Integrating Materials and Manufacturing Innovation
https://doi.org/10.1007/s40192-024-00348-4

THEMATIC SECTION: HARNESSING THE POWER OF MATERIALS DATA

Semantics-Enabled Data Federation: Bringing Materials Scientists Closer to FAIR Data

Kareem S. Aggour[1] · Vijay S. Kumar[1] · Vipul K. Gupta[1] · Alfredo Gabaldon[1] · Paul Cuddihy[1] · Varish Mulwad[1]

Received: 30 October 2023 / Accepted: 23 February 2024

Abstract

The development and discovery of new materials can be significantly enhanced through the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and the establishment of a robust data infrastructure in support of materials informatics. A FAIR data infrastructure and associated best practices empower materials scientists to access and make the most of a wealth of information on materials properties, structures, and behaviors, allowing them to collaborate effectively, and enable data-driven approaches to material discovery. To make data findable, accessible, interoperable, and reusable to materials scientists, we developed and are in the process of expanding a materials data infrastructure to capture, store, and link data to enable a variety of analytics and visualizations. Our infrastructure follows three key architectural design philosophies: (i) capture data across a federated storage layer to minimize the storage footprint and maximize the query performance for each data type, (ii) use a knowledge graph-based data fusion layer to provide a single logical interface above the federated data repositories, and (iii) provide an ensemble of FAIR data access and reuse services atop the knowledge graph to make it easy for materials scientists and other domain experts to explore, use, and derive value from the data. This paper details our architectural approach, open-source technologies used to build the capabilities and services, and describes two applications through which we have successfully demonstrated its use. In the first use case, we created a system to enable additive manufacturing data storage and process parameter optimization with a range of user-friendly visualizations. In the second use case, we created a system for exploring data from cathodic arc deposition experiments to develop a new steam turbine coating material, fusing a combination of materials data with physics-based equations to enable advanced reasoning over the combined knowledge using a natural language chatbot-like user interface.

Keywords Knowledge graphs · Data federation · Materials data architecture · Semantics · FAIR principles · Rapid materials discovery

Introduction

Across disciplines ranging from biosciences to the physical sciences—astronomy, geology, and earth sciences—considerable investments continue to be made into infrastructure that eases the handling of large and often complex scientific data. A key rationale behind the development of such infrastructure is that they help researchers and other experts efficiently, accurately, and collaboratively answer questions over and derive insights from the data, wherein the value of the data lies beyond a single research project or a geographic location. Materials science and manufacturing processes, as with other disciplines, have embraced data-driven approaches to accelerate scientific discovery and innovation. Specifically, artificial intelligence (AI) and machine learning

✉ Kareem S. Aggour
aggour@ge.com

Vijay S. Kumar
v.kumar1@ge.com

Vipul K. Gupta
vipul.k.gupta@ge.com

Alfredo Gabaldon
alfredo.gabaldon@ge.com

Paul Cuddihy
cuddihy@ge.com

Varish Mulwad
varish.mulwad@ge.com

[1] GE Aerospace Research, One Research Circle, Niskayuna, NY 12309, USA

Published online: 09 April 2024

(ML) look increasingly poised to revolutionize materials science research as evidenced by the recent use of AI/ML approaches to accelerate the development of traditional materials [1, 2] as well as their use to discover significantly more complex materials such as high-entropy alloys [3, 4].

Materials data exhibit high diversity not only in their content—from scalar parameters to time series to spectral to categorical to image data—but also in the various means by which this rich content can be sourced, processed, linked, and ultimately be analytically exploited to yield benefit. It is well-documented that the path to materials discovery via collaborative, data-centric approaches, as proclaimed by the Materials Genome Initiative (MGI) [5], is currently bottlenecked by a lack of access to well-structured, high-quality data that prevent materials scientists from harnessing the data holistically [6–8]. To this end, the first generation of materials data platforms such as the Materials Project [9] and Materials Data Facility [10] made concerted efforts to standardize and democratize the way materials data is captured and stored. By providing a centralized data repository equipped with standardized data representations, formats, and web-based access and sharing interfaces, these platforms address the siloed nature of materials science research prevalent within specific sub-disciplines. The updated MGI Strategic Plan recognized the progress made by the materials science community in capturing and managing data in siloes and now sees the next grand challenges as developing a "National Materials Data Network" to integrate distributed data stores into a federated system, encouraging the adoption of the FAIR data principles, and incentivizing the community to adopt and use the federated data infrastructure [11]. We believe that our infrastructure is one approach that could be taken to address the updated MGI Strategic Plan's goals.

Critical scientific discoveries, not limited to just the materials domain, typically build on a more holistic analysis that spans contextualized data from across multiple sub-disciplines and demands more rigor and repeatability in experimental processes. In the context of scientific data, guiding principles such as FAIR [12] prescribe a set of standard approaches and practices to manage data complexity and ultimately make data more usable and equitable. While there is no single agreed upon implementation that operationalizes all the FAIR principles, there have been several attempts to reduce some of them to practice within different disciplines [13–15], each with varying levels of success. Infrastructure dedicated to the FAIR data capture and management of materials data has also been previously demonstrated [16–18].

From a FAIR data management viewpoint, materials science and engineering present many challenges beyond just accessibility. Transforming raw unstructured materials data into meaningful structured quantitative representations of potential linkages between processing, structure, properties, and performance (PSPP) of parts is critical to the successful application of AI/ML techniques for discovering new materials and enhanced properties. For describing data with sufficient contextual detail and for capturing linkages across multimodal data, a FAIR materials data infrastructure should leverage ontologies and semantic knowledge graph technology to structure the complex information [19, 20].

We developed a data repository for additive manufacturing materials and process development data generated from a range of sources at multiple GE sites. We are actively enhancing our federated data infrastructure to encompass materials and manufacturing modalities commonly found in the aerospace and power generation industries. To date, we have developed components of this infrastructure for capturing, storing, and linking data to make it accessible to users for a variety of exploitation techniques including analytics and visualization [21–24]. To make materials data findable, accessible, interoperable, and reusable, we adopt three key architectural design philosophies:

1. **Federated data storage**: Materials data physically resides in one or more decentralized data repositories, each of which are optimized to manage and serve specific types of data.
2. **Knowledge graph-based data integration**: Knowledge graphs model materials data and any contextual metadata in a graphical structure using terms and concepts familiar to domain users who interface with the data. This provides a semantic layer that abstracts the complexities of the underlying federated data repositories away from the consumers of the data.
3. **FAIR data access and reuse services**: An extensible ensemble of data access and reuse services make it easy for materials scientists and domain experts to explore the data, and for data scientists and developers to provision the right data for their use, be it for training a machine learning model or for generating a report.

Through this approach to FAIRification of materials data, we minimize the cognitive load of scientists who no longer need to know where the data are stored or how it needs to be accessed. By maintaining data and contextual semantics using a knowledge graph, we make data more findable and accessible to scientists by allowing data search using domain-driven terminology. From a no-code data access standpoint, scientists simply need to select a set of concepts (and their attributes) from a knowledge graph representative of FAIR data; we then automatically translate such a request into a knowledge graph query that retrieves the corresponding data. Through machine-actionable representations of the data and logical data integration using linked data principles [25], we further enhance its interoperability and reusability across applications.

Merely organizing materials data using a data federation system mediated by a knowledge graph is only one part of the solution. Efficiently discovering relevant data from such a FAIR knowledge graph of materials data is still an open challenge. Besides data storage, the MGI strategic plan [26] emphasized the need for better accessibility and discoverability of data. Hence, as part of our infrastructure's FAIR data services layer, we focus on novel data access and reuse services that enable materials scientists to automatically leverage different kinds of useful contextual metadata to find and use the right data.

In the following sections, we summarize the FAIR principles, explain what knowledge graphs are and how they can help make data FAIR, and then further explain our materials data infrastructure and detail how the set of reusable services simplify the discovery of data from our knowledge graph. We then describe two successful applications built with this architectural approach to make materials data highly accessible to materials scientists within the GE company. In the first case, we created a knowledge graph for additive manufacturing (AM) alloy screening and process parameter optimization and making a large pedigreed AM dataset accessible to parameter developers for gaining insights and optimizing process parameters for new AM materials faster [22]. In the second case, we created a knowledge graph for a coating technology development program to increase the reliability of the steam path by fusing a combination of materials data and physics-based equations to enable advanced reasoning over the combined knowledge [23]. In both success stories, we developed web-based user interfaces, one based on a variety of data visualizations and the other through a natural language chatbot interface, that allowed materials scientists to interact with and utilize the FAIR materials data without having to worry about the specifics of the underlying storage infrastructure.

Background

FAIR Data

The FAIR principles are a set of guidelines aimed at enhancing the management and sharing of scientific data in support of data-driven research and collaboration [12]. An acronym for Findable, Accessible, Interoperable, and Reusable, a summary of the FAIR principles [27] are as follows:

- **Findable**: Data should be easy to discover, for both humans and machines. Metadata and unique identifiers should be used to ensure data can be located efficiently. Clear naming conventions and standardized keywords are essential.
- **Accessible**: Data should be readily available, preferably with clear, open-access policies. Metadata should include information about how to access the data, whether it is through a repository, an API, or other means.
- **Interoperable**: Data should be compatible with various systems and tools. It should use common data standards and formats to facilitate integration and analysis across different domains.
- **Reusable**: Data should be well-documented and structured, making it understandable and usable by others. This includes providing context, describing methods, and ensuring data quality.

Adhering to the FAIR principles is of importance in today's data-driven world for several reasons.

Data Explosion We are witnessing an unprecedented explosion of data across most disciplines, from genomics to climate science. To harness this wealth of information, data must be findable. Researchers need efficient methods to locate and access relevant datasets among this vast ocean of data.

Reproducibility In an era of skepticism and concern over the reproducibility of scientific research, adhering to FAIR principles promotes transparency. Well-documented and reusable data make it easier for others to verify and replicate research findings.

Collaboration Many of today's most pressing challenges require collaboration. FAIR data principles ensure that data from diverse sources can be integrated and analyzed together, enabling insights that were previously impossible.

As the volume and importance of data continue to grow, the FAIR principles for scientific data serve as a vital foundation for responsible data management and utilization.

Knowledge Graphs

A knowledge graph is a structured representation of knowledge that captures relationships between entities within a domain in a semantic and graph-based format that both humans and machines can comprehend. By structuring information as a graph of nodes (entities) and edges between them (relationships), knowledge graphs allow for efficient traversal, query, and exploration of information, and support logical reasoning over the facts in the knowledge graph to draw conclusions and derive new knowledge. The main elements of a knowledge graph are as follows:

1. **Entities**: The fundamental objects or concepts in the knowledge graph, representing everything from people and places to abstract concepts or data points.
2. **Attributes**: Describe properties or characteristics of entities. For example, if an entity represents a material,

attributes could include their name, density, and melting temperature.

3. **Relationships**: Define how entities are connected. Establish semantic links between entities, indicating how they are related to one another. For instance, a material entity can have relationships with other entities, such as "has element" or "processed by."

In practice, knowledge graphs can be realized in multiple ways—the traditionally leading approaches being semantic knowledge graphs and property graphs. Of the two, semantic [or Resource Description Framework (RDF)] graphs are inherently more knowledge-centric in that they rely on World Wide Web Consortium (W3C) standards [e.g., representing knowledge in the form of RDF triples and Web Ontology Language (OWL) axioms] to prioritize interoperability. Property graphs, while more data-centric, can potentially have a semantic overlay atop the graphical data to manage domain knowledge in a more flexible, performant manner. While RDF-based knowledge graphs may have certain limitations with regard to natively representing different forms of materials data in a scalable manner, we still prefer to use such semantic knowledge graphs due to the rich interoperability on offer, the out-of-the-box standardized ability to logically reason over such graphs, and modern semantic virtualization approaches which help circumvent issues of scalability. In theory, however, the application of knowledge graph technology to FAIR data management may potentially be realized using property graphs as well.

Ontologies provide the underpinning for the most popular open semantic knowledge graphs in use today (DBpedia,[1] Wikidata[2]). An ontology is used to define a set of concepts (or classes) within a domain of interest—which become nodes in a knowledge graph, the attributes (or properties) of those concepts, and any links (or relationships) between concepts—which become edges in the graph. Also commonly referred to as a “semantic model,” an ontology by itself is an important form of knowledge as it represents how experts within a domain think about their field and codifies the vocabulary they use to describe the concepts in that field. Several ontologies have been proposed to represent and capture materials data,[3] although the breadth of coverage and maturity of the models varies widely [28–30]. More information on materials ontologies can be found in a survey paper by Zhang et al. [31]. In addition to a semantic model of a domain, a knowledge graph usually also comprises data that represent specific instances of knowledge.

Knowledge graphs constructed from semantic models are well-aligned with key FAIR principles. In comparison with FAIR data modeled using traditional, simpler data models, FAIR data represented through semantic knowledge graphs enable relatively greater findability and reuse, as described below:

- By allowing domain experts in materials science and manufacturing to model their data and any contextual information about the data (i.e., metadata) using a structure and terminology familiar to them, knowledge graphs permit efficient search and information retrieval. By encoding domain taxonomy within semantic models, retrieving data of interest becomes dramatically simplified, and knowledge graph queries can be constructed via no-code or visual drag-and-drop techniques resulting in enhanced *Findability* of data.
- When constructed in accordance with linked data principles [25], semantic knowledge graphs assign unique identifiers (URIs) to all entities, ensuring that information about a specific entity can be maintained consistently and unambiguously paving a path to better data *Interoperability*.
- To address siloed data, knowledge graphs can further be virtualized such that instance data corresponding to the semantic model is maintained “externally” within one or more underlying data stores, resulting in a logical data architecture referred to as a semantic data fabric. Ontologies are then used to mediate querying and access to the data via a technique known as ontology-based data access (OBDA) [32]. By providing a single common interface and federated mechanism to interact with data across many storage systems and APIs, and by fusing this data on-demand in the context of curated domain knowledge, knowledge graphs can improve overall *Accessibility* to and *Reusability* of data by abstracting away many complexities related to data management.

Materials Repository Architecture

Our materials repository architecture is comprised of three layers as shown in Fig. 1. At the bottom is a federated data storage layer wherein multimodal materials data are captured and stored within distinct repositories, where each repository is dedicated to a specific type of data. Atop this is a knowledge graph that serves as a layer of abstraction over the storage layer and is also responsible for annotating the materials data with semantic context to make the data FAIR. The topmost, consumer-facing layer is responsible for facilitating easy access to FAIR data structured as a knowledge graph. This layer comprises a library of smart data access services to further abstract away the complexity of interacting with a

[1] https://www.dbpedia.org/

[2] https://www.wikidata.org/

[3] Many of which can be found at https://matportal.org/ontologies

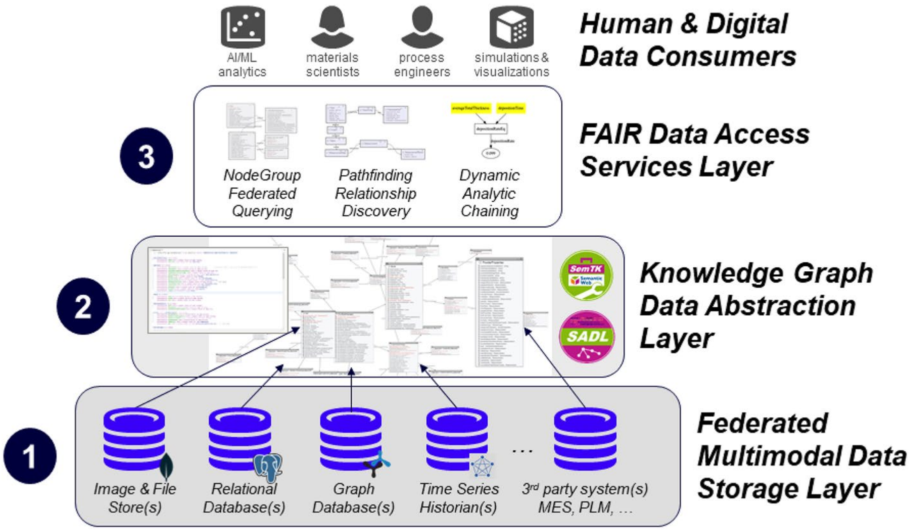

AI/ML analytics
materials scientists
process engineers
simulations & visualizations
Human & Digital Data Consumers
3
NodeGroup Federated Querying
Pathfinding Relationship Discovery
Dynamic Analytic Chaining
FAIR Data Access Services Layer
2
SemTK
SADL
Knowledge Graph Data Abstraction Layer
1
Image & File Store(s)
Relational Database(s)
Graph Database(s)
Time Series Historian(s)
3rd party system(s) MES, PLM, …
Federated Multimodal Data Storage Layer

Fig. 1 Overall architecture of our knowledge graph-driven FAIR data infrastructure for materials and manufacturing data

semantic knowledge graph from materials scientists, engineers, and data scientists.

Federated Materials Data Storage

The multitude of sub-disciplines associated with materials science and manufacturing generate, consume, and analyze very heterogeneous forms of data—experimental data results are often maintained in spreadsheets or other structured tabular formats for analysis and visualization. The experimental metadata, on the other hand, may be captured either in the form of unstructured text or as semi-structured key-value pairs. Many computational materials science tools generate output in nested, semi-structured JSON-like formats, some others in CIF, XML, POSCAR, or text file formats. Materials characterization routines may target optical or SEM micrographs and other forms of image data, or spatial and spectral data in text file formats. Monitoring the manufacturing of a part in real time produces large volumes of time-series data at high rates.

Early community-scale materials data platforms either (i) only supported a limited set of data types—e.g., by transforming all experimental metadata into some common internal format, thereby imposing some structure on the data, or (ii) maintained all forms of data as generic files or binary objects devoid of any structure—e.g., a set of micrograph images archived and stored as a file on an open data repository like Zenodo. Common to both approaches, all data were stored and served out of single centralized repositories.

Given the diversity in materials data, it is impractical to reduce all types of data to a single structure, even when the schema allows for some flexibility. On the other hand, stripping away any structure inherent in the data would end up making the data less findable and interoperable. Thus, in contrast with prior approaches, our federated materials repository integrates multiple data storage technologies and systems such that each different type of data is captured in a repository optimized for that data format—e.g., structured tabular data are stored in a relational database, time-series data in a specialized time-series database system, images in an object store or a specialized array database, and so on [22]. Not only does this approach minimize the overall storage footprint and maximize the query performance for each data type, but it also allows us to leverage the data durability and protection mechanisms within each underlying storage system out of the box to guard against potential data loss and expensive data regeneration. Besides federating data storage systems, we will extend our repository to include accessing specialized 3rd party systems as well as external sources of data that may only be accessible via APIs or other endpoints.

Knowledge Graph for Materials Data FAIRification

Our federated data storage layer can be extended as needed to store newer forms of materials data. However, data stored in this way still does not adhere to FAIR principles, and data across the different stores are not linked by default. Our solution to this problem is the second layer of our architecture—a knowledge graph integration layer—which models the data structure and relationships within and across the data repositories. The knowledge graph layer enables data to be semantically annotated with rich, contextual metadata and linked to other data based on the metadata—both critical steps toward ensuring FAIR compliance. The knowledge graph integration layer also models and captures metadata describing the underlying data stores and how to query and interact with each data store. Through this layer of abstraction, our materials repository provides a single logical interface encompassing a wide range of diverse, distributed data sources. The knowledge graph itself is maintained within a

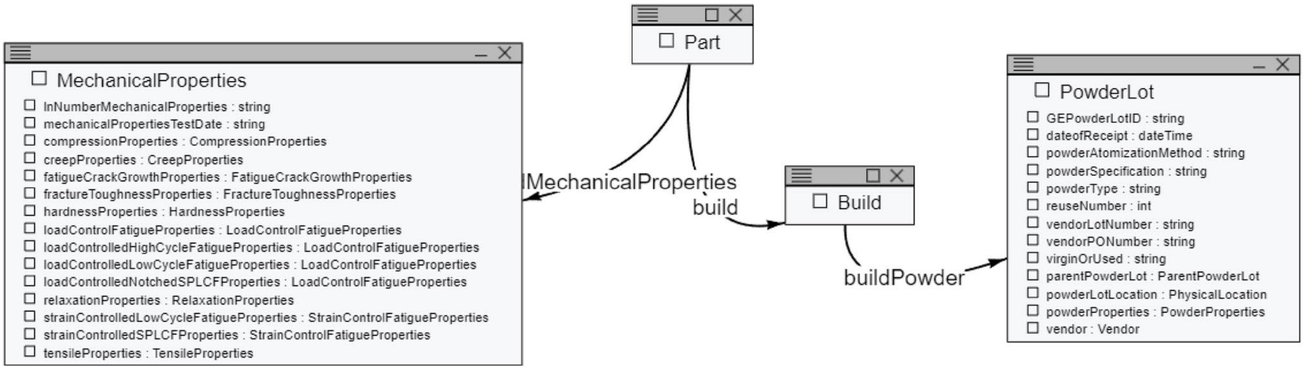

□ Part
□ MechanicalProperties
□ lnNumberMechanicalProperties : string
□ mechanicalPropertiesTestDate : string
□ compressionProperties : CompressionProperties
□ creepProperties : CreepProperties
□ fatigueCrackGrowthProperties : FatigueCrackGrowthProperties
□ fractureToughnessProperties : FractureToughnessProperties
□ hardnessProperties : HardnessProperties
□ loadControlFatigueProperties : LoadControlFatigueProperties
□ loadControlledHighCycleFatigueProperties : LoadControlFatigueProperties
□ loadControlledLowCycleFatigueProperties : LoadControlFatigueProperties
□ loadControlledNotchedSPLCFProperties : LoadControlFatigueProperties
□ relaxationProperties : RelaxationProperties
□ strainControlledLowCycleFatigueProperties : StrainControlFatigueProperties
□ strainControlledSPLCFProperties : StrainControlFatigueProperties
□ tensileProperties : TensileProperties
lMechanicalProperties
build
□ Build
buildPowder
□ PowderLot
□ GEPowderLotID : string
□ dateofReceipt : dateTime
□ powderAtomizationMethod : string
□ powderSpecification : string
□ powderType : string
□ reuseNumber : int
□ vendorLotNumber : string
□ vendorPONumber : string
□ virginOrUsed : string
□ parentPowderLot : ParentPowderLot
□ powderLotLocation : PhysicalLocation
□ powderProperties : PowderProperties
□ vendor : Vendor

Fig. 2 Small extract of additive manufacturing ontology showing four classes, their relationships, and attributes (properties) of two classes—Powder Lot and Mechanical Properties

graph database system or semantic triple store within our federated storage layer.

To facilitate development and use of knowledge graphs for managing materials data, we build on open-source semantic technology solutions. First, to ease development of ontologies by non-experts, we use the Semantic Application Design Language (SADL) [33]. SADL is a formal, structured English-like language and development environment for authoring semantic models that allow domain experts to read, write, and edit ontologies without requiring extensive training in semantic web technologies such as the W3C's OWL [34]. Second, the Semantics Toolkit (SemTK) [35] facilitates rapid and scalable development and deployment of knowledge-driven applications once a semantic model has been established, including enabling data across federated data stores to be seamlessly accessed as though they were residing in a single storage system. Together, these two technologies, SADL and SemTK, were used extensively in the success stories described in Sect. "Success Stories."

Semantic Application Design Language

The Semantic Application Design Language (SADL) is a structured, formal English-like language for authoring semantic models, allowing non-semantic domain experts to author ontologies that model their domain of interest using their community's commonly accepted vocabulary. SADL is available as an open-source Eclipse plugin and can be used to directly compile OWL files, the de facto standard for instantiating models in a semantic triple store for capturing linked data.

Within a semantic triple store, ontologies are frequently populated with data that capture details about specific instances or entities within the domain. For example, an additive manufacturing ontology may define the structure and attributes of a powder lot and be populated by instance data capturing details about each powder lot maintained at a manufacturing site. Figure 2 shows a visual representation of a subset of such a semantic model, including classes representing 3D-printed parts, the build used to produce the part, attributes about the specific powder used for the build, and mechanical properties of the part.

Semantics Toolkit

The Semantics Toolkit (SemTK) is a framework that simplifies the rapid creation and utilization of knowledge-driven services and applications. SemTK offers features and functions that simplify the use of and interaction with knowledge graphs, that can make it dramatically easier for materials scientists to interact with, explore, manage, and use knowledge graphs to ingest, retrieve, visualize, check the validity of, and derive benefit from materials data. These include:

- SPARQLgraph,[4] a user interface with features for browsing ontologies, for exploring and interacting with the knowledge graph, for building and storing queries, and for building and storing ingestion templates to map data ingestion files to a knowledge graph. Both queries and ingestion templates can be stored in SemTK via the UI and embedded in knowledge-driven applications.
- Ontology-based Data Access (OBDA) [32]: SemTK currently supports relational data, time-series data, file storage, and can be extended to flexibly support other data sources and data types as well. SemTK enables data to be transparently stored and queried in its most suitable location, while enabling it to be referenced in semantic domain terms and to be linked with other disparate datasets. Supporting disparate storage is important because many data types critical to materials science (e.g., time-

[4] SPARQL is a standard query language for querying graphs in W3C-compliant semantic triple stores.

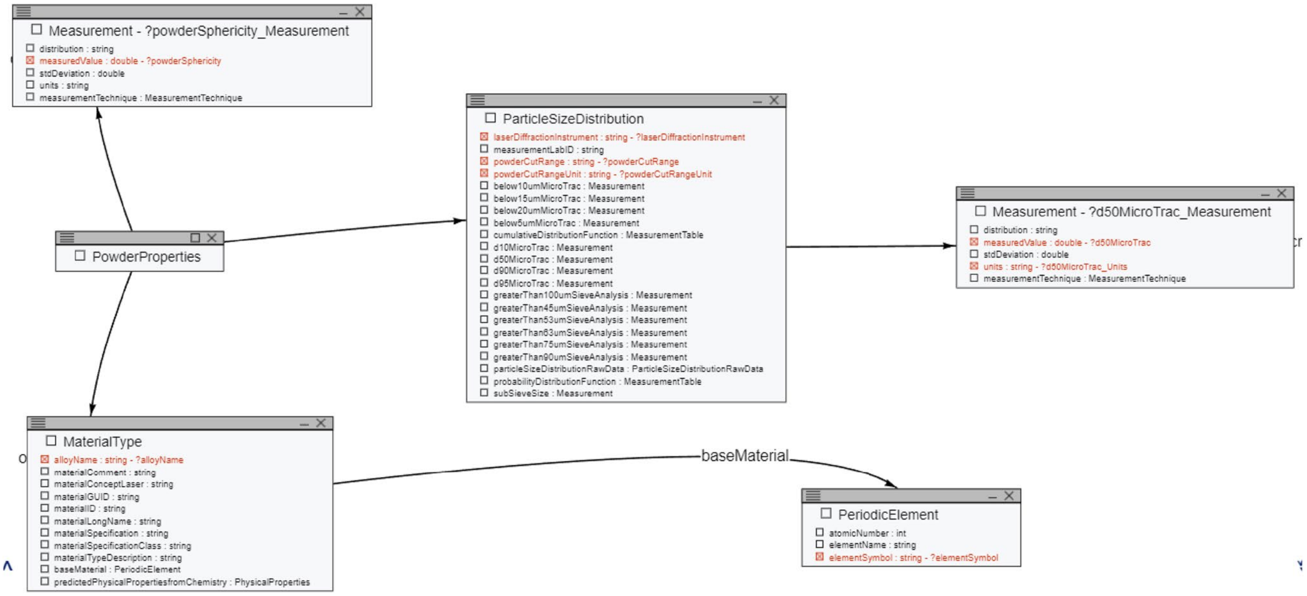

☐ Measurement - ?powderSphericity_Measurement
☐ distribution : string
☒ measuredValue : double - ?powderSphericity
☐ stdDeviation : double
☐ units : string
☐ measurementTechnique : MeasurementTechnique
☐ PowderProperties
☐ ParticleSizeDistribution
☒ laserDiffractionInstrument : string - ?laserDiffractionInstrument
☐ measurementLabID : string
☒ powderCutRange : string - ?powderCutRange
☒ powderCutRangeUnit : string - ?powderCutRangeUnit
☐ below10umMicroTrac : Measurement
☐ below15umMicroTrac : Measurement
☐ below20umMicroTrac : Measurement
☐ below5umMicroTrac : Measurement
☐ cumulativeDistributionFunction : MeasurementTable
☐ d10MicroTrac : Measurement
☐ d50MicroTrac : Measurement
☐ d90MicroTrac : Measurement
☐ d95MicroTrac : Measurement
☐ greaterThan100umSieveAnalysis : Measurement
☐ greaterThan45umSieveAnalysis : Measurement
☐ greaterThan53umSieveAnalysis : Measurement
☐ greaterThan63umSieveAnalysis : Measurement
☐ greaterThan75umSieveAnalysis : Measurement
☐ greaterThan90umSieveAnalysis : Measurement
☐ particleSizeDistributionRawData : ParticleSizeDistributionRawData
☐ probabilityDistributionFunction : MeasurementTable
☐ subSieveSize : Measurement
☐ Measurement - ?d50MicroTrac_Measurement
☐ distribution : string
☒ measuredValue : double - ?d50MicroTrac
☐ stdDeviation : double
☒ units : string - ?d50MicroTrac_Units
☐ measurementTechnique : MeasurementTechnique
☐ MaterialType
☒ alloyName : string - ?alloyName
☐ materialComment : string
☐ materialConceptLaser : string
☐ materialGUID : string
☐ materialID : string
☐ materialLongName : string
☐ materialSpecification : string
☐ materialSpecificationClass : string
☐ materialTypeDescription : string
☐ baseMaterial : PeriodicElement
☐ predictedPhysicalPropertiesfromChemistry : PhysicalProperties
baseMaterial
☐ PeriodicElement
☐ atomicNumber : int
☐ elementName : string
☒ elementSymbol : string - ?elementSymbol

Fig. 3 NodeGroup from a data scientist's perspective

series data and image data) are not suitable for storage directly within a knowledge graph.

SemTK has previously been deployed in production on both internal servers as well as in the Amazon AWS GovCloud as part of a system managing gas turbine test data for GE's Power business [36]. Both SADL and SemTK have been open-sourced, and so they can be used by the reader to develop the same materials repository architectural approach.

Knowledge Graph Abstractions

The third and final layer of our materials repository architecture is a set of abstractions that sit atop the knowledge graph that greatly simplifies interacting with the knowledge graph. These abstractions include, but are not limited to, NodeGroups, pathfinding, and dynamic analytic execution, and are described next.

NodeGroups

The NodeGroup is a key building-block for querying in SemTK [37]. The NodeGroup represents a subgraph of interest and lends itself well to visualization and drag-and-drop query-building. In addition to the subgraph structure, it contains annotations that enable it to support automatic generation of multiple types of queries. The same NodeGroup can be used to generate SPARQL queries to select a table, build a results graph, delete, or insert data.

The SemTK SPARQLgraph user interface supports the creation of NodeGroups with a visual editor. Each node represents a class and shows the data and object properties available. The user can click properties or drag new classes to quickly make connections that are valid given the ontology. Properties and nodes can be flagged for return and/or deletion. They can be constrained with additional clauses of SPARQL. Properties can also be flagged to be constrained at runtime. When a runtime constraint is applied, SemTK will generate a query to retrieve the valid values given the data in the knowledge graph. NodeGroups are also designed to be easily reused across different applications.

Figure 3 shows a NodeGroup used to retrieve data where ML techniques are used to analyze and gain insights into powder properties such as flowability and apparent density, as a function of particle size distribution of a powder lot and powder sphericity for a selected alloy chemistry. Once this NodeGroup is created by a data scientist, it can be stored and retrieved by name, freeing an end user from needing to understand the data model and manually writing the query.

NodeGroups enable a separation of roles (data engineer, data scientist, and application developer) involved in the end-to-end application lifecycle and ensure consistency in how each of them accesses data. This separation alone has reduced development times substantially in several industrial scenarios.

Pathfinding

SemTK includes a pathfinding algorithm to find connections between data and concepts in the knowledge graph. This is used frequently during the creation of NodeGroups. The user can drop concepts onto the canvas, and the tool suggests multi-hop connections between the concepts, thus greatly improving the ease of building a NodeGroup. The algorithm is also a key component used to connect concepts for automated NodeGroup creation used in the question-answering interface discussed in the second success story described in Sect. "Success Stories."

Pathfinding is based on the A* algorithm, but with additional rules to limit the number of hops searched and the amount of time allowed for a given search. In both manual and automatic modes, there is an implied proximity (e.g., 5 hops) and time limit reasonable built into the interface. If a path cannot be found within those constraints, the algorithm returns a null path.

Pathfinding can operate in two different modes. In its basic mode, the algorithm uses only the model to determine valid paths. However, when data are available and the pathfinding is meant to search for existing data, it can be operated in a mode that takes existing data into account. In this mode, a property must have instances in the data to be included in the path. For the sake of efficiency, the current algorithm only searches each property individually, and the full graph pattern may still not represent existing data. Nonetheless, this algorithm is indispensable in multiple applications, providing another large improvement in the ease and efficiency with which queries can be built.

Dynamic Analytic Chaining

A third abstraction in our repository architecture is dynamic analytic chaining. In our infrastructure, analytics can take the form of physics-based equations, computational and/or ML models, allowing us to embed domain knowledge directly into the knowledge graph. Analytical models are formally described in terms of the (materials science) concepts modeled in the knowledge graph. Each description of an analytical model, e.g., a physics-based equation, formally specifies the model's inputs and outputs, explicitly linking the analytical models to the concepts and any instance data in the knowledge graph.

Our approach includes executable code such as Python functions, which allow the analytic models to be executed on demand [38]. The system can perform inference over the knowledge graph and determine if the user's questions can be answered based on the data that are available within the knowledge graph. If the data alone are insufficient, the system reasons over the combination of data and analytics and determines if a single analytic or combination of analytics chained together can answer a query. The system can organize and execute a chain of analytics in the appropriate sequence, using the available data in the knowledge graph as inputs, until an answer is derived.

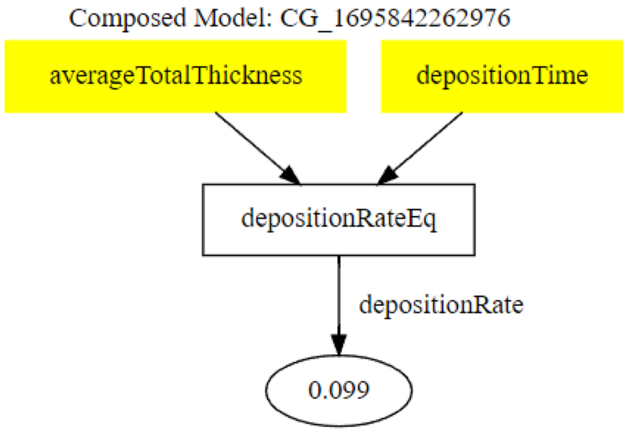


Fig. 4 Dynamic analytic chaining example

This ability to dynamically chain analytics allows physics-based knowledge as well as machine learning and computational models to be embedded with materials science data, and for materials scientists to use the knowledge and models without having to explicitly request a run sequence. We have used this capability, as will be described later, to enable materials scientists to automatically derive unknown properties of known materials and processes [23]. This approach also provides an answer which is completely explainable. A visual representation of the analytic chain and raw inputs are displayable to the user along with the calculated result. An example of this visualization is shown in Fig. 4.

Related Work/Prior Art

FAIR data infrastructure in support of data-centric scientific discovery has been proposed and explored, with diverse implementation strategies and success stories primarily published within the biomedical and life sciences [13, 39, 40], health care and clinical settings [41, 42], geosciences [14], and earth sciences [17]. Development of such infrastructure is often driven by dedicated consortia in collaboration with standards bodies [43]. Scheffler et al. [14] outline how the materials sciences, specifically, could benefit from FAIR data management. Kalidindi et al. [44] describe the implementation of different software components—data management tools, data analytics frameworks, and an e-collaboration network—for accelerated materials innovation in keeping with the TMS recommendations [45]. Kadi4Mat [46], an open-source implementation of a FAIR research data infrastructure focused on combining the best features of electronic laboratory notebooks and research data repositories to support materials scientists through the research and publication process. In [47], the authors present their "laboratory of Babel" vision of integrated materials data management for autonomous discovery, where they discuss data challenges and strategies at the laboratory-level, group-level, and community-level, with FAIR aspects coming into play primarily at the community level. He et al. [48] present a no-code approach to parse output from popular computational materials science tools to auto-generate FAIRified data. The National Materials Data Management and Service (NMDMS) platform [17], within China's Material Genetic Engineering initiative, is implemented as a country-wide repository for materials data following FAIR principles for modeling data.

Federated Data Storage. Over the past decade, data-driven enterprises are increasingly adopting a "no one size fits all" mantra [49] wherein concerted efforts are afoot to unbundle or replace traditional data stacks (e.g., data warehouses for structured tabular data and NoSQL stores for semi-structured and unstructured data) with more "fit-for-purpose," cloud-native data infrastructure. This has resulted in newfound ideologies such as composable data architectures and fragmented ecosystems such as the "modern data stack" that seek to create data infrastructure from the ground up to suit the individual data needs of a given enterprise. A core tenet of this approach is flexibility—to develop data platforms with APIs that integrate components for storing and analyzing different types of data. In an interdisciplinary research field such as materials science, given the heterogeneity of materials data, it is impractical to establish a single schema or format that covers all types of data. Hence, for storage, our philosophy is one of the data federation, wherein data of a given type are stored in a repository that is optimized to handle that type. A federation layer then transparently handles accessibility to all underlying repositories (including, potentially 3rd party/external data sources) via suitable APIs. We use as many kinds of different repositories as needed, but also as few as sufficient to handle the different kinds of materials data. Federating data is currently not as widely practiced by the materials science community. An exception among existing FAIR data infrastructure implementations is the hybrid data storage system (HDSS) layer within NMDMS [17], which comes closest to our data federation approach.

Materials Knowledge Graphs. A key requirement for implementing FAIR principles within any domain is the annotation of the scientific data with rich, contextual metadata. Ghiringelli et al. [19] present a detailed report on metadata extraction and representation for computational materials science and experimental workflows. Part of the report discusses the role of materials science ontologies for data FAIRification. Kalidindi et al. [20] also propose the need for ontologies and linked data to truly realize FAIR principles for materials data. Beyond extraction and schematization, our strategy is one of maintaining materials metadata in the form of semantic knowledge graphs. While this approach has been prevalent within the life sciences [50, 51], there have been fewer instances of FAIR materials data infrastructure based on knowledge graphs. The Materials Knowledge System (MKS) and its subsequent implementation as Python-based software (PyMKS) [18] is an early implementation of materials data structured in the form of a semantic knowledge graph but does not address federated data. More recently, the CRUX platform [52] expands materials knowledge graphs beyond materials data to additionally include metadata about resources (analysis scripts, etc.), sources (computations and experiments), and from scientific workflows. Our materials knowledge graph goes one step further and captures contextual knowledge such as the organization of materials data in a federated storage layer, and explicitly includes metadata about physical and analytical equations within the domain.

FAIR materials data discovery. Once FAIR materials data is organized in the form of a semantic knowledge graph for interoperability and machine-readability, one still needs to suitably query the graph to find and access relevant data. The NMDMS [17] platform allows discovery over federated data via full-text queries, SQL-like queries, and knowledge graph-like queries but, to our knowledge, does not maintain data as a semantic knowledge graph. The CRUX platform [52] parses keywords and materials workflow declarations into native Gremlin graph pattern queries for execution by a JanusGraph engine. Ontology-based data access (OBDA) [32] is an established paradigm in which querying of a data source is mediated by a high-level domain ontology. Ontology-based Data Federation (OBDF) [53] extends this idea by combining OBDA with a federation layer to allow querying in domain terms over linked data from across disparate data repositories. Our FAIR materials data infrastructure enables federated querying across a knowledge graph spanning data stored across multiple storage repositories. This is enabled, in part, because of novel FAIR data access abstractions and services that help auto-formulate graph queries by combining appropriate materials data with contextual metadata stored in our knowledge graph.

Success Stories

We have successfully used this architectural approach in two very different materials applications at GE (as well as in non-materials applications, e.g., [21, 36]). In this section, we give a brief description of the two materials-specific applications. In the first, we created a repository of feedstock material properties, manufacturing process parameters, physical testing, material characterization, and inspection data for rapidly introducing new alloys to the additive industry [22]. In the second, we created a repository of materials property and performance data, manufacturing process parameters, and material characterization data for a steam path coating material development use case that fused a combination of data and physics-based equations to enable advanced reasoning over the combined knowledge [23].

Additive Manufacturing New Material Introduction

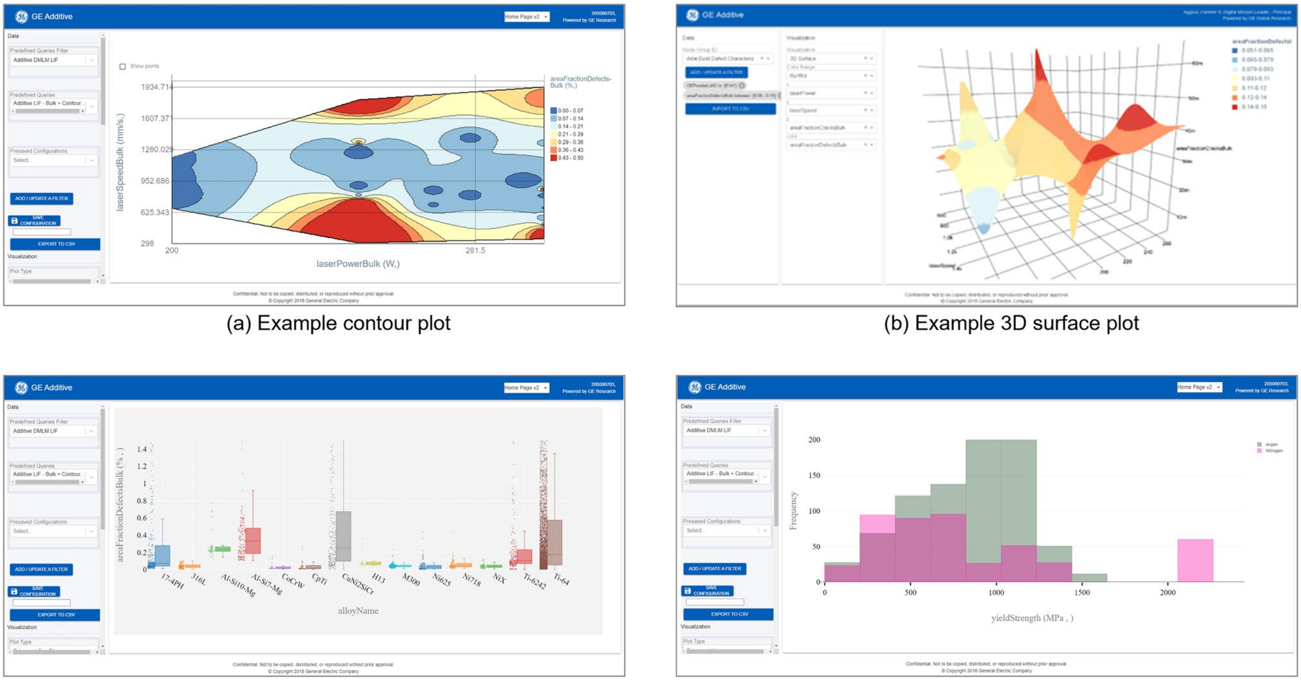
Additive manufacturing is the process of building a part by successively adding raw material via one of a variety of 3D printing techniques [54]. Additive technologies can yield lighter yet sturdier parts that produce less manufacturing waste, often through novel part geometries that are highly challenging if not impossible to manufacture using traditional techniques. This combination of greater manufacturing efficiencies and part performance is having a notable impact across almost every industry [55]. While there are many opportunities for developing new materials and part designs, the full potential of additive manufacturing has yet to be achieved. One major challenge with introducing a new material to additive manufacturing is the identification of an optimal set of process parameters (e.g., laser power, laser speed, beam spot size, and layer thickness, among many others, for a laser powder bed fusion-based process) that will produce parts quickly (fast build rate) with high quality (low number density and volume fraction of anomalies such as voids and cracks). For predicting process outcomes, scientists and process engineers use visual analytics techniques to circumvent a potentially combinatorial search space and a lack of governing equations for a given material. These analytics approaches rely heavily on finding and assembling the right datasets by drawing from multiple distinct sources.

Previously, this was an ad hoc, manually intensive process where scientists were burdened by the need to source each individual piece of the data themselves, and without a digital paper trail for tracing the genesis of an assembled dataset. Moreover, this process was brittle and prone to errors such as using incorrect data (brought about by either ill-formed queries or communication mismatches or data naming conventions that were either lacking or ambiguous) or fusing data for visualization purposes without suitable contextual information. Digitalizing such low-level "data mechanics" would help improve the overall effectiveness of our process engineering teams.

To address these drawbacks and to help deliver on the promise of additive manufacturing at GE, we developed a digital thread storage and analytics platform to capture, integrate, and extract value from the data being generated and used during the additive manufacturing lifecycle [22]. Using SADL and the SemTK framework, we constructed a knowledge graph of the additive domain to logically link diverse datasets such as material properties, build process parameters, and inspection results so that users could get a complete picture of the data and navigate their connections using familiar additive manufacturing domain terminology. For each instance of data, our knowledge graph also captures metadata describing the type of the data, where that data physically resides and codifies the queries or instructions to be used to access and retrieve that data on-demand. This way, users could get a single, unified logical view of all the data, and seamlessly navigate and interactively find data of interest from our additive manufacturing knowledge graph even when the data are diffuse, scattered among multiple storage systems. The platform leverages data organization-related metadata to automatically retrieve each data instance from its respective storage system and fuses the retrieved data based on the semantic model.

One of the first applications that this platform has been successfully used for is process parameter optimization, to accelerate the introduction of new materials into additive manufacturing. Multiple teams within GE use the platform to store, visualize, and analyze material, process parameter, and test coupon inspection data to understand what combinations of process parameters for different materials lead to the highest quality parts in the least amount of time. Through the platform's user interface (UI), users can visually explore and interact with the data, load additional data into the platform, and run multi-objective queries. On the main landing page of the UI, the user may select one of many "Predefined Queries" (i.e., NodeGroups). The user can then select one of the nine visualization types, including contour plots, 3D surface plots, box plots, and histogram plots. Examples of these four visualizations are shown in Fig. 5. Users visually analyze a combination of such plots to identify the best process parameter combinations for a new material.

Overall, the three-tiered material repository architecture—a federated data storage layer, a knowledge graph data abstraction layer, and a FAIR data access services layer—has allowed users to visually explore the materials data in a highly interactive manner through the application's user interface without requiring any knowledge of the underlying data architecture or data storage layer, greatly simplifying the management of additive manufacturing data, and allowing the materials scientists and additive engineers to worry less about data management and more about their primary concern—process parameter optimization to introduce new materials to the additive manufacturing industry more quickly and efficiently. The data access services layer dramatically simplifies not just the interaction with the knowledge graph for the end users, but also for the application builders for, e.g., developing the sophisticated user interfaces that enable a wide range of visualizations, analytics, and multi-objective data queries that are enabled within the system [22].



(a) Example contour plot

(b) Example 3D surface plot

(c) Example box plot

(d) Example histogram plot

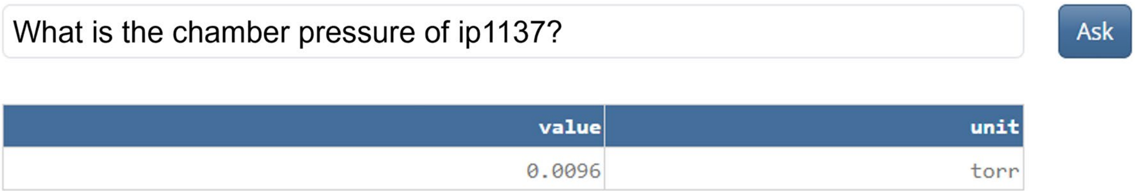
Fig. 5 Example visualizations available in the platform user interface, including **a** contour plots, **b** 3D surface plots, **c** box plots, and **d** histogram plots

AI for Accelerated Materials Development

High-pressure steam turbine blades are subject to aggressive oxidation and erosion during their operational lifecycle, prompting our service teams to seek enhancements over existing fielded coatings in use. The coatings are fabricated through 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 then subject to rigorous oxidation and erosion testing to mimic real-world service conditions for comprehensive evaluation. A materials science research program was initiated at GE to develop an enhanced coating using AI/ML approaches for predicting material properties from data about the coating's chemistry, data about the coating process such as voltages, currents, durations, and resulting coating thicknesses, and finally information about the oxidation and erosion evaluations.

While ML has made noteworthy strides as a tool for such predictions, and more broadly to advance materials science discoveries, there is an exclusive dependence on the availability of high-quality data as the primary driver for training ML models. However, most research studies looking to discover new materials or properties either (i) suffer from a scarcity of data needed to train high-precision models, or (ii) where adequate data points are present, may contain incomplete descriptors of the materials (i.e., lack a sufficient set of features to build meaningful models). Consequently, scientists are exploring other strategies that are not as dependent on data. This includes manually scavenging external sources (such as analytical equations developed and documented in the literature) and applying them in the context of a study to enhance the overall value of datasets.

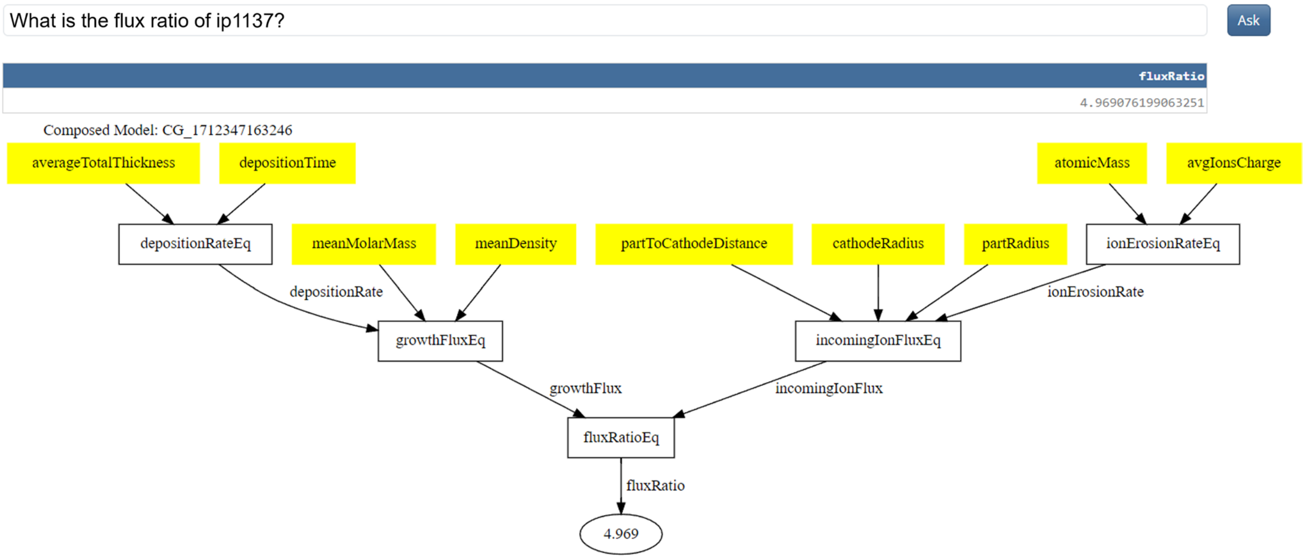
Despite their exceptional performance when operating under favorable circumstances, traditional ML approaches have significant limitations in their ability to extrapolate and gain insights in uncharted territories. Hence, inspired by DARPA's "3rd wave AI" vision [56], which calls for combining ML algorithms with domain knowledge to reason about areas never seen before, we developed a system that fuses multiple forms of knowledge into what we are calling a Compound Knowledge Graph (CKG). We combine three distinct, complementary forms of knowledge—factual, analytical, and human expert knowledge—into our CKG to enable contextual reasoning and adaptation to answer increasingly complex questions [23]. As of today, our CKG captures and links both factual scientific materials knowledge from materials science experiments as well as physics-based and data-driven ML models describing relationships between material processing, structure, properties, and performance in a knowledge graph. Within the CKG, analytical models are linked to the existing factual knowledge via the semantic description of the models' inputs and outputs. Thus, if a user requests the value of a property that is not explicitly



What is the chamber pressure of ip1137?

value	unit
0.0096	torr

Fig. 6 Question and answer interaction with the Compound Knowledge Graph in which the answer is directly in the knowledge graph and can be returned to the user. Both the numerical value and units are returned to the user



What is the flux ratio of ip1137?

fluxRatio
4.969076199063251

Composed Model: CG_1712347163246

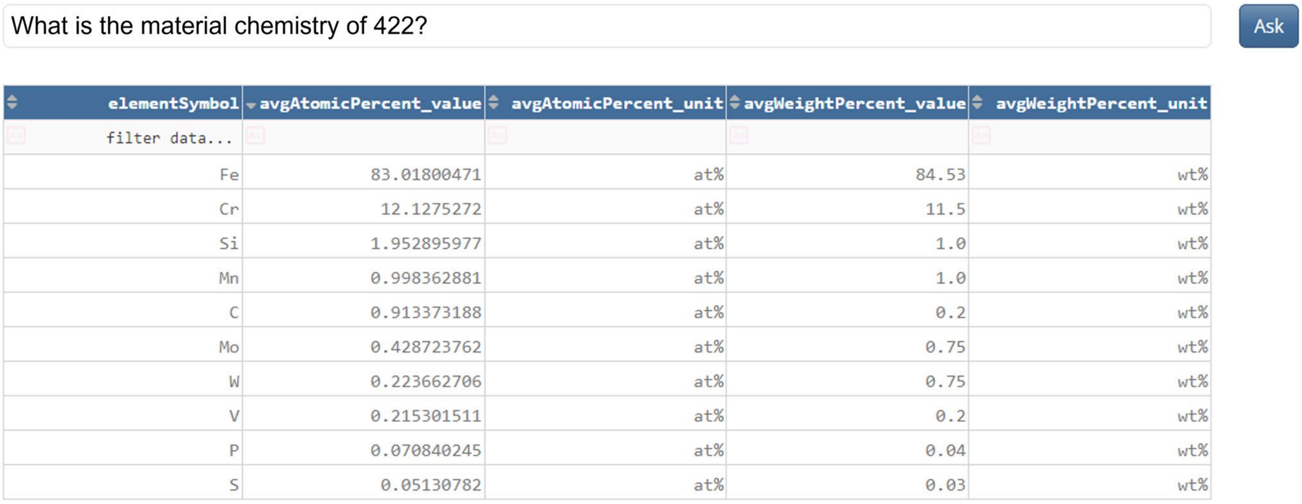
Fig. 7 Question and answer interaction with the Compound Knowledge Graph in which the answer is not available directly in the knowledge graph and so a dynamic analytical model chain must be autonomously generated to solve for the answer in real time. Once the answer is generated it is presented to the user along with a diagram showing the inputs in yellow, the analytical models being executed as white boxes, and final the output generated as an oval. In this example, five analytical models are executed from nine inputs to generate the output

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 or sequence of models and return the desired value without the user providing any guidance to the system regarding what model or models to use [38, 57].

For this particular study, we incorporated experimental data directly into the CKG as factual knowledge. This factual knowledge included descriptions of the chemistries of the substrates and coatings, data about the cathodic arc coating process such as voltages, currents, durations, and finally the test durations and remaining coating thicknesses from the oxidation and erosion testing that was performed on the samples. To enable advanced reasoning and inference, we further augmented the CKG with a collection of analytical models. In collaboration with domain experts, we identified 25 models, including nine physics-based equations from the literature and 16 models that call external services such as matminer [58]. As part of our FAIR data access services, our third abstraction—dynamic analytic chaining is used to enable on-demand composition and execution of relevant analytical models and factual knowledge in response to queries where the requested data did not previously exist in the CKG.

To facilitate interactions with the CKG, we implemented a question-answering interface that allows users to pose questions in natural language. Figure 6 shows an example of a question in which the answer is available directly in the knowledge graph, and so a simple retrieval from the CKG is all that is required and the resulting answer (both numerical value and units) are presented back to the user.

Figure 7 shows an example in which the user asks for the flux ratio of the deposition process of a part. This answer is not available directly in the CKG, and so, a dynamic analytic chain is required to autonomously derive the answer. Below the question, the system displays the answer and a diagram



What is the material chemistry of 422?

elementSymbol	avgAtomicPercent_value	avgAtomicPercent_unit	avgWeightPercent_value	avgWeightPercent_unit
filter data...				
Fe	83.01800471	at%	84.53	wt%
Cr	12.1275272	at%	11.5	wt%
Si	1.952895977	at%	1.0	wt%
Mn	0.998362881	at%	1.0	wt%
C	0.913373188	at%	0.2	wt%
Mo	0.428723762	at%	0.75	wt%
W	0.223662706	at%	0.75	wt%
V	0.215301511	at%	0.2	wt%
P	0.070840245	at%	0.04	wt%
S	0.05130782	at%	0.03	wt%

Fig. 8 Question and answer interaction with the Compound Knowledge Graph in which the answer is directly in the knowledge graph and can be returned to the user. In this instance, the answer is returned in the form of a table

of the analytical model chain that was dynamically assembled to compute the answer. The diagram shows that the system found that it could compute an answer by chaining five analytical models and using nine facts (highlighted in yellow) available in the CKG as inputs.

Finally, Fig. 8 shows a third example in which the user asks for information about the chemistry of a specific compound. In this scenario, the answer is not a single value and unit, but a table of elements, their average atomic percentages, and average weight percentages (and respective units).

In this success story, we used the identical architectural approach of a federated data store, a knowledge graph abstraction layer, and a set of service abstractions atop the knowledge graph to greatly simplify materials scientists' explorations and interactions with the data. The system uses reasoning over the semantics (metadata) of available factual data and analytical models to infer new knowledge to answer users' questions. The natural language dialog interface lets materials scientists pose questions in English and receive answers together with explainable diagrams when reasoning was performed to generate new knowledge.

Conclusions and Future Work

A data infrastructure is presented from the perspective of enabling material scientists to extract maximum value from all the potentially fragmented data at their disposal. The infrastructure is architected and implemented with the idea of applying FAIR principles to the management of materials data, and furthermore to make this data, once FAIRified, easily accessible and shareable using domain terminology. Specifically, we describe the layered architecture of our materials data repository that combines state-of-the-art practices in federated data storage and knowledge graph technology with our novel semantics-driven abstractions that simplify access to the FAIR data in the broader context of associated metadata and other domain knowledge. We applied this infrastructural approach to two materials and manufacturing-related use cases at GE, successfully enabling scientists and engineers to search for and access relevant data more efficiently and utilize the data to gain insights and build predictive ML models.

We are exploring several future directions to further expand our data infrastructure. (i) Beyond serving data geared primarily for analysis and visualization, the infrastructure should support scientific publication workflows as well—this requires expansion of our data storage layer to incorporate standard digital object identifiers (DoIs), search systems for text-based literature, and federation across common digital research data repositories. (ii) Beyond representing experimental data as factual knowledge and equations and models as analytical knowledge, our knowledge graph is being expanded to suitably capture and codify useful intuitive knowledge from domain experts. (iii) Given the successes of Large Language Models (LLMs) in enabling chat-based search interfaces, and emerging attempts to apply LLMs to classical data management problems, our federated data storage layer and knowledge graph could be suitably hybridized to additionally handle representations of the materials data and contextual metadata in some embedding space, and (iv) beyond smart interfaces for just data access, we are looking to incorporate similar abstractions and interfaces to help materials scientists collect and curate materials data in our repository—this would necessitate AI-assisted quality review frameworks that can help validate and verify the data at ingestion time, as well as based on its links to other stored data.

Declarations

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

References

1. Ren F, Ward L, Williams T, Laws KJ, Wolverton C, Hattrick-Simpers J, Mehta A (2018) Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. Sci Adv 4(4):eaqq1566. https://doi.org/10.1126/sciadv.aaq1566
2. Suzuki A, Shen C, Chennimalai Kumar N (2019) Application of computational tools in alloy design. MRS Bull 44(4):247–251. https://doi.org/10.1557/mrs.2019.70
3. Wen C et al (2019) Machine learning assisted design of high entropy alloys with desired property. Acta Mater 170:109–117
4. Hu Q-M, Yang R (2022) The endless search for better alloys. Science 378(6615):26–27. https://doi.org/10.1126/science.ade5503
5. White House Office of Science and Technology Policy. Materials Genome Initiative for global competitiveness. (2011) https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf
6. Rumble JR Jr (2017) Accessing materials data: challenges and directions in the digital era. Integr Mater Manuf Innov 6(2):172–186. https://doi.org/10.1007/s40192-017-0095-2
7. Himanen L, Geurts A, Stuart Foster A, Rinke P (2019) Data-driven materials science: status, challenges, and perspectives. Adv Sci 6(21):1900808. https://doi.org/10.1002/advs.201900808
8. Ottomano F, De Felice G, Gusev VV, Sparks TD (2024) Not as simple as we thought: a rigorous examination of data aggregation in materials informatics. Digital Disc. https://doi.org/10.26434/chemrxiv-2023-r9n12
9. Jain A, Ong SP, 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(1):011002. https://doi.org/10.1063/1.4812323
10. Blaiszik B, Chard K, Pruyne J et al (2016) The materials data facility: data services to advance materials science research. JOM 68:2045–2052. https://doi.org/10.1007/s11837-016-2001-3
11. National Science and Technology Council, Subcommittee on the Materials Genome Initiative Committee on Technology. Materials Genome Initiative strategic plan. (2021) https://www.mgi.gov/sites/default/files/documents/MGI-2021-Strategic-Plan.pdf
12. Wilkinson MD, Dumontier M, IjJ A, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE (2016) The fair guiding principles for scientific data management and stewardship. Sci Data 3(1):1–9
13. Liao X, Niehues A, deVisser C, Huang J, Doornbos C, Ederveen THA, Kulkarni P, van der Velde KJ, Swertz MA, Brandt M, van Gool AJ, 't Hoen PAC (2023) FAIR Data Cube, a FAIR data infrastructure for integrated multi-omics data analysis. medRxiv 2023.04.23.23289000; doi: https://doi.org/10.1101/2023.04.23.23289000
14. Kinkade D, Shepherd A (2022) Geoscience data publication: practices and perspectives on enabling the FAIR guiding principles. Geosci Data J 9:177–186. https://doi.org/10.1002/gdj3.120
15. NASA Google Docs (2023) How to make NASA science data more FAIR. Draft report from NASA Open Source Science Data Repositories Workshop
16. Scheffler M, Aeschlimann M, Albrecht M et al (2022) FAIR data enabling new horizons for materials research. Nature 604:635–642. https://doi.org/10.1038/s41586-022-04501-x
17. Gong H, He J, Zhang X et al (2022) A repository for the publication and sharing of heterogeneous materials data. Sci Data 9:787. https://doi.org/10.1038/s41597-022-01897-z
18. Brough DB, Wheeler D, Kalidindi SR (2017) Materials knowledge systems in python - a data science framework for accelerated development of hierarchical materials. Integr Mater Manuf Innov 6(1):36–53. https://doi.org/10.1007/s40192-017-0089-0
19. Ghiringhelli LM, Baldauf C, Bereau T et al (2023) Shared metadata for data-centric materials science. Sci Data 10:626. https://doi.org/10.1038/s41597-023-02501-8
20. Kalidindi SR, Buzzy M, Boyce BL, Dingreville R (2022) Digital twins for materials. Front Mater 9:2045–2052
21. McHugh J, Cuddihy PE, Williams JW, Aggour KS, Kumar VS, Mulwad V (2017) Integrated access to big data polystores through a knowledge-driven framework. In: IEEE International Conference on Big Data. pp 1494–1503.
22. Aggour KS, Kumar VS, Cuddihy P, Williams JW, Gupta V, Dial L, Hanlon T, Gambone J, Vinciquerra J (2019) Federated multimodal big data storage & analytics platform for additive manufacturing. In: IEEE International Conference on Big Data. pp 1729–1738.
23. Aggour KS, Detor A, Gabaldon A, Mulwad V, Moitra A, Cuddihy P, Kumar VS (2022) Compound knowledge graph-enabled ai assistant for accelerated materials discovery. Integr Mater Manuf Innov 11:467–478
24. Kumar VS, Aggour KS, Cuddihy P, Williams JW (2020) A federated, multimodal digital thread platform for enabling digital twins. Nav Eng J 132(1):47–56
25. Berners-Lee T (2006) Linked data. Available from: http://www.w3.org/DesignIssues/LinkedData.html
26. National Science and Technology Council, Committee on Technology Subcommittee on the Materials Genome Initiative (2014) Materials Genome Initiative strategic plan. https://mgi.nist.gov/sites/default/files/factsheet/mgi_strategic_plan_-_dec_2014.pdf
27. FAIR Guiding Principles: https://www.go-fair.org/fair-principles/
28. Ashino T (2010) Materials ontology: an infrastructure for exchanging materials information and knowledge. Data Sci J 9:54–61
29. Elementary Multiperspective Material Ontology (EMMO) (2020) Funded by the European Union Horizon 2020 Research and Innovation Programme. https://github.com/emmo-repo/EMMO.
30. Li H, Armiento R, Lambrix P (2020) An Ontology for the Materials Design Domain. The Semantic Web. 212–227
31. 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
32. Xiao G, Calvanese D, Kontchakov R, Lembo D, Poggi A, Rosati R, Zakharyaschev M (2018) Ontology-based data access: a survey.

In: 27th International Joint Conference on Artificial Intelligence. AAAI Press. 5511–5519.

33. Crapo A, Moitra A (2013) Toward a unified english-like representation of semantic models, data, and graph patterns for subject matter experts. Intl J Semantic Comp 7(3):215–236
34. W3C Semantic Web, https://www.w3.org/standards/semanticweb/
35. 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).
36. Williams JW, Cuddihy P, McHugh J, Aggour KS, Menon A, Gustafson S, Healy T (2015) Semantics for big data access & integration: improving industrial equipment design through increased data usability. In: IEEE International Conference on Big Data. 1103–1112.
37. Kumar VS, Cuddihy P, Aggour KS (2019) NodeGroup: a knowledge-driven data management abstraction for industrial machine learning. In: Proc. of the 3rd International Workshop on Data Management for End-to-End Machine Learning 10:4.
38. Gabaldon A, Chennimalai Kumar N (2019) Knowledge-driven model assembly and execution. Modeling the World's Systems Conf.
39. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K, Romacker M, Walgemoed P, Jimenez RC, Winnenburg R, Plasterer T, Gupta V, Hedley V (2019) Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today 24(4):933–938
40. Kemmer I, Keppler A, Serrano-Solano B et al (2023) Building A FAIR image data ecosystem for microscopy communities. Histochem Cell Biol 160:199–209. https://doi.org/10.1007/s00418-023-02203-7
41. Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH et al (2022) Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. J Biomed Semant 13:12. https://doi.org/10.1186/s13326-022-00263-7
42. Richard DR et al (2023) Umbrella data management plans to integrate FAIR data: lessons from the ISIDORe and BY-COVID consortia for pandemic preparedness. Data Sci J 22(1):35. https://doi.org/10.5334/dsj-2023-035
43. Greene G (2019) Building Open Access to Research (OAR) data infrastructure at NIST, CODATA Data Science Journal.
44. Kalidindi SR, Khosravani A, Yucel B et al (2019) Data infrastructure elements in support of accelerated materials innovation: ELA, PyMKS, and MATIN. Integr Mater Manuf Innov 8:441–454. https://doi.org/10.1007/s40192-019-00156-1
45. The Minerals, Metals & Materials Society (2017) Building a materials data infrastructure: opening new pathways to discovery and innovation in science and engineering. Pittsburgh, PA
46. Brandt N, Griem L, Herrmann C, Schoof E, Tosato G, Zhao Y, Zschumme P, Selzer M (2021) Kadi4Mat: a research data infrastructure for materials science. Data Sci J 20:1–8. https://doi.org/10.5334/dsj-2021-008
47. Pelkie BG, Pozzo LD (2023) The laboratory of Babel: highlighting community needs for integrated materials data management. Digital Discovery 2:544–556
48. He B, Gong Z, Avdeev M, Shi S (2023) FFMDFPA: a FAIRification framework for materials data with no-code flexible semi-structured parser and application programming interfaces. J Chem Inf Model 63(16):4986–4994. https://doi.org/10.1021/acs.jcim.3c00836
49. Stonebraker M, Cetintemel U (2005) "One size fits all": an idea whose time has come and gone. In: 21st International Conference on Data Engineering. 2–11, doi:https://doi.org/10.1109/ICDE.2005.1.
50. Zaveri A, Ertaylan G (2017) Linked data for life sciences. Algorithms 10:126. https://doi.org/10.3390/a10040126
51. Vogt L, Auer S, Bartolomaeus T, Buttigieg P, Grobe P, Michalik P, Stocker M, Usbeck R (2019) FAIR.ReD: semantic knowledge graph infrastructure for the life sciences. Biodivers Inf Sci Stand 3:37206. https://doi.org/10.3897/biss.3.37206
52. Wang M, Ma H, Daundkar A, Guan S, Bian Y, Sehirlioglu A, Wu Y (2022) CRUX: crowdsourced materials science resource and workflow exploration. In: 31st ACM International Conference on Information & Knowledge Management. ACM. 5014–5018. https://doi.org/10.1145/3511808.3557194
53. Gu Z, Lanti D, Mosca A, Xiao G, Xiong J, Calvanese D (2023) Ontology-based Data Federation. In: 11th International Joint Conference on Knowledge Graphs. ACM. 10–19. https://doi.org/10.1145/3579051.3579070
54. Frazier WE (2014) Metal Additive Manufacturing: A Review. J Mater Eng Perform 23(6):1917–1928
55. Tofail SAM et al (2018) Additive manufacturing: scientific and technological challenges, market uptake and opportunities. Mater Today 21(1):22–37
56. Launchbury J (2017) A DARPA Perspective on Artificial Intelligence. Information Innovation Office (I2O) DARPA. https://www.darpa.mil/attachments/AIFull.pdf.
57. 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(2):464–480
58. 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