PRACTICE AND IMPLEMENTATION PAPER

Building Community Consensus for Scientific Metadata with YAMZ

Jane Greenberg[1†], Scott McClellan[1], Christopher Rauch[1], Xintong Zhao[1], Mat Kelly[1], Yuan An[1], John Kunze[1], Rachel Orenstein[2], Claire Porter[2], Vanessa Meschke[2], and Eric Toberer[2]

[1]Metadata Research Center, College of Computing and Informatics, Drexel University, 3675 Market St 10th floor, Philadelphia, PA 19104, USA

[2]Dept. of Physics, Colorado School of Mines, 1500 Illinois St, Golden, CO 80401, USA

Keywords: Metadata standards; FAIR metadata; Consensus building; Community metadata; Domain metadata

Citation: Greenberg, J., McClellan, S., Rauch, C., et al.: Building community consensus for scientific metadata with YAMZ. Data Intelligence 5(1), 242-260 (2023). doi: doi.org/10.1162/dint_a_00211
Submitted: January 10, 2023; Received: February 20, 2023; Accepted: February 24, 2023

ABSTRACT

This paper reports on a demonstration of YAMZ (Yet Another Metadata Zoo) as a mechanism for building community consensus around metadata terms. The demonstration is motivated by the complexity of the metadata standards environment and the need for more user-friendly approaches for researchers to achieve vocabulary consensus. The paper reviews a series of metadata standardization challenges, explores crowdsourcing factors that offer possible solutions, and introduces the YAMZ system. A YAMZ demonstration is presented with members of the Toberer materials science laboratory at the Colorado School of Mines, where there is a need to confirm and maintain a shared understanding for the vocabulary supporting research documentation, data management, and their larger metadata infrastructure. The demonstration involves three key steps: 1) Sampling terms for the demonstration, 2) Engaging graduate student researchers in the demonstration, and 3) Reflecting on the demonstration. The results of these steps, including examples of the dialog provenance among lab members and voting, show the ease with YAMZ can facilitate building metadata vocabulary consensus. The conclusion discusses implications and highlights next steps.

† Corresponding author: Jane Greenberg (E-mail: jg3243@drexel.edu; ORCID: 0000-0001-7819-5360).

1. INTRODUCTION

Shared vocabulary is critical to intelligent communication between both humans and machines and machine-to-machine. This fundamental premise along with increased interest in data-driven research has motivated metadata standardization across many scientific and scholarly research domains. These developments have, in turn, resulted in a myriad of metadata standards ranging from domain specific data dictionaries and metadata guidelines to full-level schema instances shaped by constraints. The scope of metadata standards further includes ontologies, taxonomies, name authority files, and other types of knowledge organization systems (KOS) serving as value systems [1]. All of these systems combined form a rich metadata ecosystem that is a significant part of the current cyberinfrastructure.

Today's metadata ecosystem is further supported by national and global agencies, such as the International Standards Organization (ISO), International Electrotechnical Commission (IEC), Institute of Electrical and Electronics Engineers (IEEE), and National Information Standard Organization (NISO). These agencies coordinate standardization processes, including community endorsement for metadata guidelines, schemes, and models. Moreover, the processes these agencies oversee have been critical for sharing metadata vocabularies and, most recently, supporting the FAIR (findable, accessible, interoperable and reusable) data principles [2]. While these agencies have been critical to advancing metadata standards, the processes are not without challenges that seek innovative solutions.

A chief challenge is the arduous nature of the metadata standardization process, which generally requires multiple rounds of committee drafting, review, debate, revision, and voting. The process is made even more complex by the documentation requirements. These aspects present a barrier for smaller scientific communities and groups that lack resources and metadata expertise necessary to drive a full-level standardization process. In some respects, the metadata standardization process may even inhibit scientific progress, as researchers may spend a significant amount of time seeking community consensus and formal endorsement for a new standard at the expense of pursuing their science. These challenges underscore the need for a low-barrier, user-friendly means to building vocabulary consensus supporting metadata standardization. This need is most acute in academic research laboratories where student researchers rotate over the years and time constraints, as well as the absence of in-house metadata expertise, hinder metadata standardization goals. This particular scenario motivates our work with the YAMZ system (YAMZ, pronounced "yams"—suggests "yet another metadata zoo").

This paper reports on a demonstration of YAMZ as a mechanism for building community consensus in a scientific research laboratory. The present use case involves researchers in the Toberer materials science laboratory at Colorado School of Mines, where there is a need to confirm and maintain a shared understanding of the vocabulary supporting research documentation, data management, and the overall metadata infrastructure. The remainder of this paper is organized as follows. The background section reviews metadata standardization challenges and crowdsourcing factors that point toward possible solutions. Next, the YAMZ system is introduced and the uses case design is reviewed, followed by a report on the demonstration. Finally, we discuss the implications of YAMZ and highlight next steps.

2. METADATA STANDARDIZATION CHALLENGES

Metadata standardization challenges stem from a variety of factors spanning: 1) vocabulary as situated in language, 2) social factors, and 3) standards requirements and protocols. As part of exploring YAMZ, it is important to have an understanding of how these factors impact metadata standardization.

2.1 Vocabulary as Situated in Language

The most obvious challenges come from vocabulary as a component of language. Languages are complex communication systems. At the foundational level every language has many ambiguities, idiosyncrasies, inconsistencies, and nuances—all of which are further impacted by the application of grammatical rules. Every language has a vocabulary that is essentially a knowledge structure of terms representing concepts. Vocabulary terms are a chief source for naming of metadata properties; and the metadata property name generally provides the public-interfacing metadata label for humans and, at times, machines. Common challenges occur when deciding on the name (label) of a metadata property due to synonymous and homonymous terms, singular or plural word forms, lexical and dialectical variants, and an array of word forms (e.g. hyphenated, compound, or bound concept). All of these factors impact the convention for naming metadata properties. Finally, these challenges help explain why best practices include the assignment of a unique persistent identifier (PID) for each metadata property name to ensure unambiguous contextual understanding and machine readability.

2.2 Social Factors

Social factors interconnect to language and vocabulary, and also contribute to metadata standardization challenges. Standards development generally involves a committee of people who are incentivized by the need for a standard. Committee members may display some homogeneity, but differences can span academic and professional training, as well as geo-social and cultural experiences. In fact, formalized metadata standards activities may even seek more professionally and geographically diverse participation given the goals of FAIR data, particularly enabling greater interoperability. As a result, different cultural norms, dialectal preferences, and terminological biases are revealed during the standardization process. These factors may, in turn, present challenges; although, the results are quite rewarding when consensus is reached.

2.3 Standards Requirements and Protocols

The ISO, IEC, IEEE, NISO, and other agencies each have a set of requirements and protocols that guide the standardization process. While a thorough review of these details is well beyond the scope of this article, one can gain insight here by reviewing a few examples. A full-fledged ISO standardization project involves six stages: 1) Proposal stage, 2) Preparatory stage, 3) Committee stage, 4) Enquiry stage, 5) Approval stage and 6) Publication stage [3]. Of these, only three stages are obligatory (Proposal stage, Enquiry stage, and Publication). Limiting the process to only obligatory stages can reduce the workload, although each stage still requires time for a review. The ISO provides web access to multiple forms and templates to help

guide users through each stage. Each ISO deliverable is assigned to a standards development track. The development track, in turn, determines the timeframe from the Proposal stage to Publication stage, which can run 18, 24, or 36 months [3]. NISO has a similar protocol, which is outlined in an interactive process diagram covering the following seven steps: Idea, Proposal, Peoples, Process (voting and commenting), Publication, Maintenance, and Revision [4]. Documentation is made accessible via Dropbox with links to templates, as well as a full set of videos that offer training about standards and guide the process. The top of the process graphic (Figure 1, [4]) highlights the fact that standards development is voluntary, that most of the activity takes place out of sight—similar to the invisible part of an iceberg, and that the process to formal approval and publication can run from six months to five years.

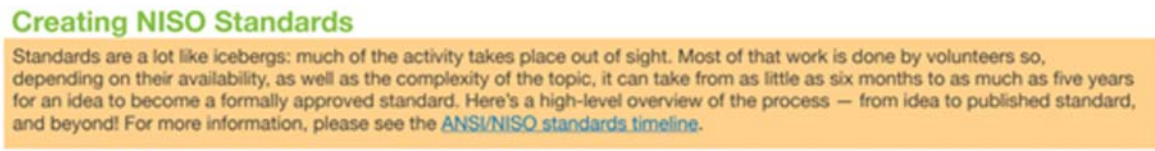

Figure 1. Creating Metadata Standards.

This time duration is also restated in the "ANSI/NISO Standards Development Timeline: Timeline for Formalizing a NISO Standard" graphic [5], which further notes that NISO is accredited by the American National Standards Institute (ANSI), and that ANSI approves all formal standards. Finally, in addition to agency requirements and protocols, there is a large and growing body of standards that guide in naming, registering (e.g., ISO/IEC 11179 "Metadata Registry (MDR)"),[6] assisting in establishing semantics relationships (e.g., ANSI/NISO Z39.19-2005 (R2010) "Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies" [7]), and machine readability and interoperability (e.g. the Resource Description Framework (RDF), Simple Knowledge Organization System (SKOS) and Web Ontology Language (OWL). These examples provide insight into a larger fabric that has allowed the metadata ecosystem to accelerate over the last few decades.

While the varied components concerning standards formation discussed above have been key to progress, it is understandable that a research scientist initially grappling with metadata vocabulary standardization may find these added layers quite complex, and may be inhibited from engaging in metadata standardization, despite awareness of the value of metadata standards. Clearly, alternative approaches to metadata standards development are needed if the process seeks to more broadly engage researchers seeking to use them. To this end, crowdsourcing platforms, as discussed in the next section, provide context for addressing this challenge.

3. CROWDSOURCING: A POTENTIAL APPROACH METADATA STANDARDS DEVELOPMENT

Crowdsourcing as a phenomenon has its foundations in the notion of "wisdom of the crowds." The term was officially coined in 2006 by Jeff Howe and Mark Roberson, two Wired magazine editors [8], although the essence of crowdsourcing has been an approach to problem solving for centuries [9]. The key

characteristic is the open call to a network of people to solve a problem. The central thesis is that the collective knowledge of a group of individuals is superior to that of a single individual.

Crowdsourcing has thrived with the evolution of the Internet. Examples such as Reddit, Wikipedia, Wikidata, and Stack Overflow are seen as reliable information sources [10, 11]. These and other crowdsourced projects help to build community-driven knowledge bases. Indeed, crowdsourcing systems are not without flaws. For example, research on Wikipedia points to gender bias [12], cultural bias [13, 14], and non-experts providing unreliable and even problematic answers to health-related crowdsource venues [15, 16]. There is pushback, however, as researchers have also developed methods to address such challenges and demonstrate where crowdsourcing contributions of non-experts can outperform those of experts [15] by providing a swift and informative response.

At the intersection of crowdsourcing, the Semantic Web, and MediaWiki, is semantic wiki technology, conceptualized and initiated between 2004 and 2005 [17, 18, 19]. Over the last few decades, semantic wiki technology inspired a number of spinoff technologies of which Freebase was among the most well-known. Freebase was shut down by Google in 2016[20], and the data is now available through Wikidata, which uses Semantic MediaWiki to add semantic functionality to the MediaWiki platform. Specifically, this technology lets a user embed structured data about entities, such as people, places, and events to support queries. Specific to materials science was the Knowledge Wiki project's "mv" prefix (matvocab.org [21]), which used the resource description framework (RDF) and other Semantic MediaWiki features, but this initiative is no longer operational. Overall, semantic wiki technology has had great appeal, but requires a learning curve and technical capacity to encode and work with structured entities. This aspect as well as funding likely led to the closure of the Knowledge Wiki project. Indeed, the full scope of challenges is difficult to gauge without a record, although recognition of noted challenges has, in part, motivated our work with YAMZ, (Yet Another Metadata Zoo) specifically through a partnership with the Metadata Research Center (MRC), Drexel University, and materials science researchers connected with the NSF supported Institute of Data Driven Dynamical Design (ID4).

4. YAMZ FRAMEWORK AND FUNCTIONALITIES

YAMZ presents a community-driven approach for collectively building a vocabulary. Frequently described as a metadata dictionary, YAMZ supports discussion, divergence, vocabulary growth, and consensus on individual terms. In contrast, Wikidata is more of a registration service for terms that are stable. YAMZ offers a framework and an underlying technology created in the context of Earth Science by the NSF DataONE Metadata Working Group (initially under the name Sea Ice) [22, 23, 24]. It is currently overseen by the MRC with a connection to the ARK (Archival Resource Key) Alliance [25]. Although YAMZ development preceded the launch of the FAIR principles, the system lends support for community consensus building and the machine-actionability of FAIR[26].

YAMZ can be used for terms from any and all domains. It can be used by researchers working on their own or by self-designated sub-communities that share interests around a project or a domain. Users

exploring YAMZ may find registered terms that suit their purposes, without needing to login to the system. Users sign in to YAMZ with their Gmail account and, once authenticated, they can elect to test, comment on, and endorse terms that apply to their work. Authenticated users can also contribute terms, which are publicly visible, to share with their colleagues. YAMZ was designed to make it easy for people to add missing terms that are essential to their community. Each term is assigned a unique ARK persistent identifier for precise long-term reference (this is especially helpful for terms that have the same spelling). Finally, individual communities can assign tags that mark their own terms. A new feature also permits a community to bulk import terms and tag their community terms. Figure 2 presents the homepage of a logged in YAMZ contributor. The user is notified when a term they contributed is voted on or watched. The "My terms" section is a list of all terms contributed by the current user. The "Watching" list shows the terms that the current user, in this case a material scientist, contributed to and that the user is currently watching. The user will receive alerts concerning any commenting if they opted in.

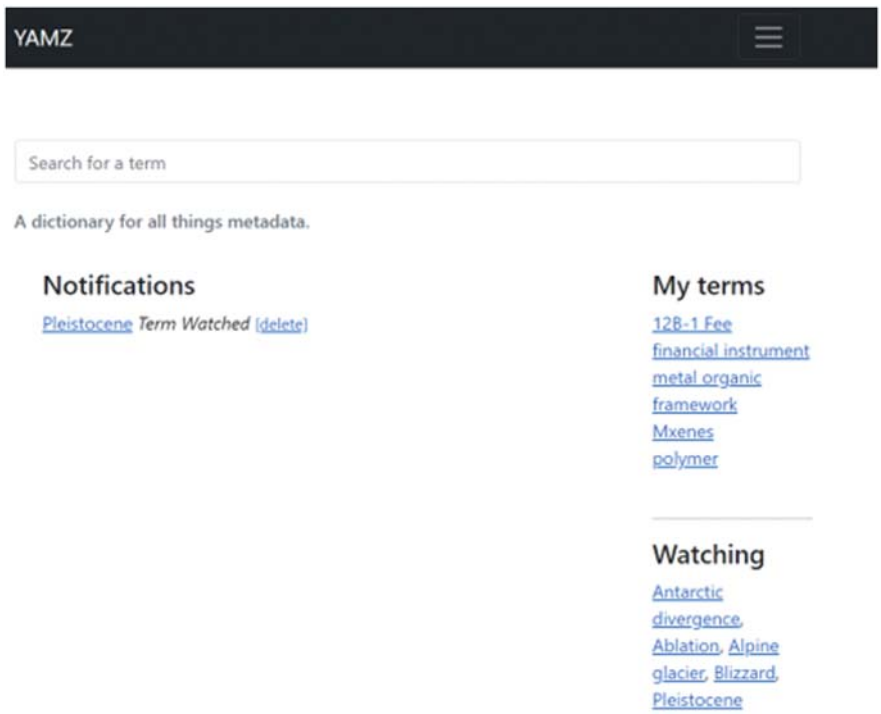

Figure 2. Profile for YAMZ User in Materials Science.

YAMZ currently includes close to 5000 terms and over 160 users across earth science, materials science, biology, humanities, and other disciplines. A consensus-based heuristic is applied to help general users, as well as specific communities, find agreement on terms and their meanings. Users can vote a term up or down to determine consensus, which is tallied to reflect the relative acceptance of the proposed term's definition and usage examples. The ranking of a term can help community members cohere around terminology and definitions that can form the basis of a metadata standard.

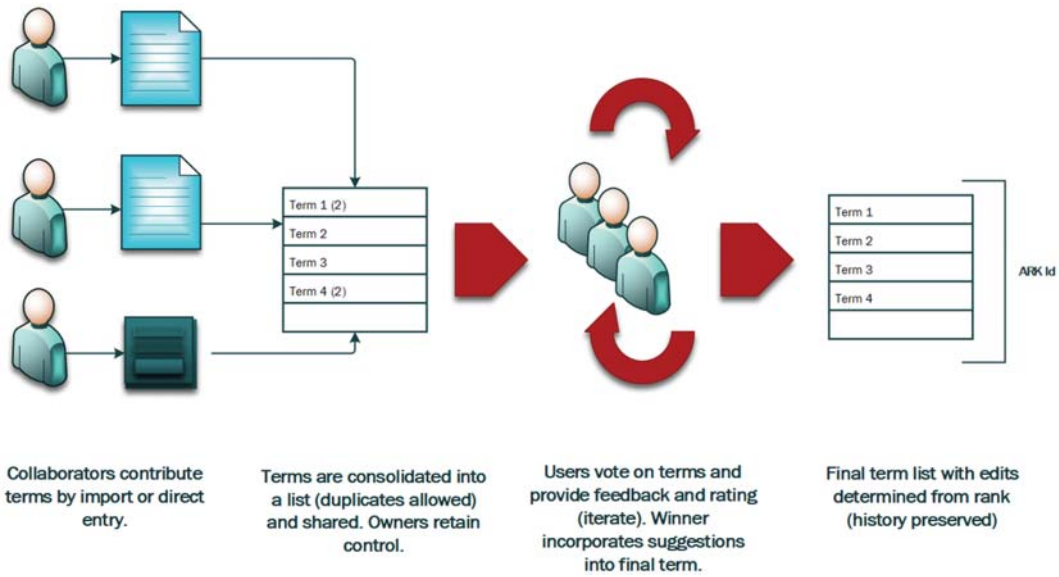

Figure 3. Creating Metadata Standards.

The YAMZ general consensus-building workflow follows four high-level sequential steps presented in Figure 3: 1) First, collaborators contribute entries by direct entry into an HTML form, or by uploading a structured document (e.g., CSV file, or tab delimited file); 2) Second, each new term is tagged by the community collaborators and receives an ARK identifier. The ARK is assigned whether or not the term contributed becomes the term endorsed by community consensus; 3) Third users (e.g. in a lab) will see each other's terms and have the opportunity to comment and vote on the definitions of the terms they favor (terms can evolve through iterative rounds of user feedback and contributor edits); 4) The final step is determining the preferred term. Here, the highest scored term becomes the default reference for the lab, however the other contributed terms are still available for linking in a document via ARK ID. A final point to make here is that YAMZ may have several instances of a term, each registered by various individuals who may, or may not, be part of the same community. YAMZ allows for the various instances and their definitions with the distinction made by the endorsed definition and the ARK.

5. DEMONSTRATION OF YAMZ

The YAMZ demonstration reported here involved the following three phases: 1) Sampling terms for the demonstration, 2) Engaging graduate student researchers in the demonstration and 3) Reflecting on the demonstration. The phases are reported here.

5.1 Phase 1: Sampling Terms for the Demonstration

The sampling was conducted to identify candidate terms that researchers could submit to YAMZ. We gathered laboratory procedure documentation specific to synthesizing and analyzing thermoelectric

materials. The documents were preprocessed, numerical expressions, as well as ngrams that ranged from 1 to 3 word segments, were removed, and the corpus was processed using Orange data mining application to produce a "word cloud" ranking terms based on frequency, and "keyword extraction" based TF-IDF metrics. The top 20 words were selected as candidate terms, from which "graphite foil spacers" and "grit" were randomly selected, and the term "melt" was selected based on discussions with lab members.

5.2 Phase 2: Engaging Graduate Student Researchers in the Demonstration

This phase involved distributing the sample of three terms (graphite foil spacers, grit, and melt) to the three graduate researchers from the Toberer Lab at the Colorado School of Mines specializing in thermoelectric materials. Prior to the demonstration, three materials science graduate students set up YAMZ logins. For the sake of reporting, we call these participants LM 1, LM 2, and LM 3 where "LM" denotes "Lab Member."

As a first step in this phase, each lab member contributed a definition for a term of their choice. The lab members were allowed to define the same term string). After each term was entered, the LMs explored the voting and commenting functions of YAMZ. The first term chosen to collectively explore was "melt," which refers to a common laboratory practice in the Toberer Lab. Figure 4 shows the initial entry by LM 1. The initial definition given is, "A way of inducing a solid-to-liquid phase transition by applying heat to material" and the tag of AMS is applied for the materials science group.

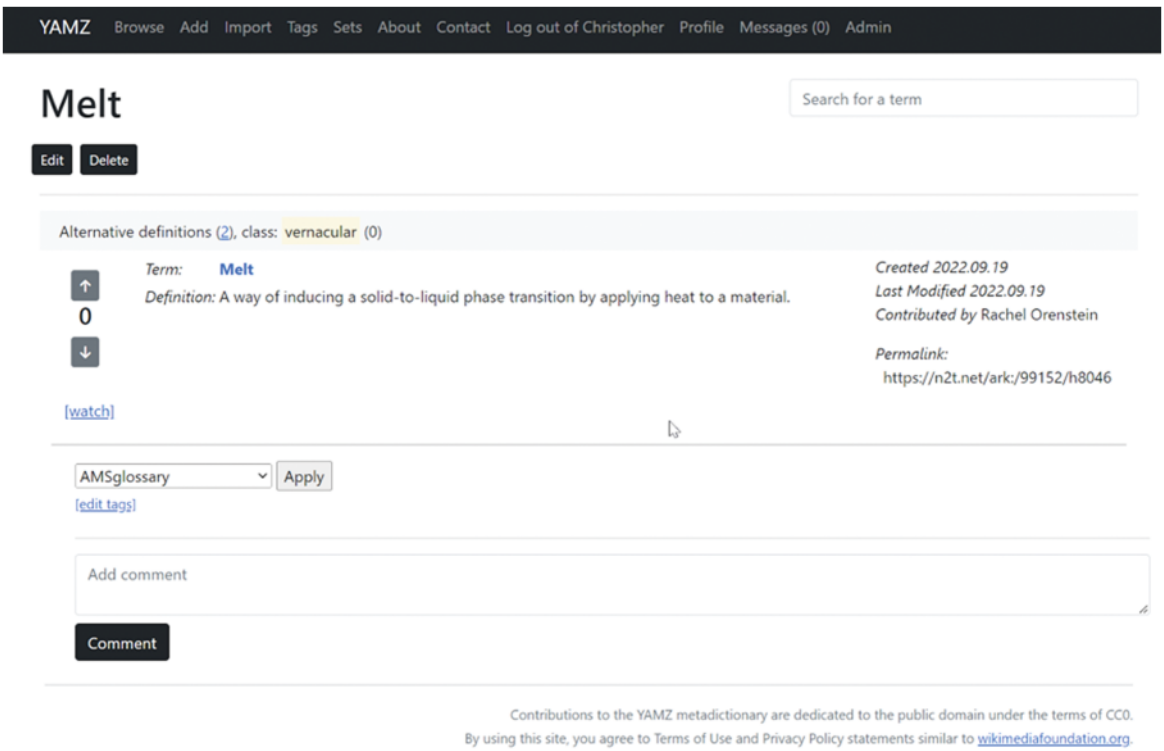

Figure 4. Initial Entry for the term "melt".

After this definition was entered, LM 2 reviewed the definition and added the comment, "Melt may also refer to a material in a fully liquid state," as illustrated in Figure 5. Figure 5 also shows a follow-up comment by LM 1, who states, "Vanessa, agreed," and, further that LM 1 modified the original definition, "Melt may also refer to a material in a fully liquid state," to "A way of inducing a solid-to-liquid phase transition by applying heat to a material." This stands as the final definition for this entry of "melt", identified by the ARK ending in "h8046".

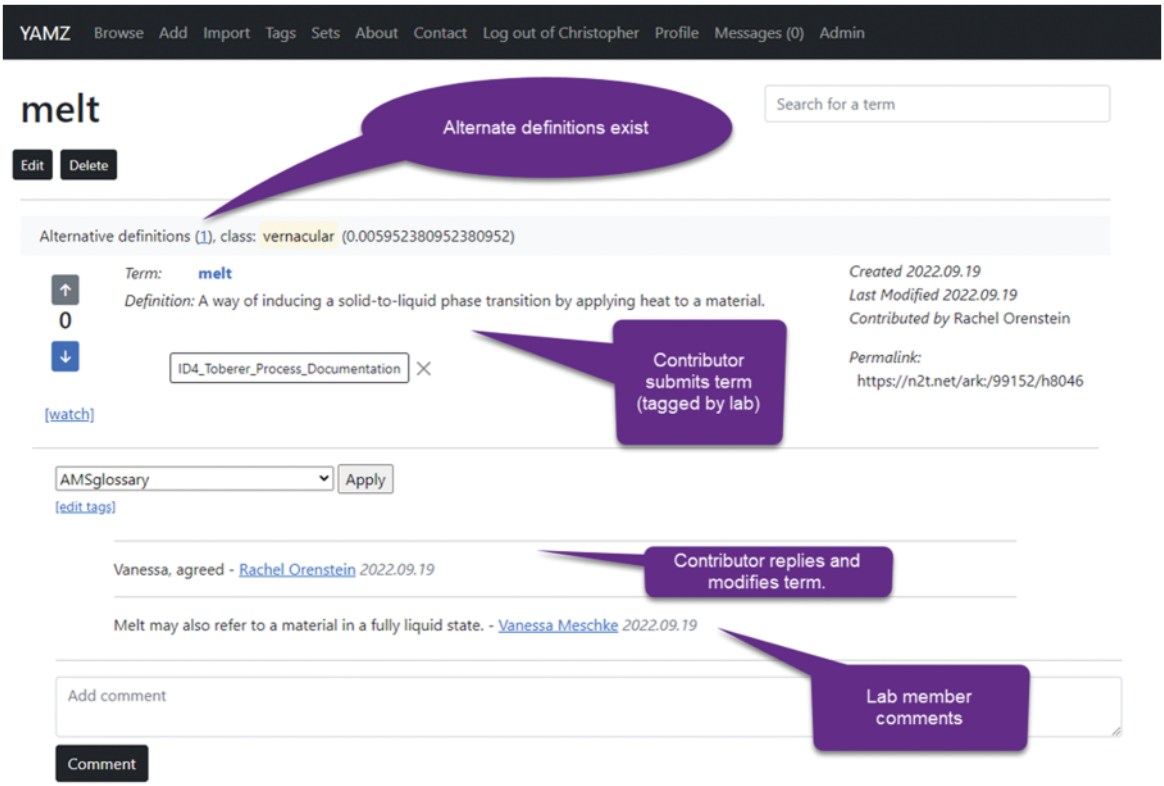

Figure 5. Commenting on an initial entry for the term "melt".

Figure 5 reveals another entry, as indicated by the "Alternative definition," preceding a hyperlinked number, directly under the "Edit" and "Delete" button below the word "melt." As indicated here, there are separate entries for "melt". Comments were added to the first instance of "melt," providing more context and showing some agreement among the lab members. Both entries for "melt" are captured in Figure 6. On the top half of Figure 6, we have the entry that LM 1 modified after LM 2 made some suggestions (for the ARK ending in "h8046"). The lower half of Figure 6 captures the alternative entry identified by ARK "h8057," with the definition given as, "verb: induce a phase transition from solid to liquid via heat or pressure; noun: a material in the liquid phase." This entry also includes verb and noun examples.

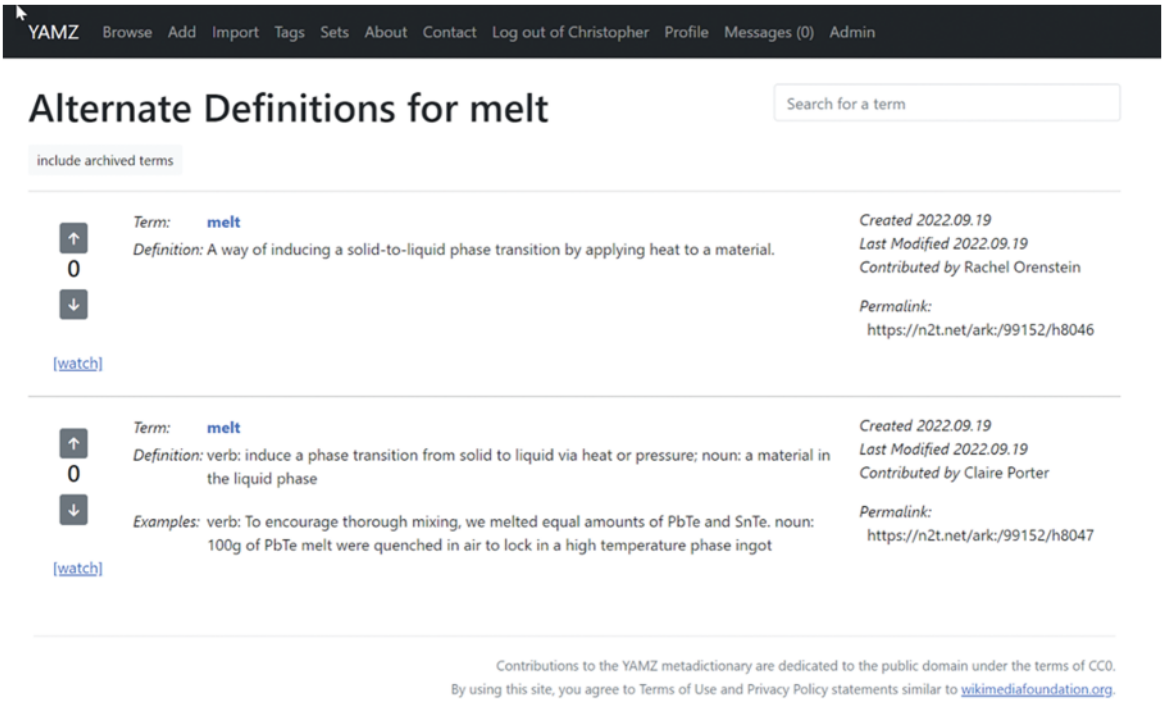

Figure 6. Two entries for the word "melt": ARK ID h8046 and h8046.

The different entries for the word "melt"present a compelling case for YAMZ and consensus building. Figure 7 provides a simple example of how YAMZ supports voting. In this example, a positive score of 2 for h8046 demonstrates consensus and a preference over the -1 score for h8047. The hyperlink to the word "watch" also indicates that the provenance of these YAMZ entries can be followed by anyone interested. The voting mechanism and collection of multiple entries and comments would have this same pattern but be more extensive if a group of academic research lab members worked with YAMZ to build a metadata standard. This demonstration took place over an hour, with some initial preparation; however, a more full-on effort would likely run over the course of a month or two, so lab members would have time to reflect and contribute.

Figure 8 and Figure 9 present the initial entry, commenting activity, and revision for the terms "grit" and "graphite foil spacers", respectively. Both figures capture a rich dialog and show how YAMZ can enable lab members to work together to standardize terms that are important to their research efforts.

The final component of the demonstration is captured in Figure 10, which includes the overall voting of LM 1, LM 2, and LM 3, for the three candidate terms that were entered. These scores reflect the consensus-based heuristic for the entries. It is important to emphasize that in this demonstration the terms are all listed as being in the vernacular class, which means they can still evolve. With more time for dialog, a stable consensus can emerge, and a YAMZ term can move to the canonical class, where they are officially

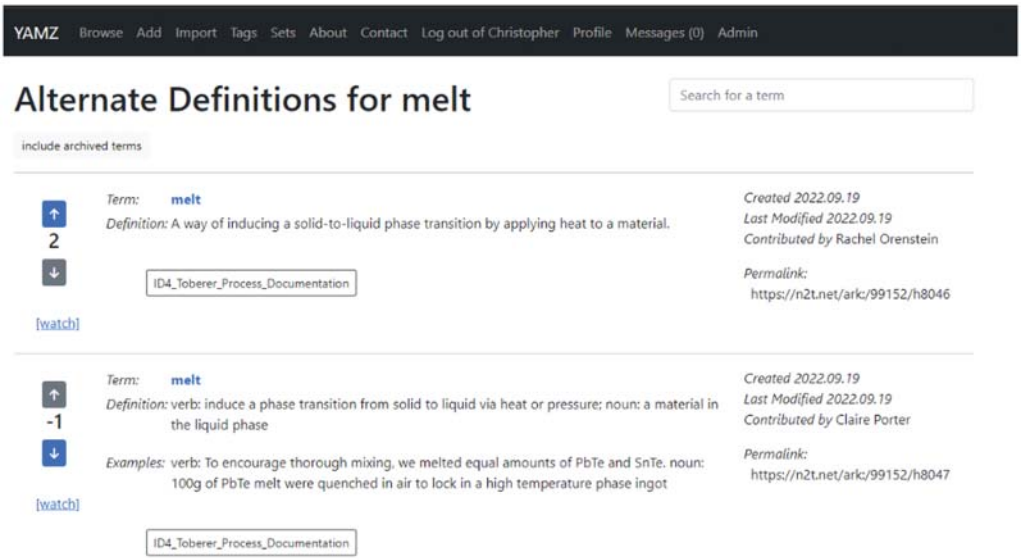

Figure 7. Voting Demonstration for different entries of "Melt".

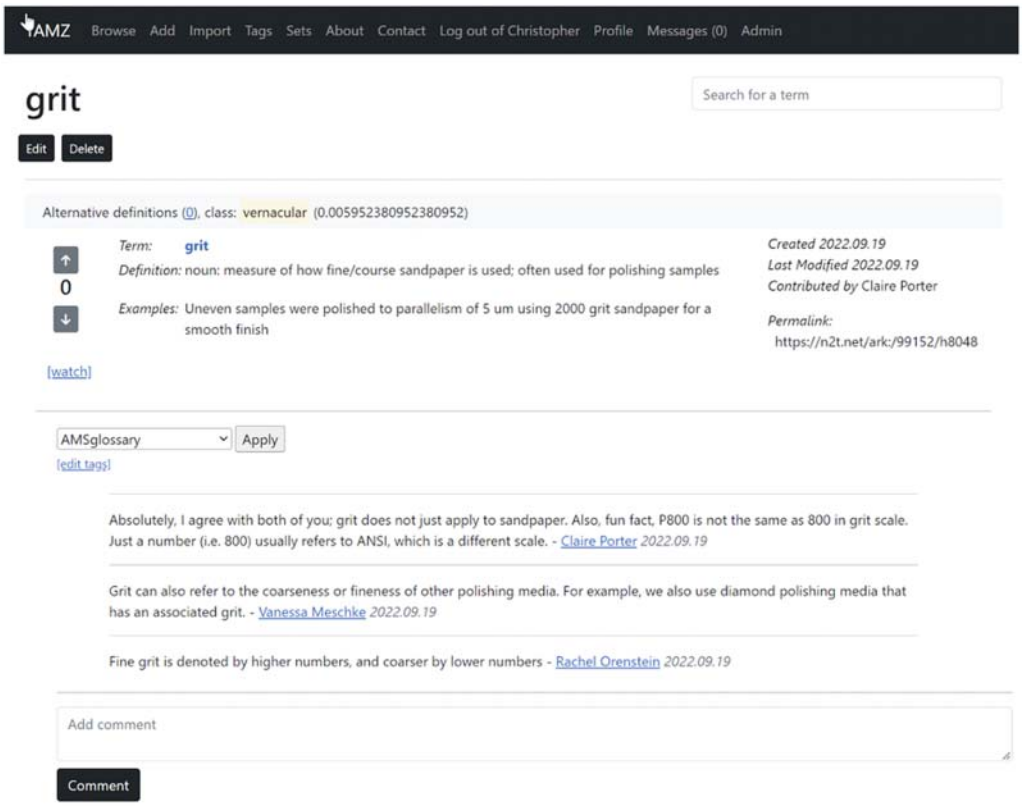

Figure 8. YAMZ demonstration for "Grit".

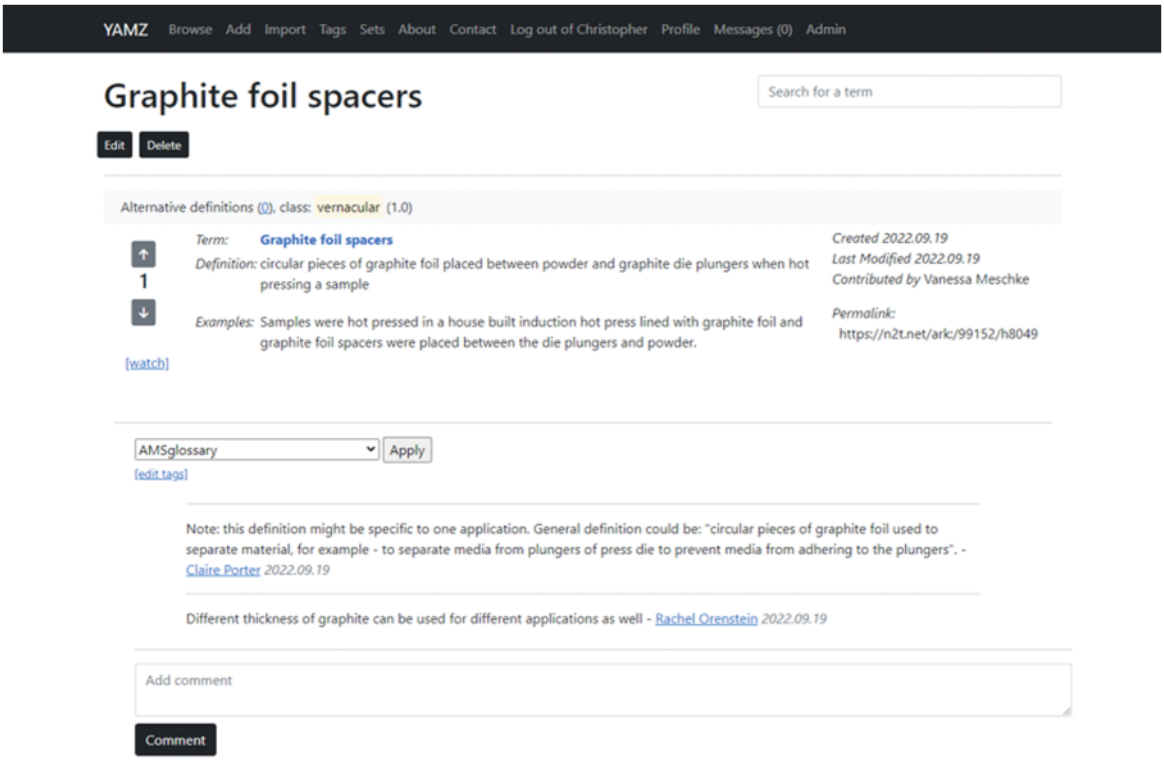

Figure 9. YAMZ demonstration for "Graphite foil spacers".

endorsed. Terms that enter the canonical class may still expect to be deprecated in the distant future because language is a living thing. Nonetheless, it is important for deprecated terms in YAMZ to persist into that future to aid in interpreting historical documents and data. Even if terms do not become canonical, they remain in the vernacular indefinitely (unless the contributor removes them).

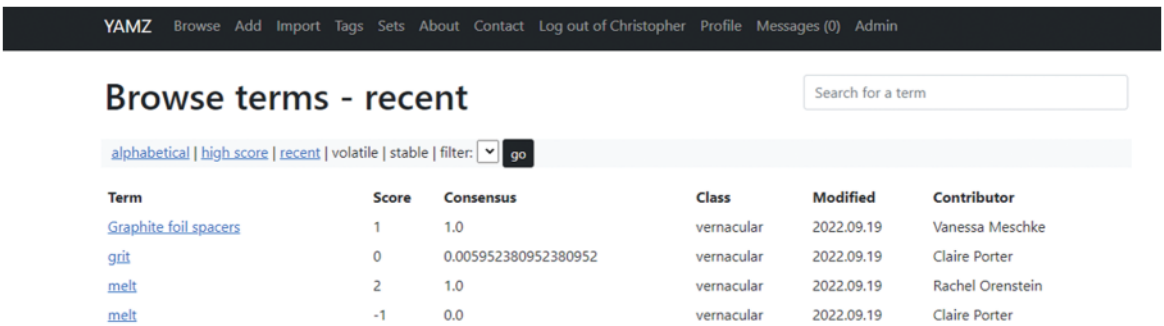

Figure 10. Final Result of Consensus Building Demonstration in YAMZ.

After each term was entered, participants tested the voting and commenting functions of YAMZ. As this was an informal demonstration of the YAMZ platform, participants discussed with researchers some of their experiences with the researchers.

5.3 Phase 3: Reflecting on the Demonstration

The last phase of the demonstration involved LM 1, LM 2, and LM 3 from the Toberer lab reflecting on the YAMZ demonstration together with MRC team members. The conversation pointed to the positive, user-friendly design of YAMZ. The use of ARKS as a PID was also viewed positively as it permits access to a stable entry for a particular definition. The commenting provenance recorded in reverse chronological order was, however, seen as made aspects of the demonstration difficult to follow. Specifically, displaying the most recent comment first was seen as a bit confusing.

The relationship between expertise and voting was raised. Stack Overflow was brought up as a model where more weight is given to participants who provide useful feedback more frequently. Earlier versions of YAMZ had considered this heuristic [22]; and it could be that a senior researcher or lab director's vote could have more weight compared to a novice or early-stage researcher. This proposition raises a set of interesting questions about how expertise is determined. Another consideration is that often senior members may not have the time to engage in a YAMZ-like activity, or be more removed from the day-to-day research. These are all issues to be considered in future development of YAMZ.

5.4 Conclusion: Implications and Next Steps

The metadata standardization process, overseen by key agencies, has greatly contributed to today's rich metadata ecosystem of standards. While there are many positive outcomes, there are also challenges for researchers who have time constraints and lack the metadata expertise necessary to participate and contribute. The complexity of the metadata standards environment points to the need for more user-friendly approaches for generating metadata standards—whether local project or institutional standards, or broad international standards. This is particularly true for those working in academic research labs who wish to adhere to FAIR principles and generate high quality metadata. This need is even more profound with the increased interest in data-driven research and AI.

YAMZ presents a framework and a technology that can assist researchers in consensus building, which is foundational in developing a metadata standard. This paper reports on a YAMZ demonstration with members of an academic research laboratory. The demonstration involved three key phases: 1) sampling terms for the demonstration, 2) engaging graduate student researchers in the demonstration. and 3) reflecting on the demonstration. The results, including records of the voting and dialog among lab members, show the ease with which YAMZ can facilitate consensus.

The work described in this article involved three lab members exploring YAMZ with three candidate terms. The exploration was conducted over a one-hour time period and serves, simply, as demonstration. A next step includes implementing a full-scale study, whereby users will work YAMZ over close to a month's

time and engage in formal voting and endorsement to identify core set of terms. Future work is also needed to address aspects revealed in the reflection, primarily interface design clarifying the reverse chronological ordering of the comments, and providing flexible display options. Research is also needed to explore reputation mechanisms and term transitions. Furthermore, MRC team members need to work with participating research scientists to evaluate the approach in the context of FAIR. Overall, the demonstration shows YAMZ to be an innovative, low bar system supporting consensus building around vocabulary integral to metadata standard development. The demonstration shows how a crowdsourced approach may help academic research labs advance their metadata activities. Finally, the YAMZ approach may have farther reaching implication offering an alternative or complementary approach to the formal standardization process.

ACKNOWLEDGEMENTS

The demonstration work reported here is supported by National Science Foundation-Office of Advance Cyberinfrastructure (NFS-OAC) 2118201, the Ronin Institute/U.S. Research Data Alliance (RDA), and the Institute of Museum and Library Services (IMLS) RE-246450-OLS-20.

REFERENCES

[1] Riley, J.: Understanding metadata. Washington DC, United States: National Information Standards Organization 23, 7–10 (2017)
[2] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 1–9 (2016)
[3] Stages and resources for standards development. Available at: https://www.iso.org/stages-and-resources-for-standards-development.html. Accessed on February 2, 2023
[4] Creating NISO Standards. Available at: https //www.dropbox.com/s/p5yinoc7xooj6gw/Creating%20NISO%20Standards.pdf?dl=0. Accessed on February 2, 2023
[5] ANSINISO Standards Development Timeline Timeline for Formalizing a NISO Standard. Available at: https //www.niso.org/standards-timeline. Accessed on February 2, 2023
[6] ISO/IEC JTC1 SC32. ISO/IEC 11179, Information Technology – Metadata registries (MDR) (2007)
[7] ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Available at: https://www.niso.org/publications/ansiniso-z3919-2005-r2010. Accessed on February 3, 2023
[8] Howe, J., Others: The rise of crowdsourcing. Wired Magazine 14(6), 1–4 (2006)
[9] Kietzmann, J.H.: Crowdsourcing A revised definition and introduction to new research. Business Horizons 60(2), 151–153 (2017)
[10] Aniche, M., Treude, C., Steinmacher, I., Wiese, I., Pinto, G., Storey, M.A., & Gerosa, M.A.: How modern news aggregators help development communities shape and share knowledge. In Proceedings of the 40th International conference on software engineering, pp. 499–510 (2018, May)
[11] Wazny, K.: "Crowdsourcing" ten years in A review. Journal of Global Health 7(2) (2017)
[12] Bjork-James, C.: New maps for an inclusive Wikipedia: decolonial scholarship and strategies to counter systemic bias. New Review of Hypermedia and Multimedia 27(3), 207–228 (2021)

[13] Greenstein, S., Devereux, M.: Wikipedia in the Spotlight. Kellogg School of Management Cases, 1–18 (2017)
[14] Callahan, E.S., Herring, S.C.: Cultural bias in Wikipedia content on famous persons. Journal of the American Society for Information Science and Technology 62(10), 1899–1915 (2011)
[15] Beck, S., Brasseur, T.M., Poetz, M., et al.: Crowdsourcing research questions in science. Research Policy 51(4), 104491 (2022)
[16] Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)
[17] Tazzoli, R., Castagna, P., Campanini, S.E.: Towards a semantic wiki wiki web. In 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan (2004, November)
[18] Krötzsch, M., Vrandecic, D., Völkel, M.: Wikipedia and the Semantic Web. The Missing Links, WikiMania (2005)
[19] Schaffert, S., Gruber, A., Westenthaler, R.: A semantic wiki for collaborative knowledge formation. na. (2005).
[20] Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., Pintscher, L.: From freebase to wikidata: The great migration. In Proceedings of the 25th International Conference on World Wide Web, pp. 1419–1428 (2016, April)
[21] Jaykumar, N., Yallamelli, P., Nguyen, V., et al.: KnowledgeWiki An OpenSource Tool for Creating Community-Curated Vocabulary, with a Use Case in Materials Science (2016)
[22] Greenberg, J., Murillo, A., Kunze, J., et al.: Metadictionary: advocating for a community-driven metadata vocabulary application. DC-2013, Lisbon, Portugal (2014)
[23] Kunze, J., Greenberg, J., Callaghan, S., Guralnick, R., Janee, G., Murillo, A., ... & Robertson, T.: SeaIce a Cross-Domain Crowd-Sourced Metadata Vocabulary. In: TDWG 2013 Annual Conference (2013)
[24] DataOne https //notebooks.dataone.org/author/cjpatton/
[25] https //arks.org/about/ retrieved on Feb 20th 2023
[26] Rauch, C.B., Kelly, M., Kunze, J.A., et al.: FAIR Metadata: A community-driven vocabulary application. In: Metadata and Semantic Research: 15th International Conference, Revised Selected Papers. pp. 187–198 (2022)

AUTHOR BIOGRAPHY

Jane Gtreenberg, Ph.D. is the Alice B. Koreger Professor at Drexel University's College of Computing and Informatics. She is also the director of the Metadata Research Center, Drexel University. Her research and teaching focus in the areas of automatic metadata generation, big metadata, knowledge organization systems (KOS), linked data/Semantic Web, ontologies, information economics, and data sharing/open data.

Scott McClellan is a second year PhD student in the College of Computing and Informatics at Drexel University. He is currently performing research on controlled vocabularies and ontologies with special focus on the materials science space. Prior education experience includes an MSI in Library and Information Science, also from Drexel University.

Christopher Rauch is a third year PhD student at the College of Computing and Informatics at Drexel University with a J.D. from Rutgers University. His research focuses on the personalization of recommender systems.

Xintong Zhao is a PhD candidate in the College of Computing and Informatics at Drexel University. She is broadly interested in Text Mining, Big Data Analytics and Data-Driven Knowledge Discovery. Her research contributes to advancing the discovery of valuable information across large amounts of unstructured data and facilitates decision making.

Dr. **Mat Kelly** is an assistant professor at Drexel University's College of Computing and Informatics. He holds a Ph.D. in Computer Science from Old Dominion University. His research focuses on digital preservation with an emphasis on the computational aspects of personal and private web archiving. He has also published in the areas of persistent identification, metadata, linked data, semantic analysis, scientometrics, and information visualization. More information can be found at https://matkelly.com.

Dr. **Yuan An** is an associate professor at Drexel University's College of Computing and Informatics. Central to his research interest is to discover the semantic relationships between the items in different data sources. He has applied advanced data analytics methods to different domains including healthcare, biomedicine, and materials science. He is currently working on knowledge graph augmented materials discovery.

John Kunze is a pioneer in the theory and practice of digital libraries. With a background in computer science and mathematics, he wrote BSD Unix tools that come pre-installed with Mac and Linux systems. He created the ARK identifier scheme (arks.org), the N2T.net scheme-agnostic resolver, and contributed heavily to the first standards for URLs (RFC1736, RFC1625, RFC2056), library search and retrieval (Z39.50), archival transfer (BagIt - RFC8493), web archiving (WARC), and metadata (RFC2413, RFC2731, ANSI/NISO Z39.85).

Rachel is a 3rd year PhD student in the Toberer group at the Colorado School of Mines. Her background is in phase boundary mapping and analysis of bipolar electronic and thermal properties of thermoelectric materials.

Claire Porter is a fourth year PhD student in the Toberer lab at Colorado School of Mines. She specializes in synthesizing semiconductors for thermoelectric applications, as well as designing instruments to explore what limits electron mobility in these materials.

Vanessa is a PhD candidate in the Toberer group at the Colorado School of Mines. Her work has focused on exploring the electronic properties of thermoelectric materials and automating data and metadata collection in the lab.

Eric Toberer is a Professor of Physics at the Colorado School of Mines. His work focuses on solid state materials for energy, including thermoelectrics, photovoltaics, and optoelectronics.