of vocabulary development. This need has motivated the development of MatSci-YAMZ and the work presented in this paper.

OVERVIEW

The primary motivation for developing the MatSci-YAMZ application was to explore the role LLMs might play in collaborative vocabulary development within the materials science community. Materials science subdomains draw from other disciplines, e.g., physics, chemistry, engineering, and generally focus on the process-structure-property-performance relationships of their results, which may be either computational or experimental in nature (Olson, 1997). A system such as MatSci-YAMZ might improve vocabulary usage among material scientists by providing more robust alternative definitions or greater disambiguation of terminology. Implementation of these features occurred in three phases: requirements and model selection, implementation and coding, and fine-tuning (Figure 1). MatSci-YAMZ also introduces provenance tracking of changes made to LLM-generated definitions, which allows us to develop insights into the lifecycle of terminology.

Figure 1 Overview of implementation phases.

PHASE 1: REQUIREMENTS AND MODEL SELECTION

The initial phase of development focused on defining specific features and requirements for the project. We chose the existing application YAMZ based upon its prior applicability to studying materials research vocabulary (McClellan *et al.*, 2024). The MatSci-YAMZ application is coded in JavaScript and uses PostgreSQL for database management, both of which permit flexibility to upgrade or alter backend functionality. JavaScript allowed for the faster integration of the LLM into the workflow using Ollama, an open-source tool that enabled us to run LLMs directly on our local hardware. PostgreSQL also supports provenance tracking, an additional feature that allows us to record information regarding changes to all definitions.

MatSci-YAMZ incorporates and extends features of the original YAMZ application. The application uses Google authentication to provide users access to enter definitions, comment, and vote. Users provide a term, definition, and example, which are all required fields, and they can also

DOI: 10.5334/dsj-2026-023

electively comment or vote on either their own or other users' definitions. The user-generated MatSci-YAMZ definition view for the term 'material' can be seen in Figure 2a. The system automatically populates an LLM-generated definition based upon the first user-generated term's example, Figure 2b. There is only one LLM definition for a given term to reduce the quantity of terms users have to engage with.

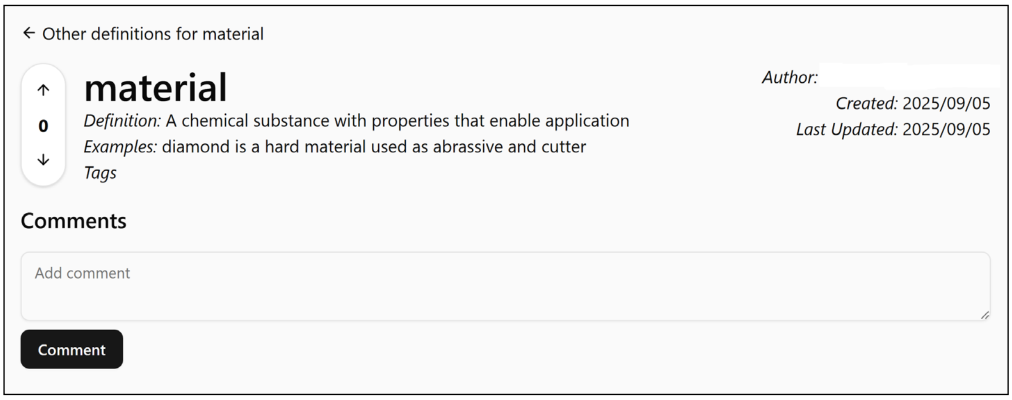

Figure 2a MatSci-YAMZ user-generated definition.

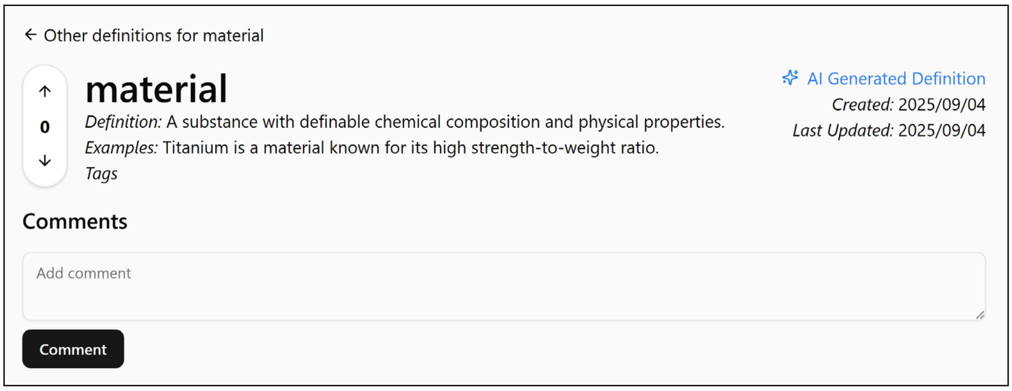

Figure 2b MatSci-YAMZ LLM-generated definition.

To locate a suitable LLM, we examined several models and evaluated their performance based upon several characteristics, including:

- Definition length in words
- Example length in words
- Time for LLM to produce definition
- Overlap of LLM definition with a pre-existing reference definition from (Callister and Rethwisch, 2020)

We used Ollama to download and execute four models to test performance: Gemma3:27b, deepseek-r1:70b, gpt-oss, and qwen3:32b (DeepSeek-AI *et al.*, 2025; Merhej *et al.*, 2025; Open-AI *et al.*, 2025; Yang *et al.*, 2025). We then examined a few commonly used materials science terms to compare model outputs with standard textbook definitions. These included 'Anisotropy', 'Dislocation', 'Grain Boundary', and 'Creep'. Tables 1a and 1b summarize LLM definitions for 'Anisotropy'.

Each model was loaded, and the system message was set to the following: 'You are to define materials science terms. Keep definitions concise and don't be conversational, just respond with a definition and an example using the term with the given definition'. Each test term was then submitted to a given model to produce a definition and example of the term being used. This evaluation served primarily as a calibration exercise to ensure models could generate materially relevant definitions and to identify basic performance characteristics. When applicable, models were tested with 'thinking mode' both enabled and disabled. 'Thinking mode' is a feature that allows a model to generate and use intermediate reasoning steps before producing its final answer (Wu *et al.*, 2024).

A set of criteria was established to guide LLM model selection for MatSci-YAMZ. Criteria selection was guided by preliminary testing, computational constraints, and a balance between semantic quality, accessibility, and computational efficiency. Larger models demonstrated the capacity for more detailed definitions, although they generally require significantly longer processing times and higher computational costs. We established basic selection criteria that included 1) alignment with established definitions from key materials science resources, 2) conciseness and clarity of the AI-generated definitions, and 3) response time in relation to the previous two metrics. As can be seen in Table 1a, the models produced definitions that generally aligned with the reference definition for 'anisotropy' as defined by Callister and Rethwisch (2020). With the definitions produced, to establish model efficiency, the number of tokens for each definition and example pair was added together and divided by the duration it took the model to produce them, and this yielded the token per second (TPS) metric. As can be seen in Table 1b, definitions produced with 'thinking mode' enabled yielded longer definitions and examples, but the qwen3:32b and deepseek-r1 took much longer to produce, while the gpt-oss model, which only operates in 'thinking mode', functioned much faster at 26.25s, yielding 5.37 TPS. The non-thinking mode models generally produced shorter definition example pairs, ranging from 40–52 tokens, while simultaneously taking less time to produce definitions. The gemma3:27b model, which has no 'thinking mode', yielded 5.24 TPS, outperforming the qwen3:32b and deepseek r-1 models. Our ultimate choice came down to gpt-oss and gemma3:27b, and the latter was chosen for its speed, producing definition and example pairs in 8.56s compared to 26.25s for gpt-oss. Speed was valued here over the computational cost.

Table 1a LLM definitions for benchmark term 'Anisotropy', for which a standard textbook definition is 'exhibiting different values of a property in different crystallographic directions' (Callister and Rethwisch, 2020).

MODEL NAME	THINKING MODE	LLM DEFINITION	LLM EXAMPLE
gemma3:27b	no	Property of a material exhibiting different properties when measured in different directions	Wood is anisotropic; it splits easily along the grain but is strong across it.
deepseek-r1	no	The property of a material that exhibits different physical properties in different directions.	A carbon fiber composite material may have high strength along the direction of the fibers but lower strength perpendicular to them, demonstrating anisotropy.
	yes	Anisotropy refers to the dependence of a material's properties on direction, where these properties vary when measured from different directions.	Wood exhibits anisotropy as its strength is greater along the grain than across it; similarly, graphite shows varying hardness in different directions.
qwen3:32b	no	The property of a material to exhibit different physical or mechanical properties in different directions.	Graphite shows anisotropy because it is soft and easily cleaved in one direction but rigid in another.
	yes	A material property that varies with direction, resulting in different physical or mechanical behaviors depending on the orientation within the material.	Wood exhibits anisotropy in mechanical strength, being significantly stronger along the grain direction (longitudinal) than across the grain (transverse).
gpt-oss	yes	The directional dependence of a material's physical properties (e.g., electrical, thermal, magnetic, or mechanical).	The anisotropy of graphite causes it to conduct electricity strongly along the planes of carbon layers while remaining electrically insulating perpendicular to those planes.

Table 1b LLM tokens and TPS for benchmark term 'Anisotropy'.

MODEL NAME	THINKING MODE	TOKENS (DEFINITION + EXAMPLE)	TOKENS PER SECOND	TOTAL DURATION (SECONDS)
gemma3:27b	no	40	5.24	8.56
deepseek-r1	no	52	1.87	30.84
	yes	424	1.57	270.1
qwen3:32b	no	46	3.79	14.37
	yes	372	3	124.11
gpt-oss	yes	141	5.37	26.25

PHASE 2: DEVELOPMENT

The development phase took place in two distinct stages. The first stage involved tailoring MatSci-YAMZ to the requirements in Phase 1. The application was written in JavaScript using the NextJS React framework as the front end, as this combination allowed us to more efficiently coordinate development of both. For our database, we utilized PostgreSQL, which was also used in the original YAMZ application, and the schema currently employed is available in Figure 3. Apart from the LLM integration, MatSci-YAMZ includes version control or tracking to note historical changes to definitions within the PostgreSQL database.

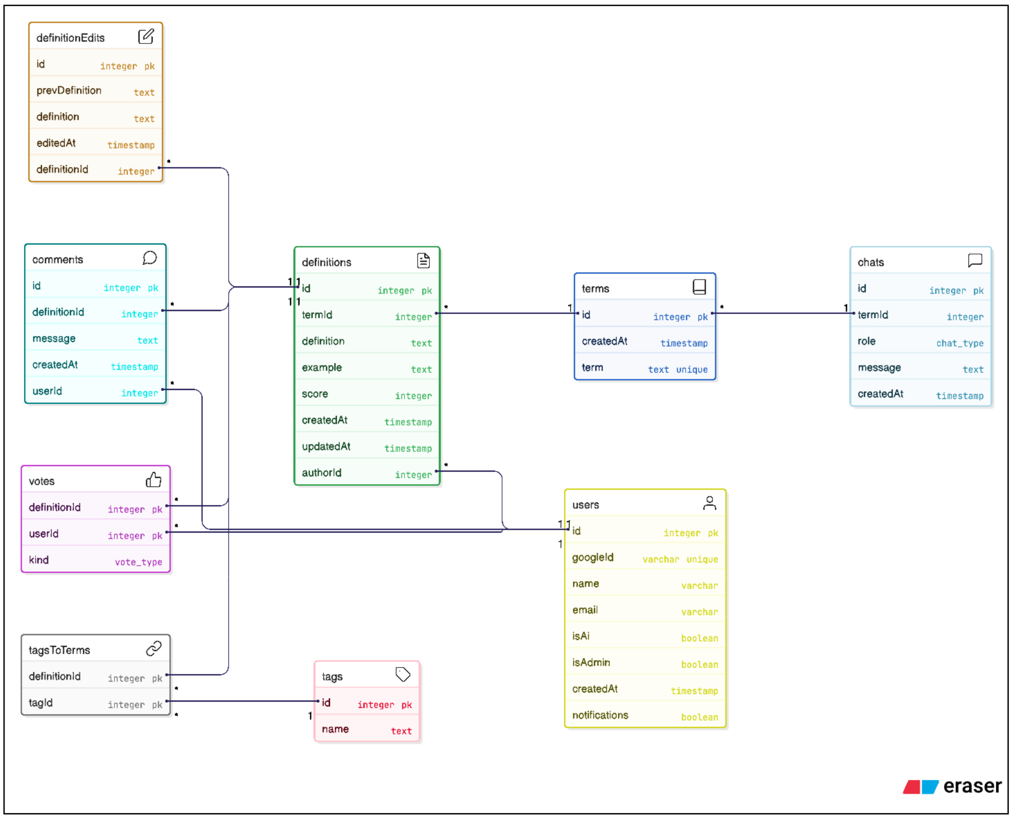

Figure 3 Database schema for MatSci-YAMZ PostgreSQL backend.

The second stage of coding focused on LLM inclusion in the preexisting YAMZ application. To practically implement the LLM into the MatSci-YAMZ workflow, we chose Ollama to run the model for its ease of use. The first benefit it provided was its simplicity to set up. Ollama, combined with Docker operating system virtualization, allowed us to quickly download and run models with multiple billions of parameters by automatically setting up GPU acceleration, increasing the performance of all models. When it came to integrating Ollama into the application, we took advantage of its REST API, which permitted us to run the LLM on one server and the web application on another. Not only did this ultimately increase the reliability of the entire application, but it also ensured that users were less likely to experience delays while the computationally expensive models were running in the background. While the Ollama API is extensive, only one endpoint was needed. This endpoint allowed us to run the model, set the system prompt, and choose how long it would stay in memory.

The coding and implementation stage was also where we developed a framework that would allow materials researchers to leverage aspects of human definitions within the context of the LLM. In keeping with our LLM requirements, we retained the system prompts to produce the initial definition and added a prompt to execute changes based upon user comments to the LLM-generated definitions. This latter prompt structure is designed to activate meaningful changes to the LLM-generated definitions based upon user information. An important and novel change from the original YAMZ application was the inclusion of version tracking, which allows administrators to see how the LLM responds to user inputs. Versions of LLM-generated definitions are stored in a PostgreSQL table, allowing researchers to run standard queries to study and compare changes across multiple definitions if necessary. The version tracking allows administrators to see a record of human and LLM interactions that produce user-visible LLM-generated definitions.

PHASE 3: FINE-TUNING

This phase focused on fine-tuning the features developed in phase 2. We tested the application with a small group of materials scientists to ensure it functioned according to the requirements described in Phase 1 (Greenberg *et al*., 2025). To accomplish this, we engaged several materials researchers who entered definitions and offered informal feedback. The purpose of these tests was to determine if the application functioned properly and that the LLM returned possible definitions related to materials science. The original workflow of MatSci-YAMZ required an administrator to manually run the LLM process. Based on feedback from testers, it was updated to automatically produce definitions. Further testing with additional users will be necessary to assess questions about the quality and relevance of definitions or user experience issues. In addition to user comments, we also wanted to ensure that the workflow functioned properly and that versions of definitions were captured in the PostgreSQL database. We detected no anomalies in small-scale use, and the Ollama interface functioned properly.

DISCUSSION

MatSci-YAMZ, a vocabulary development application for materials science, uses two complementary approaches. *First,* it supports collaborative definition building among domain experts, as established in the original YAMZ framework; *second*, it incorporates LLM-generated definitions to expedite and refine terminology development. The integration of LLMs into MatSci-YAMZ provides a new pathway for researchers to examine ambiguous or underspecified terminology. Additionally, LLM-generated definitions help identify terminology gaps, inconsistencies, and ambiguities in human-authored definitions, helping to identify terms that require further clarification or refinement. Support for human commenting and voting, as noted by the standard YAMZ features, also guided the evolution of LLM outputs. This interaction represents a human-in-the-loop (HILT) approach—an iterative exchange between domain expertise and probabilistic language generation that supports the refinement and stabilization of terminology (Shneiderman, 2020).

Provenance tracking is another key feature of MatSci-YAMZ. Provenance is critical for establishing authenticity and supporting trust in evolving knowledge systems. MatSci-YAMZ tracks how a term's definition changes over time in response to user comments, which activate LLM-generated modifications. This functionality highlights that vocabulary development is a dynamic rather than a static process. Provenance data associated with LLM-generated definitions provides insight into how meaning is negotiated, refined, and stabilized within a research community. This aspect captures semantic change, which is a part of LLM development, where tracking the evolution of language outputs can reveal patterns in model interpretation and adaptation (Periti & Montanelli, 2024). Provenance tracking for human-generated definitions will be added in future versions of the application, providing further semantic insights into how community-negotiated semantics differ from that of human-AI interactions.

MatSci-YAMZ, as an application, can contribute to improving the quality and usability of metadata by supporting more transparent and context-aware terminology development. In particular, the planned integration of Archival Resource Key (ARK) persistent identifiers, combined with provenance tracking, offers a mechanism for linking specific definitions to identifiable points in time and use. This capability enhances traceability and reproducibility in metadata practices and supports more precise data description and curation (Mayernik, 2016; Erkimbaev *et al*., 2019). It is also important to acknowledge that MatSci-YAMZ and the work presented in this paper can contribute to supporting FAIR findability through establishing standardized, interoperable and reusable vocabulary.

The work presented in this paper should be understood as an initial implementation and exploration. Testing to date has been limited in scale, and further evaluation with domain experts is necessary to fully assess LLM-generated definition quality, usability, HILT dynamics, and MatSci-YAMZ's impact on research workflows. It is also important to note that the use of LLMs may introduce limitations. Definitions generated by LLMs are derived from probabilistic modeling rather than domain-specific expertise, which may result in inaccuracies, oversimplifications, or context loss. While the MatSci-YAMZ HILT framework mitigates some of these risks through expert validation, the extent to which human interaction improves definition quality remains an open question. Additionally, variation across models and prompt configurations requires further investigation.

Despite noted limitations, MatSci-YAMZ, as explored and presented in this paper, and other recent work (e.g., Greenberg *et al.*, 2025), confirm the feasibility and potential of integrating LLMs into collaborative vocabulary development workflows. This work addresses the need to explore further the role of LLMs and HILT solutions for improving domain terminology specificity and accuracy. By enabling structured interaction between domain experts and LLM-generated content, the application provides a foundation for studying semantic alignment, terminology evolution, and the role of AI in shaping scientific communication.

CONCLUSION

MatSci-YAMZ demonstrates the successful integration of an LLM and provenance tracking into the workflow of the pre-existing YAMZ application. The primary feature added to YAMZ for MatSci-YAMZ is the LLM definition generator. We integrated the LLM into the workflow using the Ollama application. The LLM formulates a definition based upon the text from the example field of a human-generated term, and updates occur when users comment on the LLM definition. This feature provides possible alternate definitions to those generated by human users. In addition to LLM integration, MatSci-YAMZ employs provenance tracking, which allows administrators to assess changes and comments to all definitions that occur over time. The use of PostgreSQL tables to store this data allows it to be analyzed in conjunction with data from various areas of MatSci-YAMZ. These features taken together allow us to use MatSci-YAMZ to study how definitions evolve in response to comments, voting, and changes to other definitions, with the goal of improving semantics within and among materials research communities.

Furthermore, MatSci-YAMZ offers a variety of options for future study and improvements. Future steps for the application focus on three primary areas: user testing, LLM expansion, and PIDs. These areas will expand our knowledge of the role of LLMs in the production of definitions as well as implications for FAIR data practices. The first of these will involve more extensive user testing to understand better how LLM definitions compare with human definitions. To investigate the effectiveness of the application and its ability to produce definitions, we have arranged a proof-of-concept demonstration to assess the definitions and changes incurred through the definition commenting process (Greenberg *et al.*, 2025). Further user tests will explore the quality of definitions based upon community voting. Another planned area of development involves expanding the use of the Ollama interface to implement alternatives or even multiple LLMs in different configurations. Another element of the LLM workflow that might improve results involves studying how different system messages or prompts for a given LLM alter the definitions provided (Shah, 2023). The final proposed area of development involves implementing the ARK persistent identifiers (PIDs), which were omitted in the current iteration of MatSci-YAMZ. This feature will allow data to conform better to FAIR standards, improving the interoperability and reusability of terminology. ARKs will also permit researchers to reference terminology, including idiosyncratic usages deployed in published articles.

MatSci-YAMZ offers a starting point for studying issues that have arisen concerning the roles of LLMs in the production of specialized knowledge within the context of materials science terminology (Zhao, 2025). By providing a space where materials researchers can interact with LLM-generated definitions and make communal decisions regarding terminology, MatSci-YAMZ provides a broader knowledge base from which definitions can be formed. However, it remains to be seen whether other LLMs or system prompts might produce more salient definitions. Furthermore, provenance tracking will allow greater insights into how definitions change over time. Collectively, the features of MatSci-YAMZ point toward the need for continued research into how LLMs might support vocabulary development within and among communities of materials scientists.

APPENDIX A: REQUIREMENTS

Ollama version: 0.11.3
LLM version : Gemma 3:27b
NextJS version : 15.3.3
Node version : 20.19.4
PostgreSQL: 12.14
Web Server: Nginx 12.14

GPU Server:

- Ubuntu 22.04.5
- 2x Nvidia Quadro RTX 4000

Docker version: 24.0.7

DATA ACCESSIBILITY STATEMENT

The code for MatSci-YAMZ can be found in the following GitHub repository, https://github.com/metadata-research/matsci-yamz.

FUNDING INFORMATION

The authors acknowledge funding from NSF grant NSF-HDR-OAC #2118201, HDR Institute for Data-driven Dynamical Design (ID4).

COMPETING INTERESTS

JG serves as a member of the Data Science Journal's editorial board. The journal's editorial policies ensure that board members are blinded to the identity of reviewers evaluating their submissions to maintain the integrity of the review process. The other co-authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

SM, AI, and JG manuscript authors. AI, CR, MK, JK, and JP assisted with application development, design, coding, and technical operations. CG assisted with subject matter specific material. All authors listed assisted with conceptualization of this paper.

AUTHOR AFFILIATIONS

Scott McClellan orcid.org/0000-0002-1524-8346
Metadata Research Center, Drexel University, US

Addison Ireland
Pennsylvania State University, US

Colton Gerber orcid.org/0000-0002-5471-0679
Toberer Group, Colorado School of Mines, US

Joel Pepper orcid.org/0000-0002-1601-8729
Metadata Research Center, Drexel University, US

Christopher B. Rauch orcid.org/0000-0003-2061-3413
Metadata Research Center, Drexel University, US

Mat Kelly orcid.org/0000-0002-0236-7389
Metadata Research Center, Drexel University, US

John Kunze orcid.org/0000-0001-7604-8041
Metadata Research Center and ARK Alliance, US

Jane Greenberg orcid.org/0000-0001-7819-5360
Metadata Research Center, Drexel University, US

REFERENCES

Bagov, I., Greiner, C. and **Garabedian, N.** (2022) 'Collaborative metadata definition using controlled vocabularies, and ontologies', *Research Ideas and Outcomes*, 8, pp. 907–928. Available at: https://doi.org/10.3897/rio.8.e94931

Bayerlein, B., Schilling, M., Birkholz, H., Jung, M., Waitelonis, J., Mädler, L. and **Sack, H.** (2024) 'PMD core ontology: Achieving semantic interoperability in materials science', *Materials & Design*, 237, Article 112603. Available at: https://doi.org/10.1016/j.matdes.2023.112603

Callister, W., Jr. and **Rethwisch, D.G.** (2020). *Callister's Materials Science and Engineering* (10th Edition, Global Edition edition.). Hoboken, NJ: Wiley.

DeepSeek-AI, Yang, D., Song, J., Xu, R., Zhu, Q., Ma, S., Bi, X., Zhang, X., Yu, X., Li, Z., Liu, A., Xue, B., Wu, B., Lu, C., Zhao, C., Zhang, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen,

J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Jin, R.C., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wui, S., Ye, S., Yuun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Sui, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z. and **Zhang, Z.** (2025). 'DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning'. Available at: https://doi.org/10.48550/arxiv.2501.12948

Draxl, C. and **Scheffler, M.** (2018) NOMAD: The FAIR concept for big-data-driven materials science. Available at: https://doi.org/10.48550/arxiv.1805.05039

Erkimbaev, A.O., Zitserman, V.Y., Kobzev, G.A. and **Kosinov, A.V.** (2019) 'Curation of digital scientific data', *Scientific and Technical Information Processing*, 46(3), pp. 192–203. Available at: https://doi.org/10.3103/S0147688219030092

Goldbeck, G., Ghedini, E., Hashibon, A., Schmitz, G.J. and **Friis, J.** (2019) 'A reference language and ontology for materials modelling and interoperability', *National Agency for Finite Element Methods and Standards (NAFEMS World Congress) 2019*. Quebec City, Canada, 17–20 June. Quebec, Canada: NAFEMS.

Greenberg, J., McClellan, S., Ireland, A., Sammarco, R., Gerber, C., Rauch, C.B., Kelly, M., Kunze, J., An, Y. and **Toberer, E.** (2025). 'Human-in-the-loop and AI: Crowdsourcing metadata vocabulary for materials science', *Metadata and Semantics Research Conference 2025*. Available at: https://doi.org/10.48550/arXiv.2512.09895

Greenberg, J., McClellan, S., Rauch, C., Zhao, X., Kelly, M., An, Y., Kunze, J., Orenstein, R., Porter, C., Meschke, V. and **Toberer, E.** (2023) 'Building community consensus for scientific metadata with YAMZ', *Data Intelligence*, 5(1), pp. 242–260. Available at: https://doi.org/10.1162/dint_a_00211

Greenberg, J., Murillo, A., Kunze, J., Callaghan, S., Guralnick, R., Nassar, N., Ram, K., Janee, G. and **Patton, C.** (2014) 'Metadictionary: Advocating for a community-driven metadata vocabulary application', *International Conference on Dublin Core and Metadata Applications: DC 2013*. Lisbon, Portugal, 2–6 September. Available at: https://dcevents.dublincore.org/IntConf/dc-2013/paper/view/209/211.html

Mayernik, M.S. (2016) 'Research data and metadata curation as institutional issues', *Journal of the Association for Information Science and Technology*, 67(4), pp. 973–993. Available at: https://doi.org/10.1002/asi.23425

McClellan, S., de Oliveira, I.M., Rauch, C., Adriaenssens, S. and **Greenberg, J.** (2024) 'Exploratory analysis of a crowdsourcing metadata tool for building terminological consensus in civil engineering', *Automation in Construction*, 166, 105627. Available at: https://doi.org/10.1016/j.autcon.2024.105627

Medina-Smith, A., Becker, C.A., Plante, R.L., Bartolo, L.M., Dima, A., Warren, J.A. and **Hanisch, R.J.** (2021) 'A controlled vocabulary and metadata schema for materials science data discovery', *Data Science Journal*, 20(1), Article 18. Available at: https://doi.org/10.5334/dsj-2021-018

Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Ramos, S., Yvinec, E., Beyer, L., Tsitsulin, A., Busa-Fekete, R., Sachdeva, N., Gao, Y., Barr, I., Parisotto, E., Eyal, M., Bhupatiraju, S., Kumar, R., Vilar, D., Luo, J., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C.J., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noldand, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orban, J., Fernandez, J., Newlan, J., Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J-B., Anil, R., Lepikhin, D., Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Hussenot, L. and **Dadashi, R.** (2025) Gemma 3 technical report. Available at: https://doi.org/10.48550/arxiv.2503.19786

Olson, G.B. (1997) 'Computational design of hierarchically structured materials', *Science*, 277(5330), pp. 1237–1242. Available at: https://doi.org/10.1126/science.277.5330.1237

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, m., Gogineni, T., Goucher, A., Gross, L., Guzman, K.G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E.Y., Leclerc, G., Lennon, J.P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Liu, L., Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhao, S. and **Zhang, E.** (2025) gpt-oss-120b & gpt-oss-20b model card. Available at: https://doi.org/10.48550/arxiv.2508.10925

Periti, F. and **Montanelli, S.** (2024) 'Lexical semantic change through large language models: A survey', *ACM Computing Surveys*, 56(11), Article 282. Available at: https://doi.org/10.1145/3672393

Plante, R.L., Becker, C.A., Medina-Smith, A., Brady, K., Dima, A., Long, B., Bartolo, L.M., Warren, J.A. and **Hanisch, R.J.** (2021) 'Implementing a registry federation for materials science data discovery', *Data Science Journal*, 20(1). Available at: https://doi.org/10.5334/dsj-2021-015

Shah, C. (2023) From prompt engineering to prompt science with human in the loop. Available at: https://doi.org/10.48550/arxiv.2401.04122

Shneiderman, B. (2020) 'Human-centered artificial intelligence: Reliable, safe & trustworthy', *International Journal of Human–Computer Interaction*, 36(6), pp. 495–504. Available at: https://doi.org/10.1080/10447318.2020.1741118

Subcommittee on the Materials Genome Initiative. (2021) *Materials genome initiative strategic plan*. Office of Science Technology and Policy, National Science and Technology Council. Available at: https://www.mgi.gov/sites/mgi/files/MGI-2021-Strategic-Plan.pdf

Tóth, Á. and **Abdelzaher, E.** (2024) 'BERT may help in lexicographic sense delineation', *International Journal of Digital Humanities*, 6(3), pp. 291–311. Available at: https://doi.org/10.1007/s42803-024-00093-2

Wang, C., Liu, X., Yue, Y., Guo, Q., Hu, X., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Hu, X., Qi, Z., Gao, W., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z. and **Zhang, Y.** (2025) 'Survey on factuality in large language models', *ACM Computing Surveys*, 58(1), pp. 1–37. Available at: https://doi.org/10.1145/3742420

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 'tHoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhoa, J, and **Mons, B.** (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), Article 160018. Available at: https://doi.org/10.1038/sdata.2016.18

Wu, T., Lan, J., Yuan, W., Jiao, J., Weston, J. and **Sukhbaatar, S.** (2024) Thinking LLMs: General instruction following with thought generation. Available at: https://doi.org/10.48550/arxiv.2410.10630

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z. and **Qiu, Z.** (2025) Qwen3 technical report. Available at: https://doi.org/10.48550/arxiv.2505.09388

Zhao, X. (2025) *Large-scale materials knowledge extraction using LLMs and human-in-the-loop*. PhD dissertation. Drexel University. Available at: https://doi.org/10.17918/00010968

TO CITE THIS ARTICLE:
McClellan, S., Ireland, A., Gerber, C., Pepper, J., Rauch, C.B. Kelly, M., Kunze, J. and Greenberg, J. 2026 MatSci-YAMZ: Integrating Human and Artificial Intelligence Vocabulary. *Data Science Journal*, 24: 25, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2026-023

Submitted: 08 December 2025
Accepted: 22 June 2026
Published: 13 July 2026