Towards a Content-Based Material Science Discovery Network

Emily LeBlanc, Marcello Balduccini, and William C. Regli
Applied Informatics Group
Drexel University
Philadelphia, PA
{ecl38, mb3368, regli}@drexel.edu

Abstract

Many publicly available databases exist for managing materials scientific data. These databases contain information from a wide variety of work, and the information is typically encoded in some proprietary format aimed at highlighting the goals of their particular backgrounds and purposes. In order to accelerate the rate at which new materials are discovered, these databases must be federated to provide materials scientists with the means to efficiently access large quantities of highly relevant data. This position paper advocates the design of a content-based material science discovery network that can allow for more intelligent reasoning over the databases than current implementations can afford. We will discuss the gains of using a hierarchical ontology for describing metadata that captures the various layers of the materials science domain. We will then discuss our approach in a content-based networking context.

Introduction

The ability to engineer new materials to exhibit effective and valuable performance characteristics is central to the discovery and development of advanced technologies. High performance materials have impacted the progress of civilization, from the creation of stone tools that aided ancient humans in their survival in a hostile landscape, to the technological revolution sparked by the microprocessor. It naturally follows that continuing to facilitate discovery and innovation in the materials science field will have a significantly positive impact on the present and future advancement of civilizations. Such support additionally provides more immediate benefit to areas of research such as healthcare, security and energy. To these ends, initiatives are in motion (NSTC 2011) with the aim to accelerate the rate at which new materials and alloys are developed and deployed, with a significant emphasis on reducing the time and financial costs of engineering them.

One of the major requirements of the effort to foster materials innovation is to ensure that researchers and engineers have access to all publicly available materials scientific information that can inform the design and analysis of a candidate material's performance characteristics and mechanical properties. This includes, and is not necessarily limited to, information about the classification of materials and alloys, the quantitative properties describing the materials and their microstructures, as well as available manufacturing processes. Organizations from universities and corporations have independently compiled large quantities of searchable data, however the interfaces differ significantly. For example, Harvard University's Clean Energy Project (Olivares-Amaya et al. 2011) database of organic chemicals supports search by parameters describing the performance of solar cells, or by name and substructure of molecules. The University of Tokyo's MatNavi (NIMS)[1] database provides an interface in which the user navigates a directory structure in order to access a number of other databases with a variety of querying methods. The returned results from these databases also tend to vary greatly in format and clarity. A problem clearly exists in the lack of a standard information exchange model among these databases - consulting and deciphering information from each database will be a time consuming task. A basic software solution that links the resources together in a single interface may be able to expedite the research process, but is insufficient to aid in the discovery process. At the very least, the solution to this problem requires the development of a sophisticated ontological framework overlaying the existing databases to allow researchers to submit a single query to get relevant results. Further, the sheer volume of available data begs a solution that automates some of search functionality to expedite the development process.

We advocate the development of a content-based material science discovery network that makes use of a hierarchical ontology for associating relevant metadata to database content. In the section that follows, we will discuss in greater detail the problem of searching for information across independently managed resources and related efforts to federate the scientific metadata. Next, we will include a detailed description of a multi-level ontology with hierarchical structure within each level. The following section will introduce the concept of content-based networking and how this type of approach can facilitate faster discovery and development of new materials. Finally, we provide an example scenario to highlight the suitability of our approach for the problem

[1] http://mits.nims.go.jp/index_en.html

at hand and some closing discussion.

Related Work

McLeod and Heimbigner (1985) formally defined a federated database system as one which "define[s] the architecture and interconnect[s] databases that minimize central authority yet support partial sharing and coordination among database systems." Federating materials science databases requires a language that can be used not only to communicate with all existing collections, but may also be extended to integrate future resources that are essential to a fuller understanding of the domain. Ontological frameworks have been proposed to make use of semantics to reveal linked information among heterogenous databases. Ashino et al. (2006) propose an ontology to aid in the materials selection process. Their work, however, does not attempt to capture the entire domain of material science. The PLINIUS ontology (van der Vet, Speel, and Mars 1994), focusing on ceramic materials, automated some search for keyword matches within the content of their database. This method of search can be resource consuming, as every document in every database must be searched to find a match. As new databases are integrated into a system such as this, the time required to fulfill a request increases and scalability may become an issue. The most advanced of these efforts is the MatOnto project (Cheung, Drennan, and Hunter 2008), which proposes a framework aiming to enable materials scientists to "search, retrieve and integrate data from heterogeneous and disparate data sources". The ontology has a rich structure for describing materials, including classification of properties, processes, and there are even terms to describe the properties of a type of material's crystalline structure. However, the materials themselves are apparently flatly organized, e.g. any two materials in the system share a sibling relationship. For example, within the set of cobalt alloys, talonite is a variety of stellite. In a flat ontology, this relationship is not represented (see Figure 2 below for a visualization of this relationship). This approach does not allow for more intelligent reasoning over a hierarchical organization. This reduces the task of locating relevant materials content to a keyword-based search over their metadata (e.g., as provided by Google and most other internet-based search engines). This pitfall of this approach will be addressed further in our discussion of our proposed ontological structure, and our proposed solution will be illustrated in the scenario.

Ontological Structure

To foster scientific discovery, we advocate for a more sophisticated way of organizing materials scientific metadata into a domain specific, hierarchical ontology identifying instances of classes, subclasses, properties, such that relationships can be established among these elements. The terms of the ontology are the metadata language that describes content within the repositories, and a hierarchical structure is navigated to find viable candidate materials based on the specifications of the query.

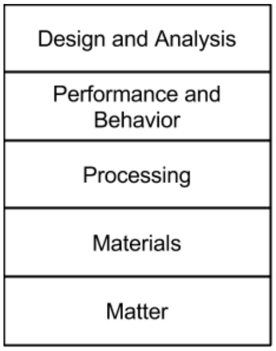

Figure 1: Layered Material Ontology

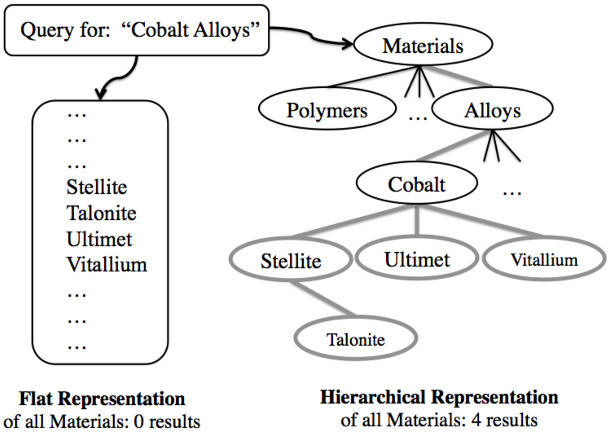

Figure 2: Flat Representation of Materials vs. Hierarchical Organization

It aids in understanding of a material science ontology to envision the domain in layers (Figure 1). The bottom layer describes matter, providing the basis for any present or future materials. Here we find elements, chemicals, compounds, and their physical properties, including their microstructures. Next, we find descriptions of the materials themselves. The material layer uses the language of matter that we have already defined to describe the very large domain of material instances. The third layer provides a description of available processes for manufacturing and enhancement of the materials. Now that we have a detailed description of feasible materials, we can encode information about their performance characteristics and behavior. This layer encodes metadata about the physical properties associated with materials such as density, electrical resistance, or elasticity. Elements from any layer may be related to elements from any layer including its own.

Let us now return to the problem of keyword-based search. As previously discussed, efforts to impose an ontological framework on the space of materials science databases organize their materials in a linear fashion, resulting in a keyword-based style search in the process of materials selection. In this context this style of search can be cumbersome, inefficient, and error-prone, as it cannot always account for difference in vocabulary or meaning across the data sources, and disregards any hierarchical organization of knowledge. More generally speaking, a gap may exist between the search terms used by the user who is performing the search, and the terms employed by the person or tool describing a certain object.

For example, consider a scientist querying for a list of all metal alloys exhibiting some desired property. The scientist will perform a search using the term “metal alloy”, while the entries in the available catalogs are tagged with their most specific alloy type, e.g. “electrum” or “vitallium”. Traditional keyword matching will fail to detect that electrum is a type of metal alloy. Although it is technically possible for the catalog entries to be tagged with multiple labels, doing so at the time of insertion into the knowledge base is unfeasible in the general case, because: (1) it results in extremely large repositories, since the extra tags must cover all of the queries that may possibly be asked, and (2) it would force a complete and difficult re-computation of the tags if the hierarchical structure of knowledge is changed at any point - for example, by introducing the notion of “precious metal alloy”. Even more sophisticated methods of keyword-matching, such as query expansion, cannot help here - ”electrum” is neither a synonym nor a morphological form of ”metal alloy”. All in all, the typical keyword matching solution described here is a fairly unsophisticated approach in the context of a hierarchical organization of terms - the scientists will likely receive a number of unrelated results and miss some relevant ones altogether.

A hierarchical organization of materials enables for more intelligent reasoning over the metadata. This concept is related to the semantic web, a branch of Artificial Intelligence. As is the case in semantic web applications, we propose to encode common-sense relationships, such as classification of materials and the propagation of properties through subclasses, such that a computer can simulate the way a human would reason over the information. By capturing subsumption relationships among materials, a reasoner is able to return a greater number of useful results while ignoring those which may be similar in name but not by classification. To extend the metal alloy example, let’s assume that we would like to ask a database for a collection of alloys of cobalt. Figure 2 describes this subclass of material first as a flat ontology, representing keyword-based matching, and then shown again as a hierarchy depicting parent-child relationships in this subdomain of cobalt alloys. Notice that the in flat representation, a broad query for “cobalt alloy” returns no results because the description does not match any of the metadata in the list. Using the hierarchical representation of metal alloys, however, a query will return all cobalt alloys even if their descriptions do not offer an exact match. An ontology of this design can significantly improve the quality of the work by navigating expansive data repositories and returning all relevant results for a given query.

Content-Based Discovery

In the approach that we propose, we employ a content-based networking approach for intelligently reasoning over multiple materials databases. The idea of content based networking was introduced by Jacobson (2009) to address the fact that network use had “evolved to be dominated by content distribution and retrieval.” His claim was that network technology was focused purely on connections between hosts, an approach that seemed less fit for the task of meeting Internet users’ demands for content. Van Jacobson proposed shifting the existing network paradigm to prioritize the *what* over the *where* by retrieving content by name rather than by location.

In a content-based network, the files (or content) of the system are addressed instead of their host machines. A receiver declares its interest in a particular type of file, for example, to the network in the form of predicates, and senders simply offer their content to the network without knowledge of who will be receiving it. The network is responsible for routing content that matches the predicates of the receivers. There is a clear correlation between this model of networking and the problem of accessing the content of numerous databases using the ontology that we have described above. Not only is the ontology providing means for sophisticated reasoning, it now provides the names by which the network may identify interests. A scientist can make the previously mentioned request for a list of metal alloys to the network, and all hosts who offer relevant content will simply push their data into the network for delivery to the interest’s originator. It is easy to see that a network of this kind is not limited to one time queries - it is also able to support persistent queries, or subscriptions. If the same scientist subscribes to any new content related to metal alloys, the system will automatically fulfill the request for new content as it arrives in the network. The consequences of this are of great benefit to the scientist, as they will receive up to date materials data as soon as it becomes available in the network. Further, he will receive only the most relevant information due to the design of the ontology.

The term “content” is not restricted to the definitions in these layers and hierarchies. In addition to querying the material catalogs, literature resources can be easily incorporated into the network such that when a candidate material is found, a collection of related publications is also returned that may aid in a better understanding of the found materials. The proposed ontology can also be extended to capture experiments for reuse by other investigators. For example, once the scientists have designed an acceptable set of potential materials, they may wish to query the network for simulated experiments that they can use to confirm or disprove that the compounds perform as they expect. Additionally, there may be a description of shapes used to aid in the design of parts. This is a new layer on top of what has already been defined, which describes design and analysis of newly discovered materials and objects made from them. Here we may find a number of known physical representations and associated experiments by which newly discovered materials can be tested in various ways.

Scenario

Let’s consider a scenario in which a team of scientists has been tasked with developing a material for a new coronary stent. The stent must be flexible enough to minimize potential scarring of arterial tissue, but must also be strong enough to support the artery and reduce future narrowing resulting

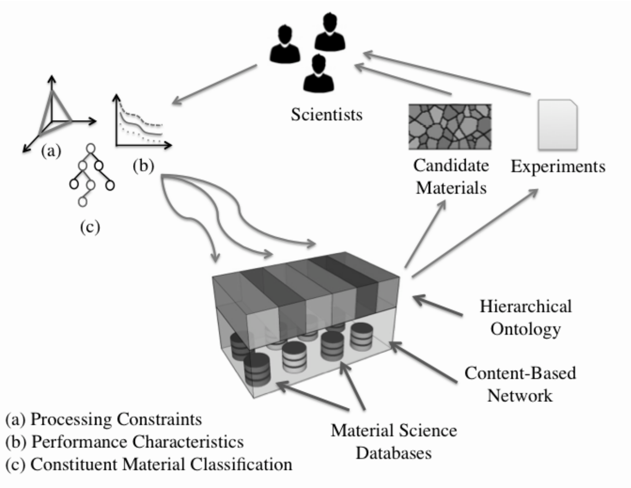
Scientists
(a)
(b)
(c)
Candidate Materials
Experiments
Hierarchical Ontology
Content-Based Network
Material Science Databases
(a) Processing Constraints
(b) Performance Characteristics
(c) Constituent Material Classification

Figure 3: Example of the data flow in a materials scientific discovery network.

from any scarring that may occur over time. To the knowledge of the scientists, no material exists that satisfies all of these properties. The team must investigate the fabrication of a new material to meet the requirements of the project. Their approach is to design a reinforced compound from existing materials that possess complementary subsets of the desired properties. They must also take limited availability of processing options into consideration when they are researching candidate materials.

The scientists have determined that they can design a stent from a metal matrix nanocomposite that will be processed by spraying the compound over a tubular mold. They are seeking out a moderately strong metal alloy for use as the reinforcement structure in the nanocomposite. In addition to providing structural support for the composite, the reinforcement alloy will ideally minimize the new material's friction coefficient. The matrix itself must be monolithic, meaning that its microstructure has a continuous crystal lattice without deformities. The resulting composite must also be sufficiently flexible as not to obstruct the natural contractions of the artery. Finally, any candidate materials must be able to undergo the process of molding by spraying. From these constraints, the scientists may formulate a query which includes specifics about material classification (including crystalline structure), desired performance characteristics, and available manufacturing options. It is easy to see that a unified ontological framework is a very suitable approach to accessing the metadata of disparate material sciences resources. Due to the complexity of this query, the scientists further benefit from our proposed hierarchical structure as they will not receive data irrelevant to what they have asked the network for, as we described in our above discussion of flat versus hierarchical representation of the domain.

Consider a situation where the scientists receive all of the relevant situation that the network presently has to offer. After some design and analysis of the returned material sets, it has been determined that the resulting materials are good, but not sufficient to perform as they had hoped. This same query can be repurposed as a subscription, perhaps with some modifications due to discoveries made about the first set of materials. If a new collection of lightweight, monolithic alloys is submitted to some database, that database will respond to the persistent query by pushing its new collection on to the network, which is then routed back to the scientists. Furthermore, this subscription can be extended to ask for a set of experiments to perform that will validate the new materials and their part in the design of the stent, thus expediting the design and analysis process for the scientists. It is clear that in our scenario, discovery is facilitated by the organization of the ontology in concert with the content-based network. Figure 3 provides a visualization of the data flow in a materials science discovery network as we have proposed here.

Conclusion

The reality of such a system has three major implementation-level requirements. Firstly, the ontologies and reasoning must be used at all levels of the representation, ranging from classified materials to the experiments used to analyze them. Secondly, the databases need to be federated in a context-based network structure with support for both ad hoc and persistent queries. Thirdly, the reasoning must be capable of evaluating the content in the databases in order to detect inconsistencies and gaps in knowledge. We believe that the careful design of the proposed network will be of great benefit to scientist investigating the borders of technology, and can contribute to the further advancement of the materials science field.

References

Ashino, T., and Fujita, M. 2006. Definition of a web ontology for design-oriented material selection. *Data Science Journal* 52–63.

Cheung, K.; Drennan, J.; and Hunter, J. 2008. Towards an ontology for data-driven discovery of new materials. In *Semantic Scientific Knowledge Integration*.

Heimbigner, D., and McLeod, D. 1985. A federated architecture for information management. *ACM Transactions on Information Systems* 3(3):253–278.

Jacobson, V.; Smetters, D. K.; Thornton, J. D.; Plass, M. F.; Briggs, N. H.; and Braynard, R. L. 2009. Networking named content. In *CoNEXT '09*.

NSTC. 2011. Materials genome initiative for global competetiveness. Technical report, National Science and Technology Council.

Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; and Aspuru-Guzik, A. 2011. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy and Environmental Science* 4849–4861.

van der Vet, P. E.; Speel, P. H.; and Mars, N. 1994. The plinius ontology of cermanic materials. In *In the Eleventh European Conference on Articiaial Intelligence Workshop on Comparison of Implemented Ontologies*.