Strategy for extensible, evolving terminology for the Materials Genome Initiative efforts

Talapady N. Bhat[1,4], Laura M. Bartolo[2], Ursula R. Kattner[3], Carelyn E. Campbell[3], John T. Elliott[1]

1 – Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
2 – Center for Materials Informatics, Kent State University, Kent, OH 44242, USA
3 – Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
4 – email: talapady.bhat@nist.gov

Abstract

Intuitive, flexible, and evolving terminology plays a significant role in capitalizing on recommended knowledge representation models for material engineering applications. In this article we present a proposed rules-based approach with initial examples from a growing corpus of materials terms in the NIST Materials Data Repository. Our method aims to establish a common, consistent, evolving set of rules for creating or extending terminology as needed to describe materials data. The rules are intended to be simple and generalizable for users to understand and extend and for groups to apply to their own repositories. The rules generate terms that facilitate machine processing and decision making.

Introduction

One of the grand challenges [1] of the Material Genome Initiative (MGI, http://www.whitehouse.gov/mgi) is to build a national infrastructure of federated data networks to support data analysis and product development through the reuse of materials data. To enable data sharing and reuse for materials innovation, consistency in terminology usage is essential. The current terminology used to describe materials data is heterogeneous, redundant, and often ambiguous. The lack of a common, community-based terminology and an infrastructure to support community building of the terminology hinder [1] the discovery and integration of material data for improved design of advanced materials.

This article presents a 'root' and rule-based approach as a common denominator for developing reusable, consistent terminologies across groups in the materials community. We used the growing National Institute of Standards and Technology (NIST) Materials Data Repository (https://materialsdata.nist.gov) as a source to develop and implement general rules for selecting roots and creating terms. We invite other groups working in particular materials areas or establishing materials data repositories to try this approach by adopting, adapting, and evolving the root and rule-based method (see section on **Results**) to address their terminology needs for building a common vocabulary. Additionally, evaluation and endorsement of the proposed approach can help ensure materials terminology is representative of the relevant concepts and used across the materials community to link materials data, databases, repositories, and facilities for the generation of new knowledge.

Overview of Materials Database and Knowledge Representation Activities

Materials Database Development

In the late 1980s and early 1990s, activities to develop computerized networked material databases were conducted without producing lasting results. However, companies and governments continued to construct their in-house databases, as well as proprietary, commercial databases for materials development [2]. Throughout the last two decades numerous articles have addressed database developments for materials data. Recent work has recognized the need for materials data and materials data management software during product development [3], as well as the need for 'sufficient' data content with recommended informatics infrastructure requirements for materials data [4]. The Materials Database Station (MDBS) of the National Institute for Materials Science (NIMS), the world's largest materials database available on the Web for academic and industry use, employs MatNavi to help users find relevant data in the databases [5]. As part of the NIMS databases, AtomWork [6] was made available on the Web providing phase diagram, crystal structure, X-ray powder diffraction, and property data of inorganic materials retrieved from scientific published papers. The material data environment (MDE) is a "system structure based on a division between the primary, numeric data and their metadata" and further separates metadata into material structure information and data source information [7]. A UK funded project focused on the design of a flexible database for metadata and file system for storing different types of materials data as images, raw data, and documents [8]. Most recently, the Uniform Description System for Materials on the Nanoscale [9] identified four broad major information categories, *General Identifiers; Characterization; Production; Specification*, that are used through the nanomaterials community to describe a nanomaterial as completely as possible. In discussion of the General Identifiers category the authors recognized the importance of practitioners creating formal and informal terminology to refer to aspects of the objects of interest, especially to be able to aggregate items of interest into classes.

These prior efforts made substantive impact on the development of materials databases and descriptions. Now the MGI has infused new life into the likelihood of publically available materials databases and repositories as part of a federated network.

A recent outgrowth of the 2011 report of the U.S. Office of Science & Technology, *Materials Genome Initiative* [10] and the 2013 Holdren Memorandum [11] has been the establishment of data repositories as complementary platforms to databases. Data repositories are infrastructures that hold digital content so that researchers and engineers can find, exchange, and incorporate each other's data through open access, commercial, or consortial arrangements (http://en.wikipedia.org/wiki/Content_repository). Within the materials community, there are several data repository efforts including the NIST Materials Data Repository (https://materialsdata.nist.gov), AFLOW (Automatic Flow for Materials Discovery) through the AFLOW Consortium (http://aflowlib.org), Citrine Informatics (http://www.citrine.io), Materials Project (https://www.materialsproject.org/), NCSA (National Center for Supercomputing Applications) Materials Data Facility as part of its National Data Facility (http://www.nationaldataservice.org/projects/mdf.html), and the University of Michigan Materials Common as part of its PRISM Center (Predictive Integrated Structural Materials Science: http://prisms.engin.umich.edu/#/prisms/materialscommons). In addition to storing datasets, many data repositories hold metadata that is searchable and describes associated data through a common set of descriptions, such as the Dublin Core Metadata Initiative (http://dublincore.org/). With both databases and repositories, the issue of building a common vocabulary across the large, diverse materials community remains a significant and unresolved challenge.

eXtensible Markup Languages and Schemas, Semantic Web Framework and Ontologies

MGI has also highlighted growing technological advances and cultural shifts currently underway that are changing the way scientific research is conducted, exchanged, and disseminated. An application protocol for describing a class of digital objects is eXtensible Markup Language (XML) that defines a set of rules for encoding documents in a format readable by both humans and machines (http://en.wikipedia.org/wiki/XML). XML provides the capability to describe strict hierarchies for applications to exploit by controlling content and combining data through XML Schemas. Some recent materials related XML schema initiatives that have been developed include MatML, the Materials Mark-up Language [12] and ThermoML, the Thermodynamics Mark-up Language [13, 14]. While XML Schemas are well suited to storing data, the ability of XML schemas to discover and integrate data can be somewhat limited. Combining different XML schemas can be difficult and the schemas and terminology may not adequately represent the domain knowledge that needs to be expressed.

A standard model for data exchange is Resource Description Framework (RDF). In his seminal 2001 article [15], Tim Berners-Lee introduced the phrase, "Semantic Web", to describe his vision of a “web of data”. He envisioned a common framework where networks of data could be connected and processed by machines for exchange and reuse across applications and communities. The World Wide Web Consortium (W3C http://www.w3.org/),founded in 1994, is the primary international standards organization and endorsed an RDF standard that represents information as a statement or “triple” comprised of a subject, predicate and object on a graph (A second version was released 25 February 2014, http://www.w3.org/RDF/). This simple yet powerful model represents relationships between concepts (subjects and objects through predicates) as labeled and directed graphs as illustrated in Fig 1. In this example the subject “rapid temperature changes” is connected to the object “deformation” through the predicate “may cause”. The predicate can create a strong (“measures”) or weak (“may cause”) connection. A collection of triples or “RDF graphs” can be used to combine, expose, and share structured (e.g., relational database) or semi-structured (e.g., email) data across different applications and domains.

A related approach for effective data discovery and exchange has been the development of ontologies or machine-readable concept maps to describe concepts and relationships between the concepts in a particular domain, such as materials science and engineering. Web Ontology Language (OWL) , the W3C standard for defining structured, Web-based ontologies, further extends RDF’s ability to provide a rich description to enable data interoperability by facilitating the sharing of data and knowledge across different domains. (http://www.w3.org/2004/OWL/). A recent example of a materials related ontology is MatOnt, an OWL ontology about materials, their structure, properties, and processing [16].

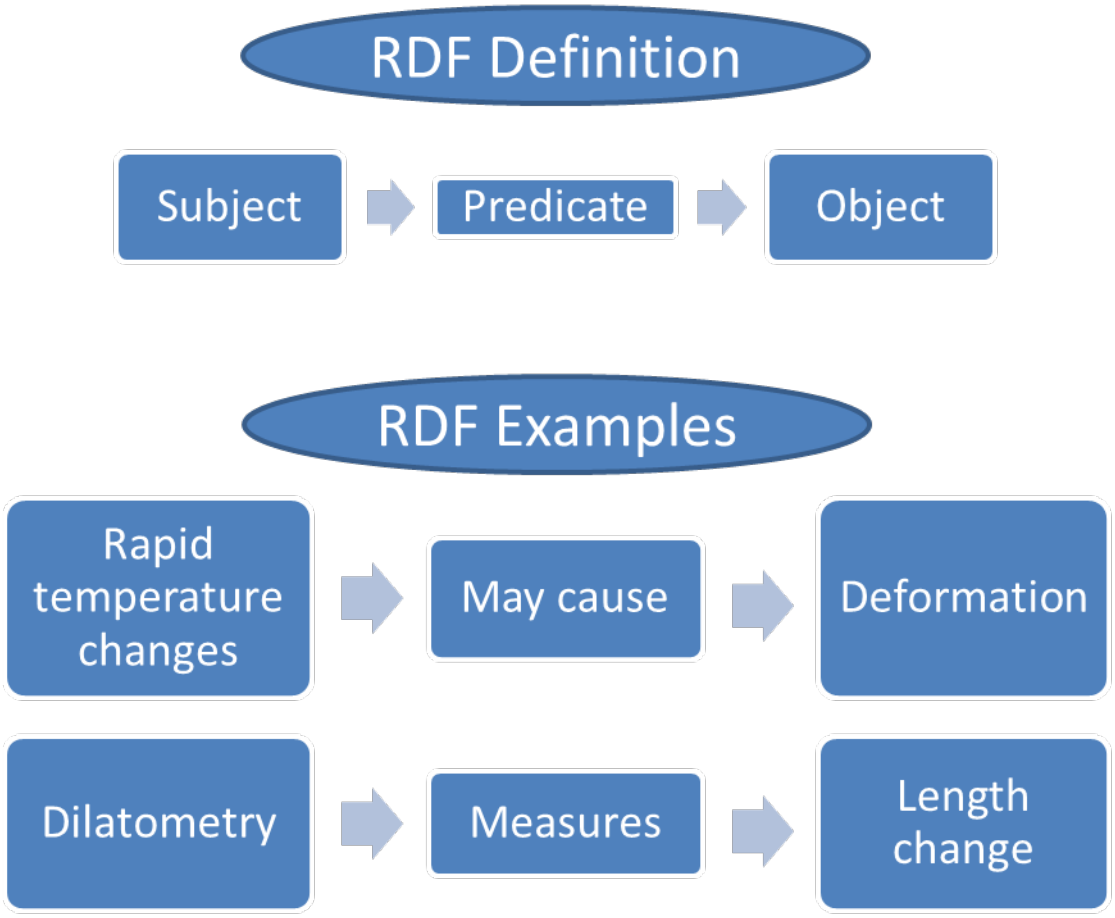

Fig. 1 An RDF graph of a triple made up of two nodes (subject and object) and a link (predicate) connecting them.

By making use of metadata, markup-languages, and Semantic Web standards, like RDF, new technologies can expand their capabilities and benefits. For example, electronic laboratory notebooks (ELNs) have the ability to create, store, and retrieve digital records of scientific and technical activities conducted in a laboratory. By incorporating Semantic Web standards, ELNs can also extend their functionality to include the capacity to process the data for remote analysis [17]. In addition to providing optimized workflow and time-saving benefits, ELNs, with semantically annotated data, could interact with distributed resources to help perform scientific procedures [18]. Technological innovations, like RDF and ELNs, are changing how science is connected and practiced. These changes increase the need for a logical, easy method for the materials community to grow a common materials data terminology. Other domain communities, such as bioinformatics, have demonstrated that the process of creating agreed upon terminology can be adopted, evolve, and accelerate rapidly through crowdsourcing over a distributed infrastructure [19, 20].

Methods: Rules-based Approach in the NIST Materials Data Repository

Root & Rules-based Method

The primary motivation of the proposed approach to developing terminology is to improve the discoverability of materials data beyond what is typically retrieved from general search engines, such as Google. This improved discoverability will be achieved through the establishment of rules-based search engines that link data, repositories, databases, facilities, and other distributed resources through more intelligent connections. To achieve this end, rules-based search engines makes use of specialized terminology, characters, and a hierarchical structure as core elements in its linguistic framework and facilitate the creation of the Semantic Web [21].

Overview of Proposed Rules-based Method

Terminology in root-based languages is self-evolving and need-based. In these languages, the community creates words on-demand, if none exists, based on community needs and preferences. In spite of the constant evolution of words in use, root-based languages are able to ensure well-defined semantics for words by establishing a flexible, easy-to-use linguistic framework to streamline the process of word creation, sharing, mapping, and evolution. Roots and rules are critical components of these languages and our proposal is to create such a foundation for MGI terminology building.

The proposed root and rules-based approach supports easy terminology building and creates a terminology that is easy-to parse and extend or mutate. These roots can be overlaid to identify related groups of terms and can be selectively replaced to create new, related terms for use in new use-cases. The root and rule-based approach provides additional enhancements to traditional term building approaches, such as: a) limiting grammar dependent semantics and local jargon in roots or terms to support interoperability within and across disciplines as well as languages.; b) tracking how suggested terms are used to better understand their semantics and to build effective use-cases; and c) restricting the creation of synonymous terms.

To date, we have experimented on building a root-based linguistic framework for scientific data documenting and sharing [22]. Following examples in the natural world for chemical and biological phenomena, we initially applied this approach to classify and search protein ligands from the Protein Data Bank [23] and the PubChem (http://xpdb.nist.gov/chemblast/pdb.pl). Subsequently, we extended the root-based approach to examine terms used in biology (http://xpdb.nist.gov/bioroot/bioroot.pl) for text-based data. Based on this examination, we established a tentative foundation to create a root and rule-based framework to manage terms in a cell-image database [24] using certain English words as roots. Now we present a root and rule-based approach and definitions of some different types of roots as applied to the NIST Materials Data Repository for use with databases and repositories developed within the Material Genome Initiative.

Specialized Terminology: Root, term, super root, and tethered root

Many Indo-European languages, such as English, utilize a limited set of highly reused, non-synonymous, short words called **roots** that can be combined to facilitate the building of new **terms**. This approach is more prominent in certain languages, such as Sanskrit, Latin, and German, and these languages, more than English, permit the creation of terms on-demand as well as the replacement of a root in an existing term by one or more other roots to create a new, related term.

The proposed MGI terminology employs these definitions and makes use of these **root** and **term** concepts.

Tethered roots are roots obtained by constraining two or more roots together without a defined special character between them. Midtown, biochemistry, supertanker are examples of **tethered roots** in the English language and their qualifiers (mid, bio or super) are not used on their own as words. Applying this logic to MGI terminology, firstprinciples, gammaprime, and longrange are proposed as **tethered roots**. Tethering restricts automated methods from parsing the roots into individual terms.

Super roots are concatenated roots to identify special cases. Peanut butter and police dog are examples of **super roots** in English language. These examples possess collectively a special meaning that is different from the dictionary meanings of the individual words that comprise these **super roots**. Peanut butter is not the same as *peanut* and *butter* and police dog is not the same as *police* and *dog*. Additionally, **roots** of a **super root** have qualifier-qualified relationships for a given context. Applying this logic to MGI terminology, elemental symbols describing an alloy, (e.g., Al_Mg) are proposed as **super roots** along with an embedded special character, the **underscore (_)**. This character is used to inform automated methods that these terms have the lowest preference for parsing in roots.

Terms can be formed by concatenating **roots**, **super roots,** and **tethered roots** with an **embedded hyphen (-)**, such as Melt-temperature, to create new terms of interest to a discipline. **Compounded terms** are formed by concatenating terms using **an embedded colon (:)**, e.g., Melt-temperature:Cu. **Compounded terms** are expected to have a high degree of specificity and thus they may be used to define entries (objects) reasonably accurately within a database. **Colons**, **hyphens**, and **underscores** are called delimiting characters and were chosen to avoid conflicts with special characters used by popular mark-up languages such as XML, and HTML (e.g., / < >). For further information, see http://www.w3.org/International/questions/qa-escapes). They also serve as a hierarchy that informs automated methods on how to parse the terms to maintain semantic relevance.

Machine Processing, Granularity and Hierarchical Structure

Figure 2 follows up the discussion in the preceding sections, "Overview" and "Special Characters", to illustrate the application of the rules-based method to represent "Differential thermal analysis to measure the melting temperature of Copper". Using this method, *Cu melting temperature measured using differential thermal analysis,* is constructed through the use of **compounded term**, **super root**, **root**, and **tethered root** combined with hierarchical processing to decipher semantics and clarity. The special delimiting characters (**colon**, **hyphen** and **underscore**) embedded in a term support machine friendly ways to introduce granularity among roots of a term, allowing tools to step through the hierarchy from colon, hyphen and underscore iteratively in order to decode their semantics. The hierarchy is maintained uniformly (left to right) across the vocabulary (i.e., compounded term, super root, and tethered root) so as to simplify their interpretation by automated methods. The meaning and specificity of the **compounded term** used in Figure 2 must be sufficient to point to an entry (object) in a database or repository and preferably use a hierarchy that is relevant to the use case as detailed in the **Rules** section (Rules 6, 7, 16). **A super root** (e.g., *Differential_thermalanalysis)* is comprised of individual **roots** (e.g., *Differential*) and/or concatenated (e.g., *Thermalanalysis*) **tethered roots** with a **classifier-classified** relationship where the roots can be separated. However unexpected loss in the meaning and clarity of the **super root** can occur when its roots are separated. A

tethered root (e.g., *Thermalanalaysis*) has a qualifier-qualified hierarchy where the hierarchical combination of **roots** (e.g., *Thermal* and *Analysis*) adds greater clarity and cannot be separated.

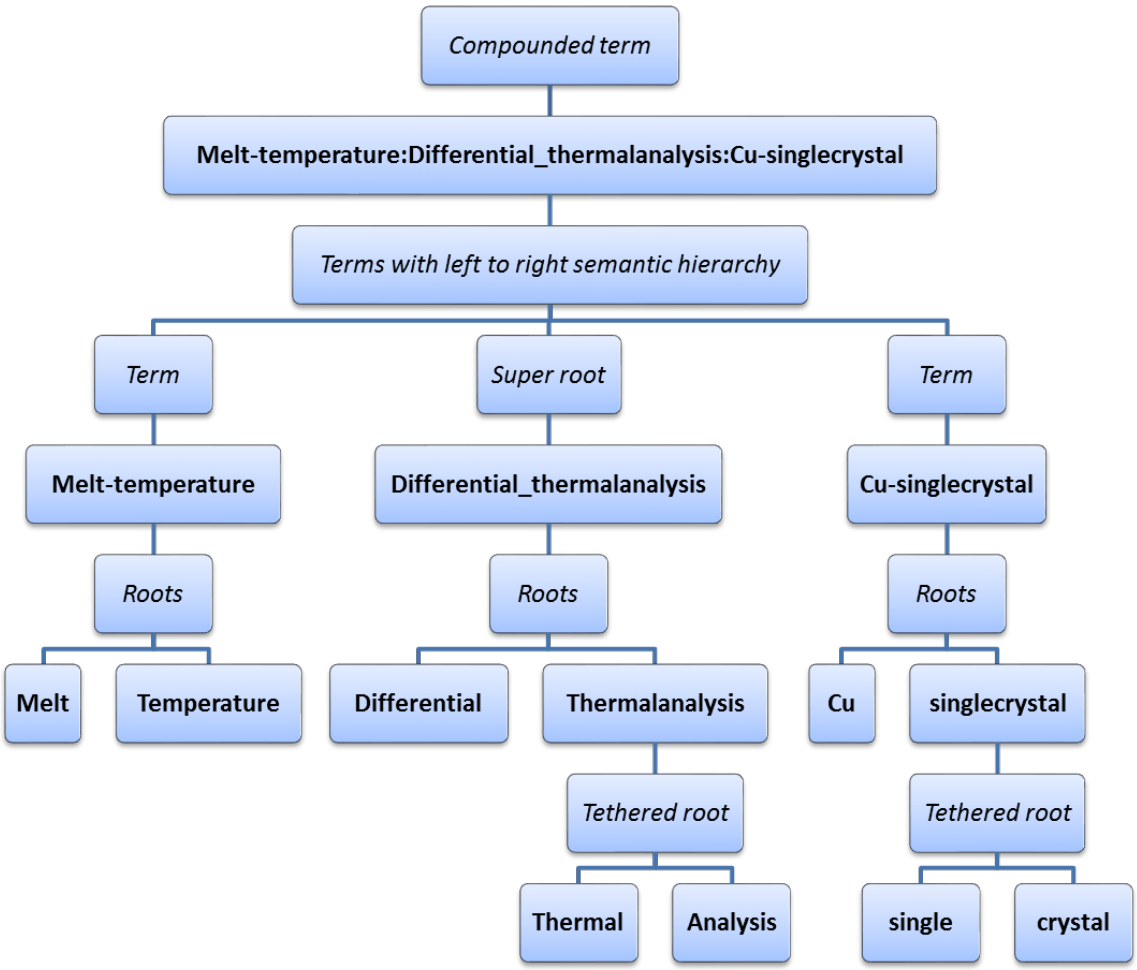

Figure 2: Semantic hierarchy and terminology for ***melting temperature of a Cu single crystal measured using differential thermal analysis***. Terms, super roots and tethered roots are shown in bold and their root or term types are shown in italics.

The NIST Materials Data Repository

The NIST Materials Data Repository [25] was established over the past two years to store contributed files with a focus on phase-based properties which include but are not limited to thermodynamics, diffusion mobilities, molar volume, and elastic modulus. The repository has implemented a uniform, flexible standards based framework using DSpace [26] for submission, storage, and dissemination. The repository assigns unique persistent identifiers for submitted data files, associates the files with valuable background information, and provides a data citation to encourage data reuse and attribution. The advantages of using the repositories include the validation of published results and as well as time savings for users. Traditional reporting of

scientific results and data occurs in journals, conference proceedings and reports often without additional information in electronic form. The user of these data would then retype the data. This process is time consuming and prone to the introduction of errors. In recent years journals started to provide these data a supplemental material with the journal article. Even then, the reporting of results rarely includes the raw data from which the results were derived making it impossible to verify or improve the results from the original analysis. For example, the CALPHAD (Calculation of Phase Diagrams) [27] method uses a number of different files, containing original experimental data, model descriptions and, possibly, auxiliary files. Using the files with the experimental data, parameters of the model descriptions are obtained by an optimization process. Files with model descriptions of binary and ternary systems are then combined to give a model description of a multicomponent system. The availability of these files will be essential for efficient updates of CALPHAD descriptions and databases.

Our application of the root and rules-based approach in the NIST Materials Data Repository is comprised of the following steps: a) Creation of a series of selection rules to identify a representative set of reusable 'roots' from existing terminology for materials data from the NIST repository; b) Creation of an initial set of rules to concatenate roots to create terms; c) Creation of adequately discriminating terms from the roots based upon the rules. Future steps will address the establishment of a publicly available web resource to house the latest set of roots and terms for review, adoption, and adaption by the materials community with the ability of community members to find accepted roots and terms and to suggest additional roots and terms to be incorporated into the common materials vocabulary via the web resource. The primary focus of this paper is to present our thoughts with respect to steps (a) & (b), to seek community input, and to provide a few examples for step (c).

Results: Proposed Rules, Brief Summaries, Examples, and Exceptions

To facilitate a community discussion, we have applied the proposed roots and rule-based approach to part of the NIST Materials Data Repository and have developed an initial set of rules and terms with examples and exceptions.

The rules and examples that follow are recommended as initial guidelines and illustrations to begin development of a common evolving rationale to be used in the materials community for selecting terms to describe materials data. We plan to apply the rules to the terms in the NIST Materials Data Repository to generate terms for the Repository for enhanced searchability. Overall the rules are divided into two categories: 1) Rules 1-10 are for users to find accepted terms and to recommend new terms and 2) Rules 11-17 are primarily for database developers and repository administrators to incorporate the terminology into their respective databases or data repositories. Though the root and rule-based approach discourages the creation of non-compliant terms, it does not prevent creation of non-compliant terms. For example, a non-compliant database of specific terms could be created as 'dialects' or local extensions to fit a particular material community need.

General Rules to Create Roots, Super roots and Tethered roots

Roots are highly reused short nouns or modified nouns (nouns modified by other words such as nouns, verbs, adjectives etc.). Roots formed by modified nouns can be of two types; i.e., super roots or tethered roots that have two or more roots written as one word.

Rules for users to create roots, super roots and tethered roots:

1. **Forming roots:**
   a. Use all roots in singular form except where plural form is used more frequently.
      Example: ***Material*** instead of **materials**
      Example: ***Property*** instead of ***properties***
      Example: ***Modulus*** instead of ***moduli***
      Exception: ***Species*** or other roots that are intrinsically plural
   b. Avoid including special characters (such as ’ : _ - = / \) as a part of a root.
      Example*: **Xray*** instead of ***X-ray***
      Example: ***Newtons*** instead of ***Newton’s*** as in ***Newtons’s law***
   c. Avoid the use of modifiers of roots
      Example: ***Gas*** instead of ***gaseous***
   d. Use abbreviations only when they are widely accepted across many related disciplines and when they are unambiguous in their meaning. See **Rule 5** for exceptions when acronyms are embedded in a super root. Use uppercase for all acronyms except for atomic symbols.
      Example: ***Au*** instead of ***gold, Cu*** instead of ***copper,***
      Example: ***SANS*** instead of ***small angle neutron scattering***
   e. For similar expressions choose a shorter equivalent as a root.
      Example: ***Composite*** instead of ***composite material***
      Example: ***Micrograph*** instead of ***microstructure image***
      Exception: ***Microstructure evolution***
2. **Forming super roots:**
   A super root is formed when the roots involved do not have a preferred discriminating power and semantics to serve as node names of a data-graph or as RDF elements except in special circumstances.
   a. Super roots are concatenated by an underscore to indicate its compound semantics and its ability to be parsed into individual roots only under unusual conditions. If a super root is comprised of roots that are not specific when considered individually, then refer to tethered root (See **Rule 3**).
      Example: ***Li_ion*** or ***Ni_diffusion*** instead of ***Li-ion*** or ***Ni-diffusion***
   b. When a root of a super root functions like a hierarchical classifier to another root then also include the classified root in the super root so that automated parsers can recognize the hierarchy. To order roots within a super root, unless there already exist a well-accepted alternate convention, use rule 6 (hierarchical classifier-classified rule).
      Example: ***Alloy_Au_Cu*** instead of ***Au_Cu-alloy*** or other ways of writing
3. **Forming tethered roots**:
   a. Create tethered roots when a root is a qualifier of another root and the semantics of any root on its own may not be of interest in a database or data repository search. Tethered roots are formed to indicate that the roots involved need be considered collectively, rather than individually, in order to derive their semantics. For this reason, roots in a tethered root are written contiguously to avoid inadvertent separation by automated methods. Since tethered roots are comprised of qualifier and qualified roots, following a general convention of root-based construction of English language words, we use their intrinsic qualifier-qualified relationship to order their roots (e.g., biochemistry.
      Example: ***Firstprinciples***, instead of ***first principles***
      Example: ***Heatcapacity*** instead of ***heat capacity***,
      Example: ***Gammaprime*** instead of ***gamma prime***
   b. A root may appear in more than one tethered roots

Example: ***Shortrange*** and ***Longrange***

Example: ***Latticeconstant*** and ***Latticeexpansion***

Tethered roots may also provide a way to avoid the use of stop words in a compounded root. That is, move the word from the right of a stop word to the left, drop the stop word, and place the qualifier before the qualified.

Example: ***Vaporizationheat*** instead of ***heat of vaporization***

Rules for creating terms:

4. **Forming terms from roots:** Terms are formed by concatenating two or more roots, super roots or tethered roots using a hyphen (-) so that automated methods may re-generate their roots when necessary. We suggest to order roots of a term by classifier-classified relationships (See **Rule 6**) which is also a general convention in English, as in police dog or technical paper unless there is a different well accepted convention.

   Example: ***Al3Ti-longrange_order***

   Example: ***Liquid-shortrange_order***

   Example: **Firstprinciples-formation_enthalpy-Al3Ti**

5. **Avoiding ambiguities and redundancies:**
   a. Avoid using ambiguous acronyms. Instead clarify their meaning by qualifying them with a classifier 'root' to form a super root or a tethered root or use the complete phrase. See **Rule 1d** for examples of using acronyms.

      Example: ***Densityfunctionaltheory*** or ***Discretefouriertransform*** instead of ***DFT***

      Example: ***Atomicforcemicroscope*** or ***Antiferromagnetic*** instead of ***AFM***

      Example: ***Finiteelementmethod*** or ***Fieldemissionmicroscope*** or ***Electromotiveforce*** instead of ***FEM***

      Example: ***FCC_A1*** instead of ***FCC*** (***face centered cubic*** or ***Federal Communications Commission***)

      Exceptions: ***SEM*** (scanning electron microscopy), ***TEM*** (transmission electron microscopy), ***VASP*** (Vienna Ab-initio Simulation Package)

   b. Avoid the inclusion of redundant words in a term

      Example: ***Xray_diffraction*** instead of ***Xray_diffraction-method***

      Example: ***Optical_microscope*** instead of ***Optical_microscope-imaging***

      Example: ***SEM*** instead of ***SEM_Microscopy***

6. **Ordering roots in a term – classifier-classified rule:** Roots (super root, tethered root and root) within a term are organized by a left to right, semantic top-down, classifier-classified hierarchy. In general, classifier and classified roots are expected to have one-to-many relationships where, in a rules-based approach, for example, the root ***alloy*** is a classifier for many materials. **Rule 16** deals with instances where a relationship is not obvious or when a relationship changes over time due to the addition of new terms. In short, the classifier-classified hierarchy is not absolute but but depends on the number of classified roots available for a given classifier root in a given use case.
   a. One way to identify classifier and classified roots in a term is to arrange the terms with an embedded hierarchical top-down, level-based classifier (for each 'classifier' term there exists several possibilities of 'classified' terms) statement with a hyphen between classifier and classified terms (e.g., ***Modelingsoftware-VASP***, ***Modelingsoftware-Abinit***). On sorting these terms, classified roots appear as the fast varying strings (***VASP***, ***Abinit***) and their classifier roots appear as the slow varying term (***Modelingsoftware***). Automated methods may use this feature to develop hierarchical data models that can be presented as data-graphs or RDF or used for auto-complete to select terms for reliable search results.

      Example: ***Phase_properties-Modelingmethod-CALPHAD***

      Example: ***Firstprinciples-Modelingsoftware-VASP***

Example*: **Alloy_Mg-yieldstrength***
Example: ***Alloy_Fe-wrought***
Example: ***Alloy_Al_Cu-precipitation_hardened***

b. When a classifier-classified relationship does not exist among the roots, e.g., for a collection of atomic elements, place them in an alphabetical order.
Example: ***Ag_Au_Cu*** instead of ***Cu_Au_Ag***
Example: ***Alloy_Ni_base_Al_Co_Cr_Ti*** to describe a superalloy.

7. **Creating roots and terms with similar, multiple, or complex meanings:** Following **Rule 1e**, use a shorter root for words with similar meaning whenever possible. A root embedded in a term can help automated methods, such as co-word analysis, natural language processing, and text-mining, to identify related semantic classes. To facilitate this process it is recommended: a) to limit the use of synonymous roots; b) if necessary, clarify the semantics of a root by appending it with a classifier-root.
Example: ***Experiment-type*** instead of ***Experimental-techniques*** or ***Experimental-method,***
Example: ***Longrange_order*** instead of ***Long-range-stacking-order***
Example: ***Thermophysical-data-source*** instead of **Reference for thermophysical properties,**
Example: ***Image-graph***, ***Image-micrograph*** instead of ***Graph***, ***Micrograph*** or ***Image***
Example***: Materialstate*** instead of ***State***
8. **Reusing roots to create terms:** Create terms by combining roots so that terms have clear semantics. Avoid terms that are broad and general in meaning. Create terms that can serve as 'semantic expressions' in use-cases. A rule of thumb is to attempt to form terms with three roots and, if needed, combine between two and five terms to form suitable semantic expressions.
Example: ***Cu-lattice_constant*** instead of two (***Cu*** and ***lattice_constant)*** separate term(s) with an additional root ***(Cu-crystal-lattice_constant*** (***lattice constants*** are always for ***crystals***)).
Example: ***Li_ion-batteryproperty*** instead of ***Li_ion*** and ***battery-property***
9. **Creating compounded terms that identify a group of objects in the material database:** Compound terms serve as 'use-cases' defining semantic expressions of terms and they are formed by concatenating two or more terms using a **colon (:)** as a special delimiting character. Compounded terms that are overly specific are unlikely to be reused. It is advised to limit the number of terms in a compounded term to between two and five terms. Compounded terms may point to persistent identifiers (PIDs), such as DOIs (Digital Object Identifier) for query purposes. Compounded terms may be used by database providers or repository administrators to cluster, identify, and display related items using messages like 'related to items that you have viewed'.
a. Use classifier-classified hierarchical **Rule 6** to decide the order of terms in a compounded term.
Example: ***Crystal-structure:Laue-method:Cu-singlecrystal-FCC_A1***
Example: ***Melt-temperature:Differential_thermalanalysis:Cu-singlecrystal***
b. When creating compounded terms, give importance to 'use case-on-demand' hierarchies, which are case-based rather than fixed schema-based hierarchies. Order a term so that a term to the left has one-to-many relationship with the term to its right.
Example: ***Modelingsoftware-VASP:Crystal-structure:Cu-FCC_A1***

Example: ***Stainless-wrought-precipitationhardened:Vickers_hardness***

10. **Providing the reference of any paper that supports the use of the new term(s) you are creating**. The reference may serve as a 'definition' of the term as well may demonstrate use of the term within a context. To reference the term ***Alloy_Al_B_Ni:Interstitial_diffusion_coefficient_B*** use
    Example: Campbell CE, Kattner UR (1999) A thermodynamic assessment of the Ni-Al-B system. J Phase Equilib 20 (5): 485-496
    Example: DOI: doi:10.1361/105497199770340743

Rules for Database Developers and Repository Administrators: Publicly available resources play an important role in facilitating the use and evolution of rule and root-based approach to build terminology. For this reason we also propose a few preliminary rules primarily meant for database or repository developers.

11. **Design for readability of compounded terms:** Use uppercase for the first letter of a term and use lowercase for all the rest unless a root is a short form or a symbol e.g., ***VASP***, ***beta***
12. **Provide usage statistics for terms:** For each term in a database or repository, store its usage statistics for users to inspect, along with the terms. These frequencies may allow a user to avoid terms that are used infrequently.
13. **Provide semantic context of terms and compounded terms:** In the database, also keep and display a bibliographic reference and/or DOI to illustrate the use and semantics of the term. This reference may also be used as the basis to build use-case-specific compounded terms or segments of data-graphs.
14. **Identify new terms introduced by users as well as flag terms if no documentation is provided** (See **Rule 10**)
15. **Allow the creation of dialects:** Terms that do not follow the rules may also be created as local dialects when necessary. Dialects may facilitate a gradual evolution of rule-based terminology and the rules in a crowd-sourced environment.
16. **Curate and validate terminology and compounded terms on a regular basis:** Dialects are important components of the proposed method for terminology building. Therefore accepting or removing dialects as terminology must be facilitated by public resource providers who act as caretakers. Redefining super roots, tethered roots and classifier-classified relationships among roots are all important steps of the evolution process of the proposed term building effort. Database developers and repository administrators need to have an established mechanism for regular updates to support a smooth evolution process. Frequency of usage and the semantic context of terms are useful factors to monitor in such an evolution process.
17. **Apply new technologies that have been adopted widely:** Explore whether new data technologies may require the rules to be updated, e.g. the relatively new dynamic data citation model [29].

Discussion

The rules proposed above facilitate creating root-based terms on-demand as found in root-based languages. **Rule 9** designates a packing character to form compounded terms so that they can be readily created or parsed by automated means to obtain their roots. For example, a large portion of material data could be organized in compounded terms using a <property>:<method>:<material> hierarchy. Roots obtained by parsing terms can be used to identify related terms (for instance *Modelingsoftware-Pandat* and *Modelingsoftware-DICTRA* are related since they share the

common tethered root, *Modelingsoftware*). After parsing terms, one or more roots can be selectively replaced to create new, related terms for use in a different related 'use-case' (e.g. *Modelingsoftware-LAMMPS*). The proposed root and rule-based approach also includes the following features: a) methods to limit the use of grammar dependent semantics (**Rule 1**) and jargon in roots or terms (**Rule 1.d**) to improve their interoperability within and across disciplines as well as across languages; b) traceability to common use of suggested terms to help understand their semantics when building new use cases (**Rules 13, 14**); and c) features to limit concurrent creation of synonymous terms (**Rules 1, 2, 3, 7, 8**).

Rules 2 and **4** designate two semantic dependent special delimiting characters to pack roots: one to form super roots and the other to form terms. These characters can be used by automated tools to unpack terms into their roots based on their semantics. One could designate additional packing characters, if needed, to identify other semantic relationships among a set of roots of a term. Packing and unpacking of terms with designated packing characters could be used to collate elements with certain semantics represented by RDF (See section on *eXtensible Markup Languages and Schemas, Semantic Web Framework and Ontologies* for discussion of RDF) into terms or vice versa using **Rule 6**. Also with **Rule 6**, packing and un-packing capabilities using designated special delimiting characters of a term could be used to represent small segments of data-graphs as terms without 'stop' words that complicate syntax and are difficult to standardize, e.g. formation enthalpy vs. enthalpy of formation. Therefore, 'stop' words are often ignored by natural language-processing-methods which may lead to incomplete or erroneous results. Certain rules (See **Rules 3, 6, and 9**) for placing roots within a term can allow one to introduce hierarchical structures among roots of a term for use with Semantic-Web technologies.

For chemical structures, chemists have been using rule-based methods to name chemical compounds for decades and the International Chemical Identifier (InChI) was developed using this concept (International Union of Pure and Applied Chemistry, http://www.iupac.org/inchi). The extensive use of the InChI since its introduction [29] demonstrates that efficient and unique annotation can be achieved by combining manual annotation with rule-based methods. Here we propose a rule and root-based approach to create and curate text-based terminology for the material science and engineering community. The focus of the proposed method is not only to improve precision and recall when searching material data for research and product development but also to facilitate reuse, automation, and scalability of metadata representation to meet the needs of a federated MGI effort.

Adoption and adaption of the rules and concepts presented here by other materials groups could provide valuable feedback for improvements and identification of other possible alternatives. The trial and iteration of creating well-structured, machine-readable vocabulary can help coalesce the materials community around a consensus-based strategy for building and establishing a common materials terminology to support materials innovation. A potential application could be to use the proposed RDF-based approach in conjunction with natural language analysis to organize the huge amount of material data that are in published form (journals, reports, etc.) in digitally processed form to facilitate machine processing and decision making.

Acknowledgments: We thank Prof. E. Subrahmanian and J. Collard for useful discussion on some of the linguistic aspects of the compounded terms which will be published elsewhere.

Disclaimer: NIST does not endorse any commercial products, and use of these products does not imply endorsement by NIST.

References

1. C.H. Ward, J.A. Warren, R.J. Hanisch, *Intergr. Mater. Manuf. Innov.*, 3:22, (2014), doi:10.1186/s40192-014-0022-8.
2. *Computerization and Networking of Materials Databases*, Vols. 1 to 5, ASTM International ASTM special technical publication: 1017, 1106, 1140, 1257, 1311; 1989 to 1997.
3. D. Cebon and M.F. Ashby, *MRS Bull.*, 31, 1004-1012 (2006).
4. S.M. Arnold: *MRS Bull.*, 31, 1013-1021, (2006).
5. M. Yamazaki, Y. Xu, M. Murata, H. Tanaka, K. Kamihira, K. Kimura, in *Baltica VII: Life management and maintenance for power plants*, Vol. 2, eds. P. Auerkari, and J. Hal, (Espo, Finland: Valtion Teknillinen Tutkimuskeskus, ISBN 9789513863173, 2007), pp 193-207.
6. Y. Xu, M. Yamazaki, and P. Villars, *Jpn. J. Appl. Phys.*, 50:11RH02, 1-5, (2011), doi:10.1143/JJAP.50.11RH02.
7. D.E. Boyce, D.R. Dawson, and M.P. Miller, *Metall. Mater. Trans. A*, 40A, 2301-2318 (2009), doi:10.1007/s11661-009-9889-y.
8. M. Scott, R.P. Boardman, P.A. Reed, T. Austin, S.J. Johnston, K. Takeda, and S.J. Cox, *Inform. Syst.*, 42, 36-58 (2014).
9. J. Rumble, S. Freiman, and C. Teague, Uniform Description System for Materials on the Nanoscale, CODATA/VAMAS Joint Working Group on the Description of Nanomaterials, Beijing, 2014, *http://www.codata.org/uploads/Uniform_Description_System_Nanomaterials-Published-v01-15-02-01.pdf*. Accessed 16 April 2015.
10. National Science and Technology Council (U.S.), "Materials Genome Initiative for global competitiveness" (Executive Office of the President, Washington, DC, 2011). *https://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf*. Accessed 17 April 2015.
11. J.P. Holdren, "Increasing access to the results of federally funded scientific research" (Memorandum for the heads of executive departments and agencies, Office of Science and Technology Policy, Executive Office of the President, Washington, DC, 2013) *https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf*. Accessed 17 April 2015.
12. G. Kaufman and E.F. Begley, *Adv. Mater. Process.*, 161(11), 35-36 (2003).
13. M. Frenkel, R.D. Chirico, V. Diky, Q. Dong, K.N. Marsh, J.H. Dymond, W.A. Wakeham, S.E. Stein, E. Königsberger, and A.R.H. Goodwin, *Pure Appl. Chem.*, 78, 541-617 (2006).
14. M. Frenkel, R.D. Chirico, V. Diky, P.L. Brown, J.H. Dymond, R.N. Goldberg, A.R.H. Goodwin, H. Heerklotz, E. Königsberger, J.E. Ladbury, K.N. Marsh, D.P. Remeta, S.E. Stein, W.A. Wakeham, and P.A. Williams,.*Pure Appl. Chem.*, 83, 1937-1969 (2011).
15. T. Berners-Lee, J. Hendler and O. Lassila, *Scientific American*, 284(5), 35-43 (2001).

16. K. Cheung, J. Drennan, and J. Hunter, in *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, eds. D.L. McGuinness, P. Fox, and B. Brodaric (Palo Alto, CA: AAAI, 2008) pp. 9-14.

17. M. Rubacha, A.K. Rattan, and S.C. Hosselet, *JALA*, 16(1), 90-98 (2011)

18. C.L. Bird, C. Willoughby, and J.C. Frey, *Chem. Soc. Rev.*, 42, 8157-8175 (2013)

19. S.D. Larson and M.E. Martone, *Front. Neuroinform.*, 7:18 (2013), doi: 10.3389/fninf.2013.00018.

20. M.G. Kahn, L.C. Bailey, C.B. Forrest, M.A. Padula, and S. Hirschfeld, *Pediatrics*, 133, p 516-525 (2014), doi: 10.1542/peds.2013-1504.

21. K.M. Hettne, A.J. Williams, E.M. van Mulligen, J. Kleinjans, V. Tkachenko, and J.A. Kors, *J. Cheminf.*, 2:3 (2010), doi:10.1186/1758-2946-2-3; correction *J. Cheminf.*, 2:4 (2010) , doi:10.1186/1758-2946-2-4.

22. T.N. Bhat, *JSWIS*, 6(3), 22-37 (2010).

23. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, and H. Weissig, *Nucleic Acids Res.*, 28, 235-242 (2000).

24. A.L. Plant, J.T. Elliott, T.N. Bhat, *BMC Bioinf.*, 12:487 (2011).

25. C.E. Campbell, U.R. Kattner, and Z.-K. Liu, *Scr. Mater.*, 70, 7-11 (2014).

26. M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, and J.H. Walker, *D-Lib. Mag.*, 9(1) (2003), doi:10.1045/january2003-smith, *http://hdl.handle.net/1721.1/29465*. Accessed 17 April 2015.

27. U.R. Kattner and C.E. Campbell, *Mater. Sci. Technol.*, 25 (2009) 443-459.

28. A. Rauber, S. Pröll et al., "Scalable Dynamic Data Citation Approaches, Reference Architectures and Applications, RDA WG Data Citation Position Paper," Draft Version - 2015-03-23, *https://www.rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html*. Accessed 19 May 2015.

29. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev, *J. Cheminf.*, 5:7 (2013), doi:10.1186/1758-2946-5-7.