



# Compound Knowledge Graph-Enabled AI Assistant for Accelerated Materials Discovery

Kareem S. Aggour<sup>1</sup> · Andrew Detor<sup>1</sup> · Alfredo Gabaldon<sup>1</sup> · Varish Mulwad<sup>1</sup> · Abha Moitra<sup>1</sup> · Paul Cuddihy<sup>1</sup> · Vijay S. Kumar<sup>1</sup>

Received: 30 August 2022 / Accepted: 26 October 2022 / Published online: 8 December 2022  
© The Author(s) 2022

## Abstract

Materials scientists are facing increasingly challenging multi-objective performance requirements to meet the needs of modern systems such as lighter-weight and more fuel-efficient aircraft engines, and higher heat and oxidation-resistant steam turbines. While so-called second wave statistical machine learning techniques are beginning to accelerate the materials development cycle, most materials science applications are data-deprived when compared to the vastness and complexity of the search space of possible solutions. In line with DARPA's vision of third wave AI approaches, we believe a combination of data-driven statistical machine learning and domain knowledge will be required to achieve a true revolution in materials discovery. To that end, we envision and have begun reducing to practice a system that fuses three forms of knowledge—factual scientific knowledge, physics-based and/or data-driven analytical models, and domain expert knowledge—into a single 'Compound Knowledge Graph' in which contextual reasoning and adaptation can be performed to answer increasingly complex questions. We believe this Compound Knowledge Graph-based system can be the nucleus of a collaborative AI assistant that supports stateful natural language back-and-forth dialogs between materials scientists and the AI to accelerate the development and discovery of new materials. This paper details our vision, summarizes our progress to date on a steam turbine blade coating use case, and outlines our thoughts on the key challenges in making this vision a reality.

**Keywords** Knowledge representation · Third wave AI · Information fusion · Domain knowledge · Cathodic arc deposition

## Introduction

Materials development has historically advanced as much by serendipity as through rigorous application of the scientific method. Three examples from within GE include the development of silicon carbide, silly putty, and Lexan. In 1891, Edward Acheson was attempting to create artificial diamonds when he discovered silicon carbide, a light-weight yet extremely hard abrasive compound used today in hard ceramics such as car brakes and bulletproof plating, that is also a semiconductor and thus used in many power electronics [1]. During WWII, in 1943, James Wright was attempting to invent a synthetic rubber to help with the war effort, and inadvertently invented what became silly putty [2]. In 1953, Dr. Daniel Fox was attempting to invent a malleable protective wire coating and, in the process, invented Lexan, an extremely hard, highly durable clear plastic used in aircraft canopies, astronaut face shields, CDs, and DVDs, among many other applications [3].

---

✉ Kareem S. Aggour  
aggour@ge.com

Andrew Detor  
deter@ge.com

Alfredo Gabaldon  
alfredo.gabaldon@ge.com

Varish Mulwad  
varish.mulwad@ge.com

Abha Moitra  
moitraa@ge.com

Paul Cuddihy  
cuddihy@ge.com

Vijay S. Kumar  
v.kumar1@ge.com

<sup>1</sup> GE Research, One Research Circle, Niskayuna, NY 12309, USA

Even today, many discoveries continue to be the result of trial-and-error experimentation and the occasional surprise. As a result, developing a new material can be an arduous task that can take more than a decade for applications with complex, multi-objective performance criteria requiring a balance of properties. Over the last several decades, materials have been considered as the limiting factor for achieving significant improvements in jet propulsion technologies, to name just one application domain. During jet propulsion's history, the average rate of increase for turbine engine material temperature capability has been roughly 50°F per decade. For example, the development timeline of single crystal nickel (SX Ni) superalloys, the most used material in jet engine turbine blades, indicates that the endurance temperature increased from 1800°F for the first generation SX Ni superalloys in the 1970s to about 2050°F for the sixth generation superalloys in the 2010s [4]. Much of the SX alloy development has been incremental, e.g., adding one or two new elements (Re, Ru) or changing the content by a few percentage points to achieve a slightly better balance of properties.

Designing and developing new materials continues to be critically important for GE because we wish to continue pushing the boundaries of what is possible in the areas of the future of flight, the clean energy transition, and precision health. This includes producing the world's first commercial ceramic matrix composites (CMCs)—materials that are lightweight, durable, and highly heat resistant for more fuel-efficient aircraft engines, developing highly thermal resistant materials to improve the efficiency of gas and steam turbines, and new superconducting magnets to improve the speed and image quality of MRI medical scanners. Although with CMCs, GE increased jet engine temperatures by 150°F in one decade through systematic efforts during their peak development [5], CMC research and development dates to the 1970s when the US government first funded CMC research. Within the first 25 years, GE had run CMC turbine shrouds in multiple industrial gas turbine applications for electricity production until it shifted CMC focus to jet engines to meet high-temperature heat resistant material demand. Today, CMC parts are flying in GE's commercial jet engines all over the world. GE's CMC turbine shrouds have surpassed 10 million flight hours in the hottest section of the CFM LEAP turbofan engines as of 2021 [6]. Interestingly, the central element of CMCs is silicon carbide, the same material discovered by chance at GE in 1891.

## Materials Discovery Challenges

Materials discovery is challenging because the relationships between composition, processing, microstructure, and resulting material properties are not well understood in concert. This is due to the complexity, high dimensionality, and

nonlinearity of the fundamental underlying mechanisms that drive material behavior. While some equations governing these relationships have been discovered for certain properties, in aggregate the equations known today are only a small fraction of what materials scientists would need to develop new materials through a purely analytical approach. Similarly, advances in computational materials science (e.g., density functional theory, molecular dynamics) have made significant contributions to the field, but are still limited to simple structures, small numbers of atoms, and short time scales. A computational approach that simulates all relevant material physics remains a distant future prospect.

To address these challenges, materials scientists are now partnering with artificial intelligence (AI) and machine learning (ML) experts to tap into AI/ML techniques driving successes in other fields. Recent advances in materials development can be traced to the launch of the Materials Genome Initiative (MGI) in 2011 [7]. Since then, ML approaches have been successfully applied to a wide range of challenges, and academic leaders in the field [8–11] as well as a growing list of start-up companies [12–14] are changing the way materials science is performed. A growing infrastructure of datastores [8, 9, 15, 16], materials-specific utilities [17, 18], and easy-to-use open source software tools,<sup>1,2,3</sup> have democratized ML, making advanced data-driven approaches accessible to the materials scientist.

While the continuing progress in ML-driven materials science is exciting and productive, there are important limitations to consider. Existing approaches excel at interpolation—optimizing within a known and tested parameter space—but are fundamentally challenged when extrapolating toward new understanding. DARPA recognizes the limitations of these so-called “second wave AI” approaches due to their lack of reasoning and contextual adaptation capabilities, and their dependence on massive quantities of data for training [19]. Most materials science applications are still data-deprived when compared to the vastness and complexity of the search space. Thus, new materials discovery strategies are required that do not rely exclusively on data alone. With rare exceptions [20], current applications of ML in materials science largely neglect existing analytical knowledge and domain expertise in guiding model predictions. This renders the system of discovery isolated from the governing laws of the physical world. It is becoming apparent that the incorporation of domain knowledge will be necessary to enable a step-change in the pace of materials development and true materials discovery.

<sup>1</sup> <https://scikit-learn.org/>.

<sup>2</sup> <https://pytorch.org/>.

<sup>3</sup> <https://www.tensorflow.org/>.

As one possible approach to address this challenge, we envision a system that fuses multiple forms of knowledge into what we call a *Compound Knowledge Graph* and enables contextual reasoning and adaptation over that combined knowledge to answer increasingly complex questions. Our concept is inspired by DARPA's vision of "third wave AI" [19], in which second wave ML systems are combined with domain knowledge to allow for learning and abstraction of concepts to reason about areas never seen before. We believe the approach detailed in the following sections will accelerate the discovery of new advanced materials to help solve our most pressing societal problems.

### Vision: A Third Wave AI Assistant for Accelerated Materials Discovery

Inspired by DARPA's third wave AI concept, our vision is to develop a collaborative AI assistant that can serve as a true partner in the scientific discovery process. This assistant will be able to answer increasingly complex materials science questions, ranging from simple queries requiring lookup of data to questions where complex contextual reasoning and inference outside the bounds of the provided data is required to produce an answer. Rather than being limited to interpolating between data available within the system, our goal is to demonstrate substantial progress toward accelerating the AI-driven creative process by being able to help accomplish increasingly challenging tasks such as

1. Inferring previously unknown properties of known materials,
2. Estimating properties of new, user-defined materials, and
3. Proposing novel materials (and commensurate processing methods) to meet increasingly challenging performance requirements.

To achieve this vision, our strategy is to capture and fuse three distinct, complementary forms of knowledge—factual, analytical, and human expert knowledge—together into a single Compound Knowledge Graph that supports contextual reasoning and adaptation about as yet unexplored areas such as novel chemistries and/or unique processing combinations.

### Knowledge Representations and Knowledge Graphs

While powerful, most data-driven ML techniques produce black box models, meaning the representation of the patterns derived from the training data are opaque to humans and therefore the 'knowledge' embedded in the patterns cannot be rationalized. Knowledge-driven approaches take

a different approach and are designed from the ground up to be transparent and explainable and can be interpreted by humans to understand what knowledge is captured and how and why certain conclusions are reached.

Many diverse knowledge representation approaches exist, including those that are logic-based (e.g., semantic models), and graph-based (e.g., Bayesian networks). While they have been in use for over 50 years, semantic models have soared in popularity over the last decade with the growth in popularity of 'knowledge graphs', a term re-popularized by Google [21]. Knowledge graphs can be construed as capturing knowledge in a graphical structure in which nodes represent distinct pieces of knowledge and the edges represent connections between nodes [22]. Semantic models or 'ontologies' are used to define the classes and relationships within a domain of interest and provide the underpinning of many popular knowledge graphs in use today (e.g., DBpedia,<sup>4</sup> Wikidata<sup>5</sup>). Ontologies define a set of concepts (classes) in a specific domain (which become nodes in a graph), the attributes (properties) of those concepts, and the links or relationships (edges) between them. The semantic model defines the classes and relationships of knowledge in a modeled domain, and the data in the model represent specific instances of knowledge, but the model itself is also a very important form of knowledge as it represents how experts within a specific domain think about their field and details the terminology they use to describe the concepts in that field. A number of teams have explored using ontologies to represent and capture materials data<sup>6</sup> though the breadth of coverage and maturity of the models varies widely [23–25]. Additional information about materials ontologies can be found in a survey paper by Zhang et al. [26].

### The Compound Knowledge Graph

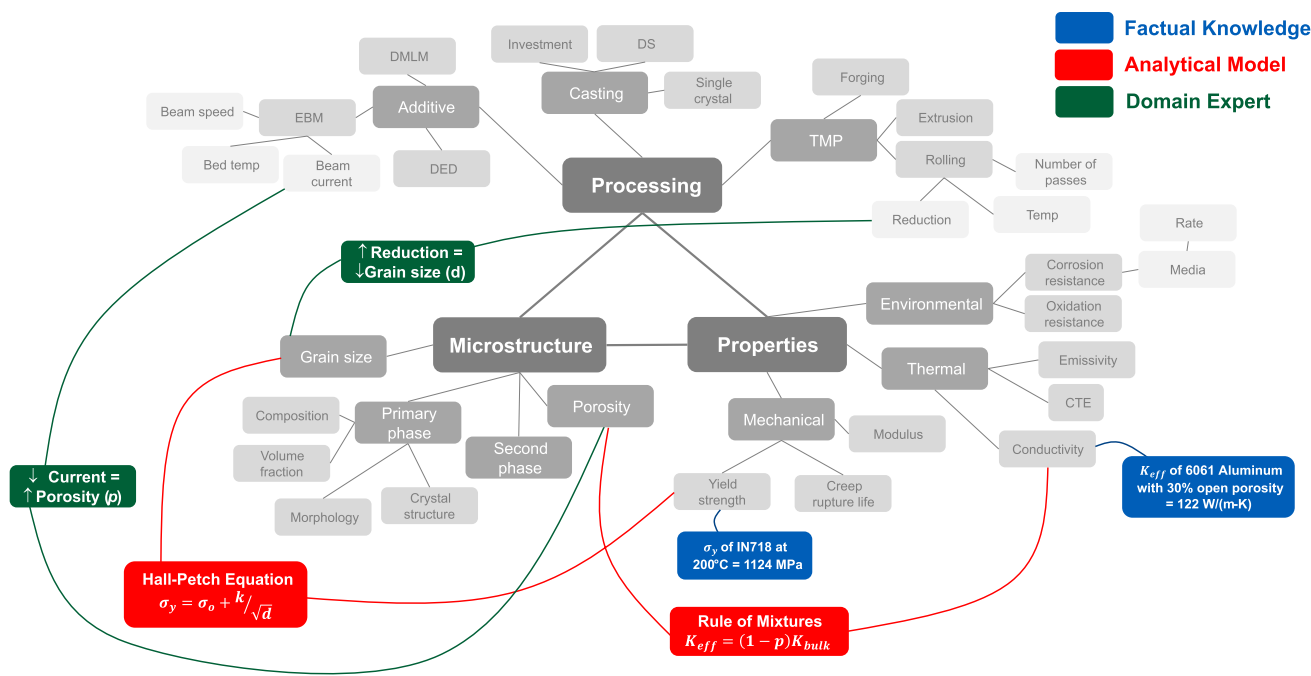
Our Compound Knowledge Graph (CKG) approach involves first capturing factual scientific materials knowledge by extracting (a priori or at query time) and fusing information from multiple complementary public and private materials data sources. Factual materials knowledge includes quantities such as material chemistry, processing details, characterization data, and properties of materials. Capturing diverse factual knowledge requires developing tools to model and align data from different internal and external sources, and building interfaces and connectors to query, retrieve, and merge data from those sources.

Second, our approach involves augmenting the factual materials knowledge with physics-based and/or second wave

<sup>4</sup> <https://www.dbpedia.org/>.

<sup>5</sup> <https://www.wikidata.org/>.

<sup>6</sup> Many of which can be found at <https://matportal.org/ontologies>.



**Fig. 1** Example of a simplified Compound Knowledge Graph containing factual (blue), analytical (red), and domain expert knowledge (green) all fused into a single unified representation. The fusion of these different forms of knowledge into a single CKG will allow a

system to reason and infer new knowledge and make recommendations about new materials to meet increasingly challenging performance requirements

data-driven machine learning models describing relationships between material processing, structure, and properties. Semantic descriptions of analytical models can be described within the CKG and used to derive properties using the corresponding known models and analytical equations, to enable the automatic derivation of unknown properties [27, 28]. Within the CKG, analytical models are linked to the existing factual knowledge by being connected directly to their specific input and output variables, which are themselves defined as properties of entities such as a material. Thus, if a user requests the value of a property that is not explicitly available through the knowledge graph, then if an analytical model or set of models exist that can be used to derive the desired property value, a reasoning engine can execute the model(s) and return the desired value. We use the term ‘reasoning’ to mean any kind of computational processing performed over existing knowledge to infer new knowledge.

Third and finally, our approach involves overlaying those two forms of knowledge (factual and analytical) with expert knowledge in the form of experience and intuition provided by domain experts. Expert knowledge encapsulating previous experiences and intuition about the relationships between different material elements and properties is the most challenging type of knowledge to gather and codify and thus is missing from most knowledge-driven systems but will be critical to enable a truly third wave AI solution. Without capturing and modeling some form of expert

intuition, a system would be unable to contextually reason about areas it has not seen before to draw original conclusions and make novel recommendations.

Once these three forms of knowledge—factual, analytical, and expert—have been fused within a single Compound Knowledge Graph, we will make this knowledge actionable by enabling reasoning and interactive question-answering by materials scientists and other interested users.

A representative example of a Compound Knowledge Graph is shown in Fig. 1. A simplified underlying materials ontology is shown in gray with branches off of the core tenets of materials science—processing, microstructure, and properties—showing some examples of the types of information that could be modeled. This is not meant to be an all-inclusive example and is intentionally limited in scope for clarity. On top of this ontology are examples of knowledge—factual (blue), analytical (red), and domain expert (green)—connecting relevant nodes in the knowledge graph. A domain expert may suggest that an increase in rolling reduction during thermo-mechanical processing (TMP) generally leads to a reduction in grain size of the resulting processed metal. This smaller grain size can be linked, via the Hall–Petch equation, to an increase in low temperature yield strength. If the specific metal of interest were the common nickel-based superalloy IN718, we might further access factual knowledge such as the yield strength at 200 °C is equal to 1124 MPa.

A second example in Fig. 1 links an electron beam additive manufacturing parameter (beam current) to porosity and, subsequently, thermal conductivity. Many more examples like this can be incorporated into increasingly detailed, inclusive, and complicated CKGs to extract useful information and make connections that would be difficult to see with one type of knowledge alone.

## Enabling Technologies

We are building upon two open-source knowledge representation and reasoning packages to build out our vision. First, the Semantics Toolkit (SemTK) [29] facilitates the development of knowledge-driven data management solutions, including enabling data across multiple external sources to be seamlessly accessed as though they were sitting in a single knowledge graph. SemTK enables data to be transparently stored in its most suitable location, while enabling that data to be referenced in semantic domain terms and linked with other disparate datasets [30]. Second, to ease the development of semantic models (‘ontologies’) by non-semantic experts, we use SADL—the Semantic Application Design Language [31]. SADL is a formal, structured English-like language and development environment for authoring semantic models that allows non-semantic domain experts to read, write, and/or provide feedback on ontologies without requiring extensive training in semantic technologies such as the World Wide Web Consortium’s (W3C) Web Ontology Language (OWL). These two technologies together—SemTK and SADL—were used extensively in the case study described in the section “[Progress Toward Our Vision](#)”.

## Sources of Knowledge

To populate the Compound Knowledge Graph, there are three primary sources from which we expect to extract different forms of knowledge, each of which will require different methods and techniques. These are given as follows:

1. Structured databases and data repositories,
2. Unstructured sources such as publications, textbooks, and technical reports, and
3. Human domain experts such as materials scientists.

These three sources each present their own unique challenges to extract meaningful knowledge and are described below in order of expected increasing difficulty.

### Structured Sources

Extracting factual data from databases and other repositories is typically the easiest, as it involves systematically

interacting with application programming interfaces (APIs) that have been built to enable the automatic querying of structured data. This is not trivial, however, because each data source typically has its own unique API and thus requires its own custom connector. Multimodal data fusion enabled by semantic models describing the data and using SemTK as the technical backbone is an effective mechanism to address this challenge as it supports the seamless querying and integration of data from multiple federated data sources [32]. To demonstrate the effectiveness of this approach, we developed a proof-of-concept fusion of data from two publicly available materials repositories—The Materials Project [8] and Materials Data Facility [15]—through a single, easy-to-use graphical user interface with SemTK retrieving user requests through a UI, translating those into queries against the two sources using their own custom APIs, and finally extracting, fusing, and returning the results on demand. This approach can be extended to other external data sources (potentially using OPTIMADE [33]), as well as other internal repositories to link factual knowledge physically or logically into the CKG.

### Unstructured Sources

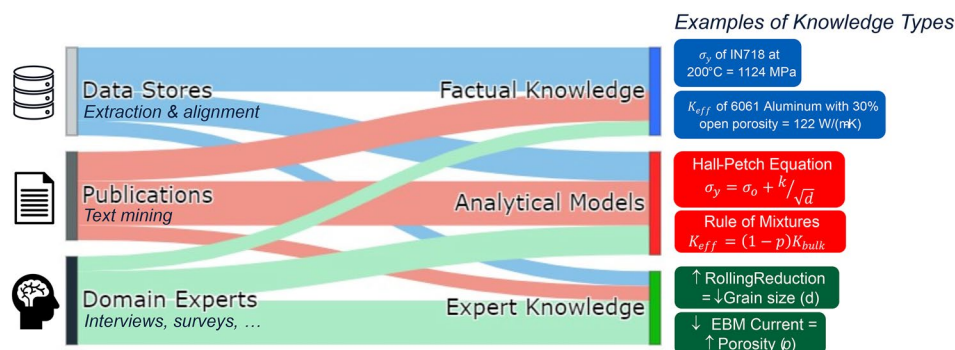
Extracting data from unstructured sources is considerably more challenging, as it requires first structuring highly variable data through techniques such as natural language processing. Textual sources such as books, conference and journal publications, and websites each have their own distinct style, formatting, as well as access controls and usage terms and conditions, making it challenging to build parsers that work across multiple sources. Information extraction modules must be built to parse unstructured text and structure it, and then natural language modules must contextualize the structured text to align it to the existing knowledge in the knowledge graph. We would need to enable the extraction of equations from text, tables, algorithms and pseudo-code in text, and raw source code, which requires the automated understanding of what variables from the attributes in the knowledge graph serve as the input to the resulting model, and what attribute or variable serves as the output.

Extracting knowledge from text, whether factual, analytical models, or expert knowledge, is a challenging problem and while not solved, it has been extensively researched, e.g. [34, 35].

### Domain Experts

The most challenging knowledge to tackle is extracting and codifying human expert knowledge. Typically, expert knowledge is approximate knowledge acquired by experts through years of experience and often expressed as “rules of thumb.” Extracting knowledge from domain experts may sound as

**Fig. 2** Examples of knowledge being extracted from different sources—structured data stores, unstructured text and documents, and human experts—to populate a Compound Knowledge Graph



straightforward as interviewing and surveying a community of expert materials scientists, but in practice is quite challenging. It requires knowing what are effective questions to pose to domain experts to elicit useful information, and, once useful nuggets of wisdom and experience have been documented, it is challenging to digitalize those nuggets and embed them into a knowledge graph. This challenge of building a computable representation of expert intuition has been studied in academic circles [36, 37] and to a lesser extent in industry but remains a largely unsolved problem.

The CKG will use the expert knowledge to bridge the gaps in factual/analytical knowledge to reach an approximate answer when a more precise answer is not available. The CKG may also use expert knowledge as a heuristic short cut when computational resources do not permit a more expensive computation using more precise knowledge. In either case, inconsistency is unlikely to appear, and if it does, precise knowledge would have preference.

Figure 2 schematically illustrates the different sources of information and how they link to the three forms of knowledge—factual, analytical, and expert. Examples from the CKG in Fig. 1 are also included for each knowledge type. It is expected that all forms of knowledge will come from all sources of information but to varying degrees as represented qualitatively by the thickness of each colored connector.

These three forms of knowledge also require different styles of representation in the CKG. Factual data will be represented as direct instance data, analytical knowledge will be represented in a computable form such as executable code or algorithms that can be turned into code on demand, and the expert knowledge will be represented as meta-level information that can be reasoned over.

## Progress Toward Our Vision

The reduction to practice of our CKG vision is motivated by an applied case study seeking to develop new coatings for high pressure steam turbine blades. Steam turbine blades are subject to aggressive oxidation and erosion in service

and the team is attempting to improve upon existing fielded coatings. This case study has all the elements of a ‘typical’ materials science project: a manufacturing process producing a coating with a certain microstructure that responds to external stimuli through its properties. The coatings are produced by a cathodic arc deposition process and characterized by one or more layers of varying chemistry with microstructure and defects quantified through microscopy. Coated test coupons are subject to oxidation and erosion testing to mimic service conditions.

While a traditional second wave AI approach is underway in parallel, an interdisciplinary collaboration between the material development team and knowledge representation and reasoning experts at GE Research is exploring how a third wave AI approach might be implemented—to adapt and reason like a human expert would—to discover new coatings outside the realm of the training dataset. The team is working on multiple aspects of our envisioned CKG-enabled system including a foundational semantic model for the CKG that fuses multiple knowledge modalities, user interfaces for Q&A dialogs and conversational interactions with the CKG, and mechanisms to capture and reason over expert knowledge. The work presented below describes progress to date on these fronts, and while much is left to be done, we believe this lays the groundwork for general implementation of a third wave AI approach for materials development and discovery.

## Modeling the Compound Knowledge Graph

Our objective in modeling the CKG was to start with as generic a foundation as possible. Our intention was not to build an ontology that exclusively addressed the steam turbine blade coating use case, rather we started by building the framework of a general-purpose materials ontology and then fleshed out the specifics required for the coatings use case. We are building a framework that can expand and grow over time to tackle more materials science challenges in the future.

```

CathodicArcDeposition (note "IPD") is a type of Processing,
  described by layer with a single value of type string,
  described by motion with a single value of type string,
  described by numCathodes with a single value of type integer,
  described by cathodeMaterial with a single value of type string,
  described by targetMaterial with a single value of type string,
  described by part with a single value of type string,
  described by cathodeFluxRate with a single value of type Measurement,
  described by cathodePreStress with a single value of type Measurement,
  described by part with a single value of type string,
  described by cleanliness with a single value of type string.

CoatingProperties is a type of Properties,
  described by nonErodedMassChangeInSteam (alias "non-eroded mass change in steam") with a single value of type Measurement,
  described by corrosionRate (alias "corrosion rate") with a single value of type Measurement,
  described by erodedMaterial with a single value of type string,
  described by growthRate with a single value of type Measurement,
  described by fluxRate with a single value of type Measurement,
  described by preStress with a single value of type Measurement,
  described by postStress with a single value of type Measurement.

MaterialType is a class,
  described by materialName (alias "material name") with a single value of type string,
  described by materialTypeDescription with a single value of type string,
  described by baseElement with a single value of type PeriodicElement,
  described by materialSpecification with a single value of type MaterialSpecification.

Chemistry is a class,
  described by chemistryMeasurementMethod with a single value of type string,
  described by elements with values of type ChemistryElement.

ChemistryElement is a class,
  described by chemistryElement with a single value of type PeriodicElement,
  described by minWeightPercent with a single value of type Measurement,
  described by maxWeightPercent with a single value of type Measurement,
  described by avgWeightPercent (alias "average weight percent") with a single value of type Measurement,
  described by minAtomicPercent with a single value of type Measurement,
  described by maxAtomicPercent with a single value of type Measurement,
  described by surfaceRoughness with a single value of type Measurement.

SurfaceRoughness is a class,
  described by surfaceRoughness with a single value of type Measurement,
  described by surfaceRoughness with a single value of type Measurement.

```

**Fig. 3** Example extract of classes and attributes from the steam turbine coatings case study SADL model

The semantic model is written in SADL, an English-like language that made it easy for the materials scientists and knowledge representation and reasoning experts to collaborate [31]. The model was written and evolved through several collaborative sessions over a period of six months. The modeling was driven largely by the materials science domain experts who understand the process important attributes, such as cathode material, bias voltage, and chamber pressure. Once we had the model in place, the data ingestion and equation mapping, while not trivial, were able to move relatively quickly. Examples of the case study ontology in SADL are shown in Fig. 3.

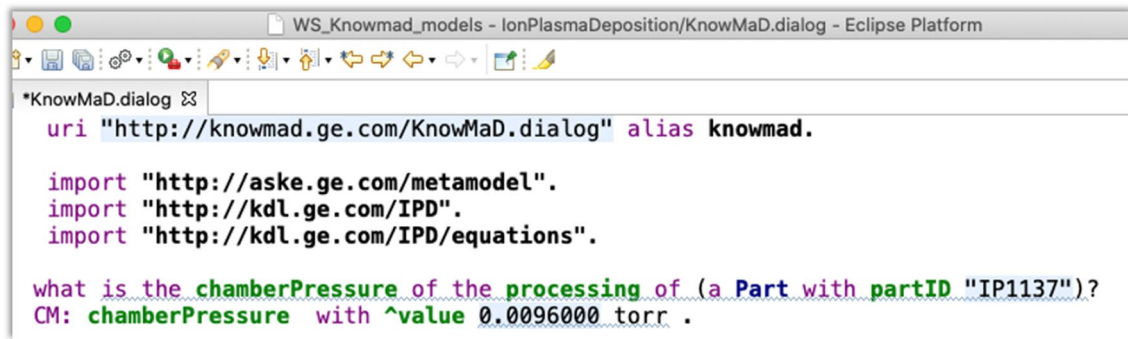
### Fusing Factual and Analytical Knowledge

Analytical models are formally described in terms of the concepts captured in the Compound Knowledge Graph. Each description of an analytical model, e.g., an equation, formally specifies the model's inputs and outputs, explicitly linking the analytical models to the factual data in the knowledge graph.

The CKG specification of an analytical model makes explicit all the references to objects that are normally left

implicit when written for human reading. For example, an equation in a book may say “deposition rate = thickness/time”, implicitly referring to a coating deposition process applied to an object. Here, *thickness* refers to a property of a particular coating microstructure of a particular part. Both the coating microstructure and the part would be represented in the CKG, as is the deposition process that was applied to the part. Similarly, *time* informally refers to the duration of the coating deposition process that was applied to the same part. This deposition process instance is also represented in the CKG, with deposition time as one of its properties. The part instance would be linked to the deposition process instance, thereby making the connection between the coating thickness and the deposition time explicit. Finally, the deposition rate would be formally represented in the CKG as another property of a deposition process associated with the same part. By formalizing these relationships, the CKG enables the correct reasoning required for the automated selection of the necessary models and their assembly into a complex model to compute a desired final property.

The deposition rate equation, encoded in SADL, is given as follows:



```

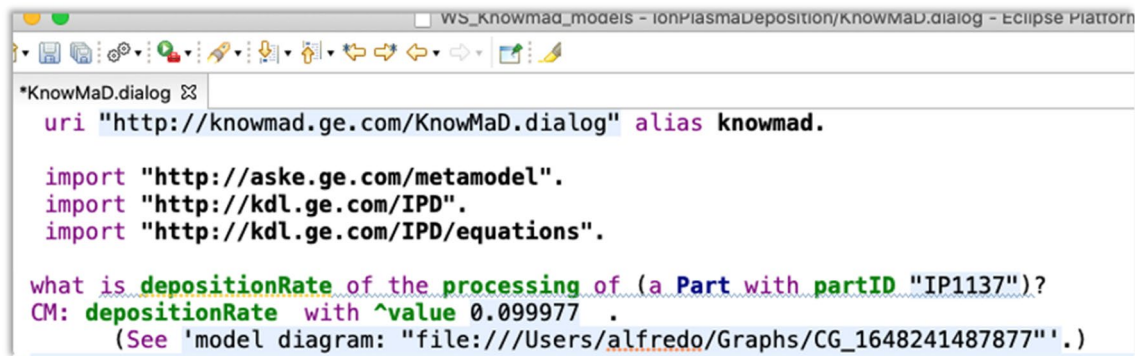
WS_Knowmad_models - IonPlasmaDeposition/KnowMaD.dialog - Eclipse Platform
*KnowMaD.dialog
uri "http://knowmad.ge.com/KnowMaD.dialog" alias knowmad.

import "http://aske.ge.com/metamodel".
import "http://kdl.ge.com/IPD".
import "http://kdl.ge.com/IPD/equations".

what is the chamberPressure of the processing of (a Part with partID "IP1137")?
CM: chamberPressure with ^value 0.0096000 torr .

```

Fig. 4 User queries the CKG in which the answer is stored directly in the graph and the answer can be returned



```

WS_Knowmad_models - IonPlasmaDeposition/KnowMaD.dialog - Eclipse Platform
*KnowMaD.dialog
uri "http://knowmad.ge.com/KnowMaD.dialog" alias knowmad.

import "http://aske.ge.com/metamodel".
import "http://kdl.ge.com/IPD".
import "http://kdl.ge.com/IPD/equations".

what is depositionRate of the processing of (a Part with partID "IP1137")?
CM: depositionRate with ^value 0.099977 .
(See 'model diagram: "file:///Users/alfredo/Graphs/CG_1648241487877"'.)

```

Fig. 5 User queries the system in which the answer is not stored directly in the CKG but an analytical model is available to solve for the answer

```

external depositionRateEq(double totThickness (averageTotalThickness of the
                        actualCoating of the microstructure of a Part {um}),
                        double depTime (depositionTime of the processing of the Part {min}))
returns double (depositionRate of the processing of the Part {"um/min"}): "http://...".
depositionRateEq has expression (a Script with language Python
with script
"def depositionRateEq(totThickness, depTime):
    return float(totThickness) / float(depTime)
").

```

Notice the use of the definite article (“the Part”) to refer to the same object mentioned before. The SADL equation precisely describes each of the two inputs and the output using properties and classes explicitly defined in the CKG ontology. The analytical model also explicitly defines a piece of Python code that can be retrieved to perform the actual calculation.

For the current case study, we identified 25 models in collaboration with domain experts, including equations from the literature and models that call external services

such as *matminer* [18]. We encoded each of the models in SADL, including Python functions implementing the equations. In what follows we show three examples, in increasing order of complexity, of how a user might interact with the CKG, ranging from extracting factual knowledge to retrieving results automatically derived from the fusion of factual knowledge and analytical equations when the desired factual knowledge does not exist.

Figure 4 shows a Q&A dialog interface where the user has queried the CKG for the chamber pressure used in the deposition processing of a particular part identified by its ID.



In this instance, the chamber pressure value for that part is already present directly in the CKG and the system retrieves and displays the value and units to the user.

The next example, shown in Fig. 5, shows a query asking for the deposition rate used in the processing of the same part. In this instance, the system does not find the requested value in the CKG, so it then considers analytical models that may be used to compute the deposition rate for the part. It finds the equation we described earlier, and it also determines that the input values for this analytical model, i.e., the averageTotalThickness and depositionTime associated with this part, are present in the CKG. It concludes that the required inputs for this single equation model are available and proceeds to execute the model, resulting in the automatic fusion of two forms of knowledge—factual and analytical—to answer the user’s question without the user having to explicitly request that this calculation be performed.

The answer in this case includes a link to a dynamically generated diagram (Fig. 6) that shows the structure of the assembled model. The inputs are highlighted in yellow. The equation is shown as a square box with the model’s name, and the output value is an oval.

The system is capable of dynamically assembling sophisticated model chains. Consider a more complex example where the user queries for the fluxRatio of the deposition process for a part. The system dynamically discovers that it can compute the flux ratio from a part’s available property values by chaining five analytic models into the model chain depicted in Fig. 7.

### Capturing and Reasoning Over Expert Knowledge

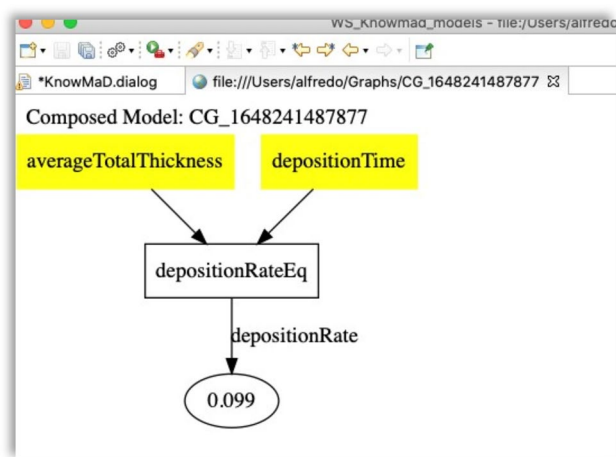
While large volumes of expert knowledge are captured in reports and other textual sources, domain experts themselves are still the best source of such knowledge because they can provide proper contextual information such as sensitivity, controllability, and valid ranges of parameters for domain knowledge. Such information is often lacking in text. Encoding expert knowledge, such as “if a coating has higher macroparticles it will tend to have a higher pre-steam erosion rate,” requires the ability to model directional relationships between attributes in the CKG. We do this by using a rule to programmatically generate a new class corresponding to each relevant property in SADL. This allows us to elevate properties to be first-class objects so we can then capture metadata about properties and relationships between the properties. Once we capture rules that relate properties, we then define rule patterns to reason over linear directional expert knowledge rules. We expect other types of expert knowledge can also be similarly encoded but significant research is needed to understand the spectrum of expert knowledge patterns that must be encoded.

### Enabling Natural Language Interactions

To make the CKG actionable, users must have easy mechanisms through which they can interact with the CKG, pose questions, and receive answers. By using a natural language dialog interface, the CKG will be able to pose questions back to the user to clarify the user’s request and help narrow down to a precise answer. Figure 8 shows example aspirational dialogs with different levels of complexity, from a simple one question-one answer back-and-forth to a more extensive dialog to provide property information.

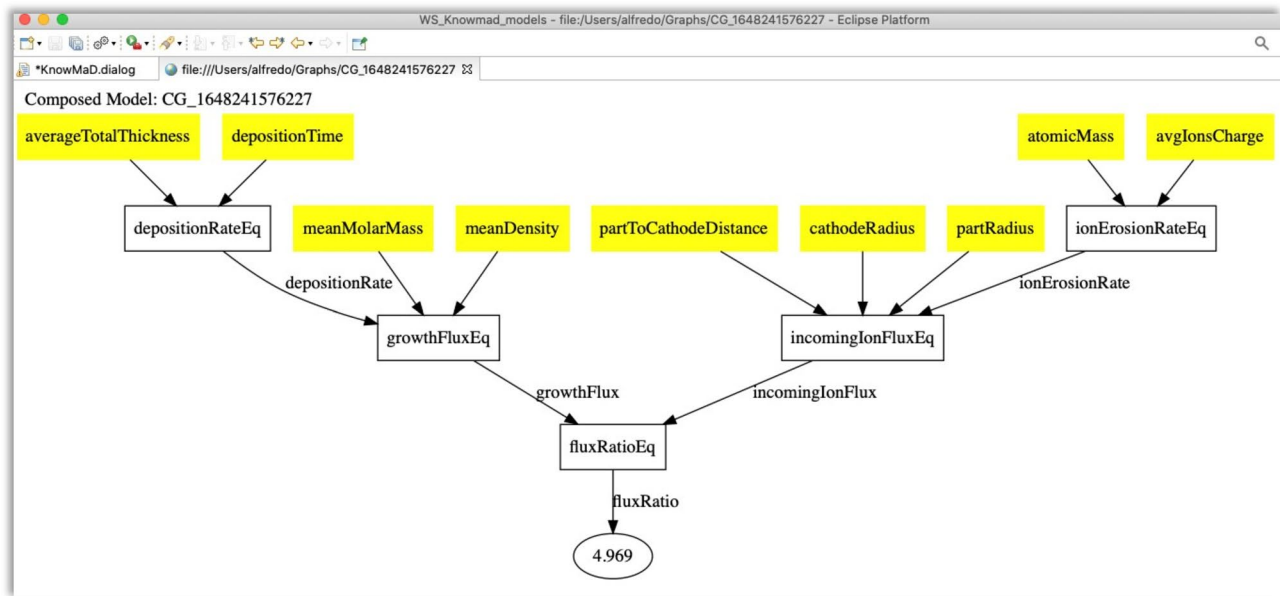
We have reduced to practice an initial natural language-powered interface to provide seamless interactions with the CKG. The natural language interface alleviates the need for materials scientists to learn Semantic Web standards such as the knowledge graph query language SPARQL<sup>7</sup> and/or requiring knowledge of the exact terminology used in the CKG ontology (as required in the examples in Figs. 4 and 5 above), thereby allowing users to pose questions using familiar materials science terms. This component is comprised of three modules: (i) a front-end web interface to capture user questions and display answers (an example of which is shown in Fig. 9), (ii) a question parser module to interpret user questions, including mapping entities in the questions to the appropriate concepts and properties in the CKG, and (iii) a query generation module which leverages the output of the question parser to generate the appropriate SPARQL queries to retrieve answers.

Consider the question “*What partID had the minimum post-steam erosion rate?*”. The question parser extracts the question word (*What*), the requested returns (*partID*),

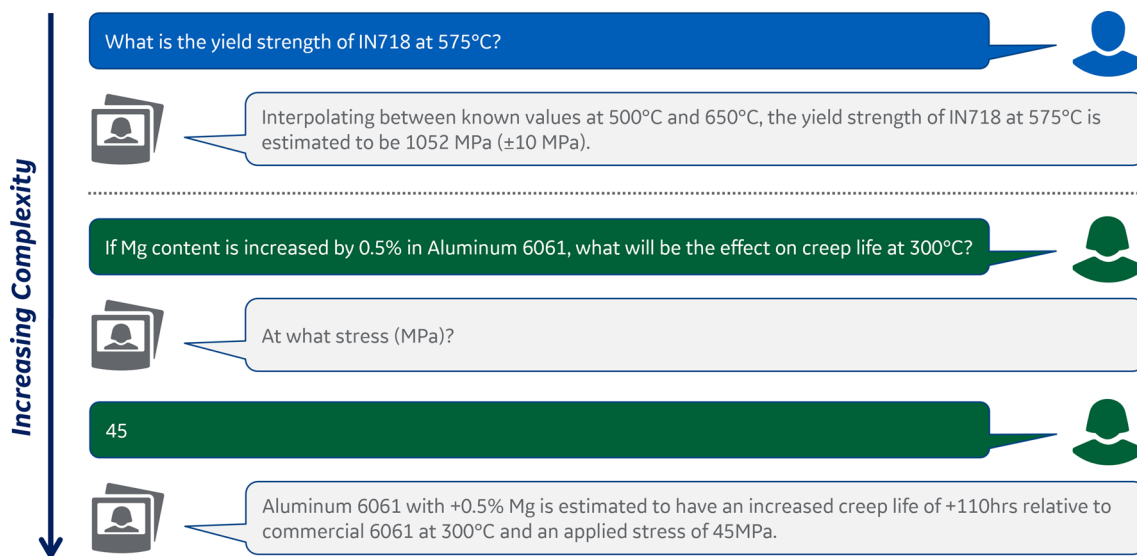


**Fig. 6** Model diagram showing the inputs in yellow, an analytical model being executed as a white box, and finally the output generated as an oval from a chain of equations. In this example, only a single equation is executed

<sup>7</sup> <https://www.w3.org/TR/sparql11-query/>.



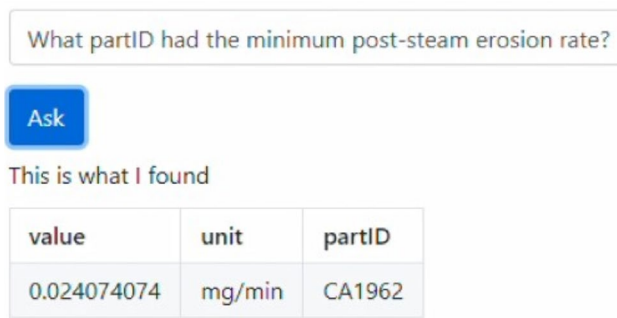
**Fig. 7** Model diagram showing the inputs in yellow, analytical models being executed as white boxes, and final output generated as an oval from a dynamically generated chain of equations. In this example, five analytical models are executed from nine inputs to generate the output



**Fig. 8** Examples of aspirational natural language dialogs between users (blue and green) and an AI assistant enabled by the CKG

constraints/conditions (*post-steam erosion rate*), and any aggregate functions (*minimum*) from the question. The parser further maps the phrases extracted as returns and constraints to either appropriate concepts, properties, or entities in the CKG. The CKG parameters needed for return and/or constraint along with the aggregate functions are passed onto the query generation module. The query generation module uses this information to automatically generate a query to

retrieve the answer from the CKG. The natural language component currently supports questions whose answers can be directly retrieved from the CKG. A natural evolution of this module is the ability to support questions for which answers do not exist in the CKG but can be obtained by assembling and executing equations and models, and eventually answering questions that require reasoning over all three forms of knowledge in the CKG.



**Fig. 9** Screenshot of CKG natural language user interface in which a user enters a question in natural English and a response is returned

## Conclusions and Future Work

Inspired by DARPA’s third wave AI vision, we aim to develop an AI assistant that can perform contextual reasoning and adaptation in areas outside the limitations of the data available in the system. This will be possible only through the fusion of three forms of knowledge—factual, analytical, and expert—into a single Compound Knowledge Graph over which we can reason and infer new knowledge to answer increasingly complex materials science questions.

We have begun reducing our vision to practice, focusing on a steam turbine blade coating use case. Materials scientists and AI experts have partnered to model the CKG in a generalizable way such that the model can be expanded to other materials use cases over time. We integrated factual data with sixteen analytical models embedded in the CKG, as well as nine attributes from matminer [18] for a total of 25 models. We built a structured query parser in SADL to answer both factual and analytical model-driven questions that required the system to automatically determine if an answer could be retrieved through a simple lookup in the CKG or if it required dynamically assembling sophisticated model chains to dynamically calculate new answers on the fly.

We interviewed materials science domain experts to extract expert knowledge and explored techniques to model those in the CKG, and finally built a natural language interface to allow users to submit natural English questions and obtain answers without requiring familiarity with the structure or terminology of the CKG.

While we have made meaningful progress toward the demonstration of our vision, we have only scratched the surface and substantial work remains. There are two major and several minor areas that will require significant research and development.

The first major challenge is we must develop strategies to both collect and semantically model domain expert knowledge so that it can be reasoned over in conjunction with the factual and analytical knowledge. This is particularly

challenging because we must be able to model many different forms of patterns/relationships that humans retain, and we must be able to model the context (circumstances) in which that knowledge applies, and the confidence or uncertainty associated with that knowledge.

The second major challenge is we must enable reasoning and inference over the different forms of knowledge in the CKG in unison to be able to answer complex questions, and we must be able to perform that reasoning in the presence of uncertainty. We need to develop mechanisms for the AI assistant to understand what question the materials scientist is asking, introspectively understand what knowledge it has at its disposal to answer the question, determine what is missing and therefore what it needs to ask back to the human for clarification and guidance, and then pose those questions back to the user in natural language to receive new information. This complex reasoning and inference challenge requires the fusion of second wave machine learning and expert domain knowledge. Further, since the reasoning depends on different nuggets of knowledge each with their own levels of uncertainty, we must understand the compounding effects of individual uncertainty propagation on the final answers/recommendations that the system provides.

There are other challenges that also need to be addressed to make this vision a reality but are not at the same level of complexity as the two mentioned previously. These include enabling data and equation extraction from structured and unstructured sources in a scalable manner so that the CKG can grow without manual intervention. This will require dynamic ontology alignment between the CKG and newly extracted knowledge, which will depend upon advanced text mining and natural language processing.

There is a significant potential for third wave AI to revolutionize materials science, but the challenges are substantial, and success will require significant investment and partnerships between government, industry, academia, and national labs. We recognize that no single organization is going to solve these grand challenges by themselves, and so we need to work together to develop open standards, tools, and ontologies to develop third wave AI solutions and apply them to the most pressing materials science challenges that we as a community and as a nation face.

**Acknowledgements** The authors thank Aida Amroussia, Scott Weaver, and Vipul Gupta for their contributions.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source,

provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Encyclopedia Britannica (2022) silicon carbide. <https://www.britannica.com/science/carbide>.
2. Rosenberg J (2019) A short history of the ball of goo called Silly Putty. ThoughtCo. <https://www.thoughtco.com/the-history-of-silly-putty-1779330>.
3. Moock J, Siu C (2003) Happy birthday: LEXAN resin turns 50! [https://www.gelexan.com/gelexan/turns\\_50.html](https://www.gelexan.com/gelexan/turns_50.html).
4. Long H, Mao S, Liu Y, Zhang Z, Han X (2018) Microstructural and compositional design of Ni-based single crystalline superalloys—a review. *J Alloy Compd* 743:203–220
5. Kennedy R (2019) Ceramic matrix composite technology is GE's centerpiece jet propulsion strategy for the 21<sup>st</sup> century. GE Aerospace | The Blog. <https://blog.geaerospace.com/technology/42869/>.
6. Massie C (2022) Meet the super material helping GE's adaptive cycle engine deliver transformational performance. GE Aerospace | The Blog. <https://blog.geaerospace.com/product/ges-adaptive-cycle-engine-deliver-transformational-performance/>.
7. U.S. National Science and Technology Council (2001) Materials Genome Initiative for Global Competitiveness. [https://www.mgi.gov/sites/default/files/documents/materials\\_genome\\_initiative-final.pdf](https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf).
8. Jain AS, Ong P, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA (2013) Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater* 1:011002
9. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* 65(11):1501–1509
10. Kauwe SK, Graser J, Murdock R, Sparks TD (2020) Can machine learning find extraordinary materials? *Comput Mater Sci*. 174.
11. Holm EA, Cohn R, Gao N, Kitahara AR, Matson TP, Lei B, Yarasi SR (2020) Overview: computer vision and machine learning for microstructural characterization and analysis. *Metall Mater Trans A* 51:5985–5999
12. O'Mara J, Meredig B, Michel K (2016) Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. *JOM*. 68:2031–2034
13. Mat3ra, <https://mat3ra.com/>
14. Multiscale Technologies, <https://multiscale.tech/>
15. Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I (2016) The materials data facility: data services to advance materials science research. *JOM* 68(8):2045–2052
16. Choudhary K, Garrity KF, Reid ACE et al (2020) The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *NPJ Comput Mater* 6:173
17. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput Mater Sci* 68:314–319
18. Ward L, Dunn A, Faghaninia A, Zimmermann NER, Bajaj S, Wang Q, Montoya J, Chen J, Bystrom K, Dylla M, Chard K, Asta M, Persson KA, Snyder GJ, Foster I, Jain A (2018) Matminer: an open source toolkit for materials data mining. *Comput Mater Sci* 152:60–69
19. Launchbury J (2017) A DARPA perspective on artificial intelligence. Information Innovation Office (I2O) DARPA. <https://www.darpa.mil/attachments/AIFull.pdf>.
20. Trask N, Martinez C, Lee K, Boyce B (2022) Unsupervised physics-informed disentanglement of multimodal data for high-throughput scientific discovery. [arXiv:2202.03242](https://arxiv.org/abs/2202.03242) [cs.LG].
21. Singhal A (2012) Introducing the Knowledge Graph: things, not strings. Google. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>.
22. Ehrlinger L, Wöb W (2016) Towards a definition of knowledge graphs. SEMANTICS.
23. Ashino T (2010) Materials ontology: an infrastructure for exchanging materials information and knowledge. *Data Sci J* 9:54–61
24. Elementary Multiperspective Material Ontology (EMMO) (2020) Funded by the European Union Horizon 2020 Research and Innovation Programme. <https://github.com/emmo-repo/EMMO>.
25. Li H, Armiento R, Lambrix P (2020) An ontology for the materials design domain. *The Semantic Web*. 212–227.
26. Zhang X, Zhao C, Wang X (2015) A survey on knowledge representation in materials science and engineering: an ontological perspective. *Comput Ind* 73:8–22
27. Gabaldon A, Chennimalai Kumar N (2019) Knowledge-driven model assembly and execution. *Modeling the World's Systems Conf*.
28. Mrdjenovich D, Horton MK, Montoya JH, Legaspi CM, Dwaraknath S, Tshitoyan V, Jain A, Persson KA. (2020) Propnet: a knowledge graph for materials science. *Matter*. 2.
29. Cuddihy P, McHugh J, Williams JW, Mulwad V, Aggour KS (2018) SemTK: a semantics toolkit for user-friendly SPARQL generation and semantic data management. Posters & Demonstrations, Industry and Blue Sky Ideas at the 17th Intl. Semantic Web Conf (ISWC).
30. McHugh J, Cuddihy PE, Williams JW, Aggour KS, Kumar VS, Mulwad V (2017) Integrated access to big data polystores through a knowledge-driven framework. *IEEE Intl. Conf. on Big Data*. 1494–1503.
31. Crapo A, Moitra A (2013) Toward a unified english-like representation of semantic models, data, and graph patterns for subject matter experts. *Int J Semant Compt* 7(3):215–236
32. Aggour KS, Kumar VS, Cuddihy P, Williams JW, Gupta V, Dial L, Hanlon T, Gambone J, Vinciguerra J (2019) Federated multimodal big data storage & analytics platform for additive manufacturing. *IEEE Intl. Conf. on Big Data*. 1729–1738.
33. Anderson CW et al (2021) OPTIMADE, an API for exchanging materials data. *Sci Data* 8:217
34. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson KA, Ceder G, Jain A (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571:95–98
35. Kumar A, Bharadwaj AG, Starly B, Lynch C (2022) FabKG: a knowledge graph of manufacturing science domain utilizing structured and unstructured knowledge source. *Proc. of the Workshop on Structured and Unstructured Knowledge Integration*. 1–8.
36. Hinrichs TR, Forbus KD (2012) Toward higher-order qualitative representations. *26th Intl. Workshop on Qualitative Reasoning*.
37. Ebert-Uphoff I, Gil Y (2015) Exploring synergies between machine learning and knowledge representation to capture scientific knowledge. *1st Intl. Workshop on Capturing Scientific Knowledge (SciKnow) at the 8th Intl. Conf. on Knowledge Capture (K-CAP)*. 1–9.