

Threshold Determination and Engaging Materials Scientists in Ontology Design

Jane Greenberg¹(✉), Yue Zhang¹, Adrian Ogletree¹, Garritt J. Tucker²,
and Daniel Foley²

¹ Metadata Research Center (MRC), College of Computing & Informatics (CCI),
Drexel University, Philadelphia, PA, USA

{janeg, yue.zhang, aogletree}@drexel.edu

² Computational Materials Science and Design (CMSD), Materials Science and Engineering
Department, College of Engineering, Drexel University, Philadelphia, PA, USA
gtucker@coe.drexel.edu, df92@drexel.edu

Abstract. This paper reports on research exploring a threshold for engaging scientists in semantic ontology development. The domain application, *nanocrystalline metals*, was pursued using a multi-method approach involving algorithm comparison, semantic concept/term evaluation, and term sorting. Algorithms from four open source term extraction applications (RAKE, Tagger, Kea, and Maui) were applied to a test corpus of preprint abstracts from the arXiv repository. Materials scientists identified 92 terms for ontology inclusion from a combined set of 228 unique terms, and the term sorting activity resulted in 9 top nodes. The combined methods were successful in engaging domain scientists in ontology design, and give a threshold capacity measure (*threshold acceptability*) to aid future work. This paper presents the research background and motivation, reviews the methods and procedures, and summarizes the initial results. A discussion explores term sorting approaches and mechanisms for determining thresholds for engaging scientist in semantically-driven ontology design and the concept of ontological empowerment.

Keywords: Nanocrystalline metals · Materials science · Semantic terminology · Ontology design · Ontological empowerment · Threshold determination · Helping Interdisciplinary Vocabulary Engineering (HIVE)

1 Introduction

Vocabularies, taxonomies, and semantic ontological systems have been a mainstay of scientific endeavors from earliest times. Aristotle’s *History of Animals* (*Historia Animalium*) [1] is among the most recognized examples. In this seminal work, animals are classified by observable properties, such as having blood or being bloodless, their living habitat, and movement processes (walking, flying, or swimming). Aristotle further introduced binomial naming; that is, the classing and naming of organisms by their genus and what we today identify as species. During the nineteenth century, Carl Linnaeus, the ‘father of modern taxonomy,’ advanced binomial nomenclature for plant specimens by

© Springer International Publishing Switzerland 2015

E. Garoufallou et al. (Eds.): MTSR 2015, CCIS 544, pp. 39–50, 2015.

DOI: 10.1007/978-3-319-24129-6_4

introducing facets, hierarchies (genus/species, supra- and sub- categories), associations, and other types of relationships that are integral components of many contemporary semantic ontologies [2].

Fast forward to today, where semantic ontologies are being integrated into our digital data infrastructure. Ontologies have a crucial role to play in aiding data discovery, reuse, and interoperability; and, most significantly, they can facilitate *new science* [3]. Development of ontology encoding standards, such as the Web Ontology Language (OWL)[4] and the Simple Knowledge Organizing System (SKOS)[5], are interconnected with the growth of Big Data and the desire to advance data science activity. Additionally, the ‘open data movement’ has motivated various communities to generate and share ontologies; there have been numerous collaborations to this end in biology, medicine and health sciences, ecology, and geology.

Materials science, as an interdisciplinary field of study, has been able to benefit from ontology work in these other disciplines; however, documented efforts targeting materials science are limited to a few examples [6]. Researchers associated with the Materials Metadata Infrastructure Initiative (M^2I^2) [7] at the Metadata Research Center, Drexel University, are addressing this shortcoming by exploring means for advancing ontological practices in the field of materials science. As part of this effort, an interdisciplinary research team of information and materials scientists are studying ways to engage domain scientists in ontology development while extending the Helping Interdisciplinary Vocabulary Engineering (HIVE) technology [8, 9, 10].

This paper reports on the M^2I^2 effort, and specifically the development of an ontology for nanocrystalline metals. The chief goal was to explore a threshold for engaging scientists in semantic ontology development. To further explain, we seek baseline data on the engagement capacities of scientists (domain experts) for aiding ontology development. More precisely, how much time and effort can we anticipate of scientists, without them feeling like ontology work is an intellectual drain.

A secondary goal was to identify means by which information scientists/non-domain experts can easily facilitate ontology design processes. To this end, we identified fairly generic, domain agnostic technologies that can be applied across various materials science sectors as well as other disciplines. We explain these technologies in our methods and reporting. The unified goal is to establish an ontology design framework that can be used across a range of disciplines and sub-disciplines.

The remainder of this paper reports on this research and is organized as follows. Section 2 presents background information on materials science and nanocrystalline metals; Section 3 provides the case for shared semantics in materials science; Sections 4-6 cover the research objectives, methods, and procedures; Section 7 presents the results; Section 8 includes a contextual discussion of the results and examines challenges and opportunities for determining thresholds for engaging materials scientists in ontology design. Section 9, the last section of this paper, presents several conclusions, notes research limitations, and identifies next steps for future research.

2 Materials Science and Engineering: Nanocrystalline Materials

Materials Science and Engineering (MSE) is the study of the intersection of materials' processing, structure, properties, and performance [11]. The goal is to improve existing materials and develop new materials for a myriad of scientific and technological applications. The origins of MSE lie within the overlapping interests of chemistry, physics, and engineering. MSE research is relevant to other engineering and scientific disciplines, as the impact of advanced materials has shown to be universally beneficial. Over the past few decades, one significant driving force behind materials research has been the emergence of nanotechnology and nanoscience [12], where both science and engineering at the atomic/molecular level are investigated. Advancements in structural, electronic, magnetic, optical, and other functional properties of materials have correlated well with advancements in nanotechnology research.

Engineering or manipulating the nanostructure of a material enables enhancement for a wide array of physical properties (e.g., mechanical, electrical, optical, etc. [13]. Nanostructured materials are characterized by the fundamental structure or building block of the *material being* on the order of *nanometers*. Nanocrystalline (NC) metals, a type of nanostructured material, has received noticeable interest due to improvements in its mechanical properties. In NC metals, the length-scale of the fundamental unit (i.e., grain or crystal) is on the order of 1-100 nanometers [14].

NC metals have been the subject of numerous studies, as their mechanical strength has been recorded in early efforts to exceed that of traditional metals with larger grains or crystals. A growing body of research confirms that additional property enhancements in NC metals show promise in more common products and applications. Specific examples demonstrate how NC metals incorporated into artificial limbs may improve human health [15]. Innovative NC driven capacities, the open data movement, and calls to accelerate materials science R&D provoke the development of shared semantics.

3 The Materials Genome Initiative and the Case for Shared Semantics

Materials are integral to our daily lives; and global efforts along with industry/academic partnerships seek to advance MSE R&D. In the United States, the Obama Administration has launched the Materials Genome Initiative (MGI) [16] to accelerate the development of new materials in areas impacting human health, energy and the environment, and social welfare. The MGI 2014 Strategic Plan [17] recognizes the significance of data in 'Objective 3, Facilitate Access to Materials Data;' and Section 3.2 specifically calls for semantics to aid discovery across data repositories.

Semantic ontologies are important for this objective; they aid scientists and data managers in discovering, using, and repurposing research data together with additional components of the research enterprise (e.g., data, models, simulations, instrumentation, software, code, etc.).

Biology, geology, medicine, and environmental science have extensive disciplinary networks of shared semantics ontologies. Two examples include the Biosharing portal [18] in the United Kingdom, which provides links to a vast collection standards including scientific ontologies; and the National Center for Biological Ontologies (NCBO) biportal [19], which houses 441 ontologies covering scientific and medical topics. Federal agencies are also responsible for maintaining terminologies that equate with semantic ontologies. A case in point is the United States Geological Survey (USGS), which maintains the Integrated Taxonomic Information System [20] and other terminologies in the geological and biological sciences. All of these facilities allow scientists to access and use semantics on a global scale. More significantly, sharing semantics supports data discovery, use, integration, and other functionalities that can promote new science.

Shared semantic ontologies have flourished in various scientific domains, although efforts in materials science and engineering (MSE) are limited. One reason for the slow uptake in this area may be that materials science and engineering research endeavors are able to leverage ontologies developed in other noted areas. Another reason is that scientist may not see the value of ontologies or a direct impact or value, making their engagement difficult. Ontology creation is a time-consuming, intellectually demanding undertaking; and scientists have limited time to devote to such efforts. To this end, ontology R&D needs to educate scientists/domain specialists as to the value of ontologies and provide mechanisms so that involvement in the ontology creation process is not too labor intensive.

Ontology work for Chronic Obstruction Pulmonary Disease (COPD) provides one example addressing these goals, driven by the practice of *ontological empowerment* [21]. User-friendly open source thesaurus software (TemaTres) was used to engage domain experts (medical researchers) in ontology design and maintenance work. In this case the domain experts had a sense empowerment by contributing to and maintaining the ontology. Moreover, the COPD ontology was seen as a valuable tool. The MSE predicament might be addressed in a similar way by facilitating domain expert engagement and leveraging the information scientist's expertise to provide a user-friendly development environment. Specific research objectives guiding the research presented in this paper are outlined in the next section.

4 Research Objectives

The objectives of this research were to:

- Explore an approach for engaging materials scientists in ontology development, including means by which information scientists may aid the process.
- Gather a threshold capacity measure, consisting of engagement time and number of terms, for domain scientists' engagement in the development of semantic ontologies.
- Develop a base-level ontology for nanocrystalline metals.
- Consider implications of this research for other areas in materials science and engineering (MSE) and other disciplines.

5 Methods

The posited research objectives were addressed using a multi-method approach involving algorithm comparison, semantic concept/term evaluation, and term sorting.

- The algorithm comparison combined term extraction results of four natural language processing open source applications. (*The phrase ‘algorithm comparison’ is used hereafter to reference this method.*) RAKE and Tagger support unsupervised algorithms, and Kea and Maui support supervised methods. Supervised methods involve training the models with documents that have been indexed by a person, indicating a gold standard.
- The semantic concept/term evaluation method followed general relevance evaluation processes, with a three tier scale of ‘valuable’, ‘not sure’, or ‘not valuable’. (*Concepts are intellectual ideas represented by single and bound terms as well as phrases. This paper generally uses “term/s,” although the discussion of algorithms uses the phrase ‘keyphrases’, consistent with broader Kea and Maui reporting in the scientific literature.*)
- The term-sorting activity was a basic clustering process, asking participants to separate and group concepts in preparation for establishing hierarchies and associations.

More details on method execution are presented in Section 6, Sample and Procedures.

6 Sample and Procedures

The research was conducted using a test corpus of 10 abstracts drawn from the arXiv repository. We generated our test corpus by searching the repository for the phrase “nanocrystalline”, selecting the 10 most recent preprints (as of May 2015), and collecting their abstracts for analysis. The following steps document the research design for the three methods.

6.1 Algorithm Implementation

To obtain our sample of terms, we needed to understand how to implement each of the algorithms and their operations.

- **RAKE** parses text into phrases (terms, bound terms, or term strings) based on given stop lists and desired keyphrase length and frequency. A candidate score is calculated for each phrase based on co-occurrence [22]. Finally, RAKE returns a list of keyphrases ranked by their scores. In this research, we generated word groups with the following constraints: each word had at least 4 characters, each phrase had at most 3 words, and each keyword appeared in the text at least once. We then selected phrases with scores higher than 5.0 as our keyphrases.
- **Tagger** is also an open source tool used for unsupervised automatic indexing [23]. Like RAKE, Tagger cleans the input text, splits it into words, rates the word according to relevance, and returns the top five candidates as keyphrases.

- **Kea** creates a model for keyphrase extraction using training files with manual indexing tags [24], and differs from RAKE and Tagger. The algorithm first splits text into sequences and calculates feature scores for all candidate phrases. A secondary step involves generating a model from keyphrases that are manually indexed and identified in the files. When extracting keyphrases from a new document, Kea parses text for candidate phrases, calculates feature values, and applies the training model to generate the keyphrases. In this research, we applied the default model in Kea package to use free indexing on our documents.
- **Maui** is similar to Kea. This algorithm first trains selected documents and keyphrases to build a model, and then uses the model to test on new data [25]. Maui selects candidate phrases by parsing sentences into textual sequences as tokens and extracting tokens based on given stop lists. For each candidate term, Maui calculates a score based on several features and put the scores into a machine learning model to learn the probability of real candidates. Compared to the Kea system, Maui only includes three basic Kea features and adds six new features. In our research, we used the Maui model created with the SemEval-2010 keyphrase set [26] for free indexing.

6.2 Term Evaluation

The terms extracted from each of the algorithms were combined into a single alphabetical list, and duplicate terms were removed. The list was distributed to three participants: one professor and two doctoral students working in the area of nanocrystalline metals. These domain experts were asked evaluate ‘if the term was valuable as a vocabulary word for disciplinary study of nanocrystalline metals.’ The following three indicators were used in the evaluation: valuable (v), not sure (ns), and not valuable (nv) for disciplinary study. These results were combined into a single set. Cases where all three ratings for a term were “v” or “nv” were easily determined as “v” and “nv.” There were no cases where all three ratings were “ns”. Table 1 shows our methodology for combining mixed ratings.

Table 1. Rating Synthesis.

| Mixed ratings | Overall rating |
|---------------|----------------|
| v, v, nv | v |
| v, ns, nv | v |
| v, nv, nv | nv |

6.3 Term Sorting

The sorting activity involved further clustering of terms under higher-level concepts for the development of hierarchies. This activity was supported by ConceptCodify

[27], which allows users to create groups with group names (functioning as top nodes or facets), and put cards into each group (instances).

7 Results

The results of this study are helpful in understanding how selected technologies can help information scientists work with domain experts, and for obtaining a measure of domain scientists' capacities for ontology design engagement. The results of this study are presented below under the designated methodological sub-headings.

7.1 Algorithm Comparison

In this research, different algorithms extracted different numbers of key phrases. RAKE generated terms ranked by their scores for each document, and we chose terms with a score higher or equal to 5.0. Therefore, we had 7 key phrases for each document, and in total we had 70 key phrases. Tagger extracted the top 5 terms with highest relevance for each document, and we have 50 key phrases from all the files. Kea indexed each document by 10 key phrases with each phrases less or equal to 5 words. Similarly, Maui indexed each file by 10 key phrases with each phrases less or equal to 3 words. Thus, Kea and Maui each generated 100 key phrases from all the files. Table 1 summarizes the outputs from each algorithm and Table 2 gives an example of key phrase extraction for each application.

Table 2. Algorithm comparison.

| Application/algorithm | RAKE | Tagger | Kea | Maui |
|--------------------------------|--------------|--------------|------------|--------------|
| Algorithm | Unsupervised | Unsupervised | Supervised | Supervised |
| Training files | N/A | N/A | Default | SemEval-2010 |
| Maximum word length of phrases | 3 | 3 | 5 | 3 |
| Number of terms | 70 | 50 | 100 | 100 |

The outputs from this activity were combined into a single dataset for the term evaluation activity.

7.2 Term Extraction and Evaluation

The term evaluation process allowed domain experts to identify terms, representing concepts, for ontology inclusion. Table 3 presents the total number of phrases extracted by different algorithms and the total number of unique phrases. Keyphrases extracted by different algorithms were saved as four independent files; we then tallied the number of keyphrases in each file, and removed duplicated keyphrases.

Table 3. Example of keyphrase extraction from one document.

| Application/algorithm | Example of keyphrase extraction for one document |
|-----------------------|--|
| RAKE | local temperature rise, grain structure stabilized, average grain sizes, nanoscale grain structure, 8.8, significant plastic deformation, stable nanocrystalline alloy, driven grain growth, intense strain localization, grain growth |
| Tagger | nanocrystalline, shear bands, evolution, strain localization, formation |
| Kea | shear bands, Grain Structures, Nanocrystalline, shear, Strain Localization, Thermally-Stable, Nanoscale, Nanoscale Grain, Nanoscale Grain Structures, Grain growth |
| Maui | thermally stable, grain structure, strain localization, nanoscale grain, shear bands, nanoscale grain structure, Thermally, Stable, nanocrystalline, localization |

Table 4. Total number of terms extracted and total number unique terms.

| Algorithm | Total number of terms/phrases extracted | Total number of unique terms |
|-------------------------|---|------------------------------|
| RAKE | 70 | 69 |
| Tagger | 50 | 50 |
| Kea | 100 | 96 |
| Maui | 100 | 96 |
| Combination of all data | 320 | 311 |

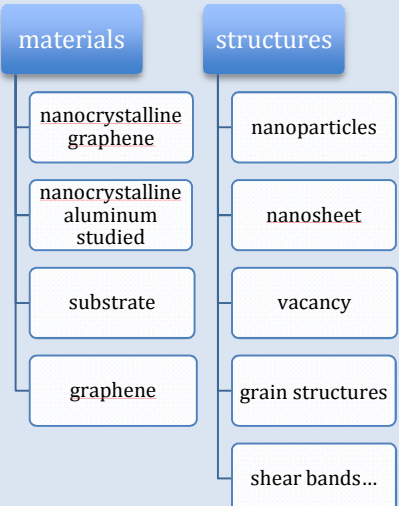
Table 4, column two presents the number of terms generated by each algorithm, and column 3 presents the number of unique terms per individual algorithm. The unique terms per individual algorithm execution were combined into a single set (311 terms); and close to 27% (83) of these terms were duplicative. That is, the term (which can include a keyphrase as abound term/s) was extracted via more than one of the algorithms. The 83 terms were removed, resulting a set of 228 unique terms for evaluation as candidate ontology terms.

The evaluation activity targeted the 228 terms and resulted in a corpus of 92 terms deemed valuable for ontology inclusion. The rating, noted above in the methods section, required at least rating of ‘v/ (valuable) with a second rating of ‘v’ or ‘ns; (valuable and not-sure). As reported above, there were no cases that had all three cases as ‘ns’ (not sure). The 92 terms deemed valuable were the corpus for the terms sorting activity, and serves as the nanocrystalline ontology source.

7.3 Term Sorting

The initial term sorting activity resulted in the identification of 9 top nodes listed on the left-hand of Table 5. Instances per node ranged from 2 to 11 for the initial ontology rendering, with an average of 6 instances per node. The right hand side presents two of the top nodes and associated instances.

Table 5. Nanocrystalline metals--ontology

| Top Nodes (Facets) | Nodes and Instances |
|--|--|
| <ol style="list-style-type: none"> 1. materials 2. structures 3. processes 4. material classes: 5. properties 6. techniques 7. descriptors 8. devices 9. physical objects |  <pre> graph TD materials[materials] --- n1[nanocrystalline graphene] materials --- n2[nanocrystalline aluminum studied] materials --- n3[substrate] materials --- n4[graphene] structures[structures] --- s1[nanoparticles] structures --- s2[nanosheet] structures --- s3[vacancy] structures --- s4[grain structures] structures --- s5[shear bands...] </pre> |

The initial sorting activity reported on here is being reviewed among the three participants. The scientist overseeing the initial activity responded to two survey prompts, and indicated that the term sorting/grouping task took roughly 20 minutes. The scientist also noted that the sorting activity was a “very straightforward task. The only factor that gave me [him] pause was the occurrence of phrases containing redundant terms (i.e. nano crystalline alumina alongside nano crystalline and alumina).” This example points to syntagmatic and paradigmatic relationships--a very common ontology design challenge. As part of next steps, the research team plans to address orphaned terms, and provide a mechanism for adding and tracking missing nodes and instances to complete the ontology, and move forward to creating SKOS encoded instances of this ontology.

8 Discussion: Threshold Determination

This research provides insight into what is an acceptable capacity for engaging scientists in ontology development. The algorithm preparation time was, as anticipated, fairly reasonable (approximately 4 work days of an information scientist's time). The

selected methods and unifying framework, merging outputs and eliminating duplicate terms, was an easy way to produce a corpus.

The term selection results were straight forward, taking each scientist roughly 10 minutes to evaluate the collection of 228 unique terms, resulting in a unified list 92 terms for ontology inclusion. An unexpected aspect of the term evaluation activity was that the two doctoral students were more direct using either ‘v’ (valuable) or ‘nv’ (not valuable), but neither used the ‘ns’ (not sure) criteria. It is difficult to gauge why they did not use this third indicator; it could be that the evaluation instructions were not clear, although all three indicators were evident on the scoring sheet. It is also possible that the doctoral students had great comfort with this activity or have been engaged in database work, and their evaluation patterns are reflective of Gruber’s classic notion of concepts (represented by terms) being either *in a world* or *outside*, with no ambiguity [28]. Follow-up is needed here to learn more about this result.

The second domain scientist task involved working with the ConceptCodify application and establishing group names (top nodes) and instances, drawing from the set of 92 terms. The scientist championing this work reported that it was relatively simple and took roughly 20 minutes. While some aspect of pause was noted, the scientist showed no frustration or sign cognitive overload, indicating that the method, number of terms, and time demand were suitable. These results point to an initial measure of *threshold acceptability*, and more data is needed to indicate where a time increase or more terms to evaluate or sort would indicate a threshold capacity.

This study is not without limitations. The nanocrystalline metals ontology, while robust with examples, is limited in scope. The sample was generated from a set of 10 of the most recent articles on nanocrystalline metals deposited in arXiv. A more extensive sample will very likely result in more terms requiring evaluation, and a larger corpus for the sorting activity. The time and intellectual demand from domain scientists will increase with a larger sample. To this end, the ontology research team is rethinking the sorting exercise and how to efficiently gather valid terms for completing the ontology. In closing, it’s likely that social networking technology, as demonstrated by YAMZ (formerly SeaIce) [29], with the thumbs up/down to garner team agreement, may offer an approach.

9 Conclusion

This study investigated a means for determining a threshold for engaging scientists in semantic ontology development. The research was conducted in the area of nanocrystalline metals, where there is limited evidence of a shared ontology. Materials scientists identified 105 terms for ontology inclusion; and an exploratory sorting activity resulted in 9 top nodes.

In reviewing the study’s objectives, the results present confirm an approach for engaging materials scientists in ontology development. The method pursued also demonstrates a way that information scientists may aid the process by generating a corpus of term. The resulting base-level ontology indicates a measure of threshold capacity for domain scientist engagement of approximately 10 minutes for evaluation, and 20 minutes for terms sorting and grouping.

Development and maintenance of semantic ontologies is crucial for advancing and accelerating MSE research. Semantic ontologies help provide insight into the full scope of a domain and enable discovery, sharing, and interoperability. In closing, ontologies, as intellectual maps, provide valued intelligence where they are applied. The M²I² will translate lessons learned here into our next stage of research, and will continue to advance ontology R&D in MSE.

Acknowledgements. We would like to acknowledge the National Consortium for Data Science (NCDS), 2014 Data Science Fellows Program; the National Science Foundation under Grant Number OCI 0940841; and thank you also to the participating scientists.

References

1. Aristotle: The History of Animals (350 B.C.E.) Translated by D'Arcy Wentworth Thompson. The Internet Classics Archive. http://classics.mit.edu/Aristotle/history_anim.html
2. Heuer, P., Hennig, B.: The classification of living beings. In: Munn, K., Smith, B. (eds.) Applied Ontology: An Introduction, vol. 9, pp. 197–217. Walter de Gruyter (2008)
3. Greenberg, J.: Philosophical foundations and motivation via scientific inquiry. In: Smiraglia, R., Lee, H.L. (eds.) Ontology for Knowledge Organization, pp. 5–12. Ergon-Verlag (2015)
4. Web Ontology Language. http://www.w3.org/standards/techs/owl#w3c_all
5. Simple Knowledge Organization System. http://www.w3.org/standards/techs/skos#w3c_all
6. Cheung, K., Drennan, J., Hunter, J.: Towards an ontology for data-driven discovery of new materials. In: AAAI Spring Symposium: Semantic Scientific Knowledge Integration, pp. 9–14, March 2008
7. Materials Metadata Infrastructure Initiative (M²I²). <https://cci.drexel.edu/MRC/projects/materials-metadata-infrastructure-initiative/>
8. Helping Interdisciplinary Vocabulary Engineering. https://cci.drexel.edu/hivewiki/index.php/Main_Page
9. Conway, M.C., Greenberg, J., Moore, R., Whitton, M., Zhang, L.: Advancing the DFC semantic technology platform via HIVE innovation. In: Garoufallou, E., Greenberg, J. (eds.) MTSR 2013. CCIS, vol. 390, pp. 14–21. Springer, Heidelberg (2013)
10. Zhang, Y., Greenberg, J., Tucker, G.T., Ogletree, A.: Advancing materials science semantic metadata via HIVE. In: International Conference on Dublin Core and Metadata Applications, São Paulo, Brazil, September 1–4, 2015
11. Callister, W.D., Jr.: Materials Science and Engineering—An Introduction, 5th edn. John Wiley and Sons (2000) (ISBN 0-471-32013-7)
12. Nano work - Nanotechnology Reports. <http://www.nanowerk.com/nanotechnology/reports/reports.php>
13. Tucker, G.J., Tiwari, S., Zimmerman, J.A., McDowell, D.L.: Investigating the Deformation of Nanocrystalline Copper with Microscale Kinematic Metrics and Molecular dynamics. *Journal of the Mechanics and Physics of Solids* **60**, 471–486 (2012)
14. Tucker, G.J., McDowell, D.L.: Non-equilibrium Grain Boundary Structure and Inelastic Deformation Using Atomistic Simulations. *International Journal of Plasticity* **27**, 841–857 (2011)

15. Affatato, S.: Ceramic-On-Metal for Total Hip Replacement: Mixing and Matching Can Lead to High Wear. *Artificial Organs* **34**(4), 319–323 (2010)
16. Materials Genome Initiative. <https://www.whitehouse.gov/mgi>
17. Materials Genome Initiative Strategic Plan: National Science and Technology Council Committee on Technology Subcommittee on the Materials Genome Initiative, December 2014. https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf
18. Biosharing. <https://www.biosharing.org/>
19. National Center for Biological Ontologies—bioportal. bioportal.bioontology.org/
20. Integrated Taxonomic Information System (ITIS). www.itis.gov
21. Greenberg, J., Deshmukh, R., Huang, L., Mostafa, J., La Vange, L., Carretta, E., O’Neal, W.: The COPD Ontology and Toward Empowering Clinical Scientists as Ontology Engineers. *Journal of Library Metadata* **10**(2–3), 173–187 (2010)
22. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents (2010)
23. Presta, A.: Tagger. GitHub. <https://github.com/apresta/tagger>
24. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: practical automatic keyphrase extraction. In: *Proceedings of the Fourth ACM Conference on Digital Libraries*, pp. 254–255. ACM (1999)
25. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: *Proc. of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 668–673. Morgan Kaufmann Publishers, San Francisco (1999)
26. Kim, S.N., Medelyan, O., Kan, M.Y., Baldwin, T.: Semeval-2010 task 5: automatic keyphrase extraction from scientific articles. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 21–26. Association for Computational Linguistics (2010)
27. ConceptCodify. <https://conceptcodify.com/>
28. Gruber, T.: *What is an Ontology* (1993)
29. Greenberg, J., Murillo, A., Kunze, J., Callaghan, S., Guralnick, R., Nassar, N., Ram, K., Janeé, G., Patton, C.: Metadictionary: advocating for a community-driven metadata vocabulary application. In: *DC-2013: CAMP-4-DATA Workshop: Proceedings of the International Conference on Dublin Core and Metadata Applications*, Lisbon, Portugal, September 2–6, 2013