npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

An infrastructure with user-centered presentation data model for integrated management of materials data and services

Shilong Liu[1,2,3], Yanjing Su[1,4], Haiqing Yin[1,5], Dawei Zhang[6], Jie He[1,2,3], Haiyou Huang[4,7], Xue Jiang[1,5,8], Xuan Wang[1,2,3], Haiyan Gong[1,2,3], Zhuang Li[1,2,3], Hao Xiu[1,2,3], Jiawang Wan[1,2,3] and Xiaotong Zhang[1,2,3]✉

With scientific research in materials science becoming more data intensive and collaborative after the announcement of the Materials Genome Initiative, the need for modern data infrastructures that facilitate the sharing of materials data and analysis tools is compelling in the materials community. In this paper, we describe the challenges of developing such infrastructure and introduce an emerging architecture with high usability. We call this architecture the Materials Genome Engineering Databases (MGED). MGED provides cloud-hosted services with features to simplify the process of collecting datasets from diverse data providers, unify data representation forms with user-centered presentation data model, and accelerate data discovery with advanced search capabilities. MGED also provides a standard service management framework to enable finding and sharing of tools for analyzing and processing data. We describe MGED's design, current status, and how MGED supports integrated management of shared data and services.

npj Computational Materials (2021)7:88 ; https://doi.org/10.1038/s41524-021-00557-x

INTRODUCTION

Increasingly, the materials community is acknowledging that the availability of vast data resources carries the potential to answer questions previously out of reach. Yet the lack of data infrastructure for preserving and sharing data has been a problem for decades. The need for such infrastructure was identified as early as the 1980s[1,2]. Since then, a large effort has been made in the materials community to establish materials databases and data repositories[3–6]. Some studies focus on specific subfield within materials science and develop corresponding databases. The Inorganic Crystal Structure Database is a comprehensive collection of crystal structure data of inorganic compounds containing more than 180,000 entries and covering the literature from 1913[7], where all crystal structure information is uniformly represented in the well-established Crystallographic Information File[8]. The Materials Project of first-principles computation provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials[9–11]. The Royal Society of Chemistry's ChemSpider is a free chemical structure database providing fast text and structure search access to over 67 million structures from hundreds of data sources[12]. National Environmental Corrosion Platform of China focuses on corrosion data and includes five major databases and 13 topical corrosion databases, which contains over 18 million data of 600 materials[13]. National Materials Scientific Data Sharing Network of China provides access to massive materials data resources that are collected from more than 30 research institutes across China[14]. Such infrastructures effectively solve problems for the target fields but are not general enough to meet the needs of the broad materials community. Moreover, the isolated management of materials data leads to information silos and impedes the process of data discovery and reuse.

As materials discovery becomes more data intensive and collaborative, reliance on shared digital data in scientific research is becoming more commonplace[15–17]. Several reports on Integrated Computational Materials Engineering have continued to highlight such need[18–20]. In 2011, the United States announced the Materials Genome Initiative to encourage communities to develop infrastructure to halve the time and cost from materials discovery to application[21]. In addition to US efforts, there are other international projects, such as the Metallurgy Europe program[22] in Europe, the "Materials research by Information Integration" Initiative (MI^2I)[23] in Japan and the Materials Genome Engineering (MGE) program in China that have been launched to develop such infrastructure.

China's MGE program originated at the S14 Xiangshan Science Forum on System Engineering in Materials Science in December 2011. The Chinese materials community reached the consensus on collaborative development of shared platforms integrated with theoretical computing, databases and materials testing at the forum. Following this forum was a series of conferences held across the nation from 2012 to 2014 for detailed strategic planning[24,25]. In 2016, the MGE program was launched by the Chinese government to change the concept of materials research to the new model of theoretical prediction and experimental verification from the traditional model of experience-guided experiment. It encourages researchers to integrate technologies for high-throughput computing (HTC), high-throughput experimenting (HTE), and specialized databases, and to develop a centralized, intelligent data mining infrastructure to speed up materials discovery and innovation[26].

The launch of the MGE program presents new challenges for modern data infrastructure. One is how to store and manage the ever-increasing amount of materials data with complex data types and structures. Better management of data benefits easier discovery and retrieval of datasets, better reproducibility, and

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing, China. [2]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. [3]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, China. [4]Corrosion and Protection Center, University of Science and Technology Beijing, Beijing, China. [5]Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing, China. [6]National Materials Corrosion and Protection Data Center, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, China. [7]Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, China. [8]Beijing Key Laboratory of Materials Genome Engineering, University of Science and Technology Beijing, Beijing, China. ✉email: zxt@ies.ustb.edu.cn

reuse of the study results[27–32]. Another is how to integrate data analysis and data mining techniques to unlock the great potential of such materials data. As materials research is evolving to the fourth paradigm of scientific research as Materials 4.0[33,34], great improvements in materials discovery have been achieved in applying machine learning techniques and big data methods to materials data[35–40]. Better integration between data and tools makes it convenient to discover the value of data.

To address these challenges, some studies take the approach to stand up very general materials data repositories that store as much data as possible without imposing strict restrictions on the structure or format. The Materials Commons provides open access to a broad range of materials data of experimental and simulation information, and allows collaboration through scientific workflows[41]. The Materials Data Facility provides data infrastructure resources and scalable shared data services to facilitate data publication and discovery[42]. The Materials Data Repository of National Institute of Standards and Technology provides a concrete mechanism for the interchange and reuse of research data on materials systems, which accepts data in any format[43]. These infrastructures provide a convenient means to preserve a wide variety of data, but do not enable straightforward searching and retrieving for data contents, or integrating with analysis tools due to the extreme heterogeneity in the stored data.

Some recent studies have recognized the importance of data standards. The Materials Data Curation System uses data and metadata models expressed as Extensible Markup Language (XML) Schema composed by researchers to dynamically generate data entry forms[44]. The Citrination platform has developed a hierarchical data structure called the Physical Information File that can accommodate complex materials data, ensuring that they are human searchable and machine readable for data mining[45]. These infrastructures provide a standardized data format to reduce the heterogeneity in the stored data, but enables only technical experts to manipulate these data formats due to the introduction of complex data types and structures.

After considering these previous efforts, we believe that the development of modern data infrastructure for MGE will hinge on two main technical requirements corresponding to integrated management of shared data and services:

(1) The infrastructure needs to provide a user-centered presentation data model[46] for materials researchers to collect and normalize heterogeneous materials data from various data providers easily and efficiently.
(2) The infrastructure needs to provide a service management framework of capabilities to integrate with various services and tools for analyzing and processing data, and cooperate with databases seamlessly, which enables service discovery and data reuse.

With these requirements in mind, we have developed the Materials Genome Engineering Databases (MGED), which is an emerging architecture with high usability. Serving as the materials data and service platform for MGE, MGED are differentiated from these previous efforts by our emerging architecture with high usability and user-centered dynamic container model (DCM). The architecture of MGED consists of four main systems: data collecting system (DCS), data exchange system (DES), hybrid data storage system (HDSS), and data service system (DSS). DCS collects and normalizes original datasets into standard container format constructed by DCM. DCM is a presentation data model that reflects the characteristics of materials data and allows effective user interaction with the database. DCM provides data schemas to represent the organization of experimental and computational data with associated metadata and data containers for the content of these data. Schemas are made from combinations of standard data types that encode data values and structures. These data types are well designed to have only ten kinds after consideration of suitableness for materials data and convenience for user interaction. Together with automation tools provided by DCS, MGED provide a low-overhead and convenient means to deposit heterogeneous materials data from various data providers. DES manages data mapping rules used for data parsing and reconstruction, and format transformation rules used for data exchange, storage, and service. HDSS is responsible for the management of storage technologies used in MGED and stores the data into corresponding databases according to its structural characteristics. DSS provides a fundamental services framework for search and discovery of data and a service integration framework for the management of various third-party analysis tools. DSS allows data and tools to be joined into a seamless integrated workflow to make data reuse and analysis more effectively. With the integrated management of shared data and tools, MGED provides researchers with an open and collaborative environment for quickly and conveniently preserving and analyzing data.

RESULTS

Architecture

MGED adopts a browser-server architecture using Python-based Django framework that allows users to easily access materials data and services on the platform through a browser. MGED is currently online and accessible at www.mgedata.cn and mged.nmdms.ustb.edu.cn.

Figure 1 provides a high-level view of MGED's architecture. MGED has four main systems: DCS, DES, HDSS, and DSS. DCS collects and normalizes original datasets from multiple data providers into standard container format. DES manages data mapping rules used for data parse and reconstruction, and handles data transformation among formats used for exchange, storage, and service. HDSS combines several storage technologies and stores each category of data occurred on the platform separately in the corresponding database. DSS manages a variety of data services including fundamental services like data search and third-party enhanced services like data mining.

Figure 1 also indicates two main materials data flows: the data colleting flow from data providers and data using flow from data consumers.

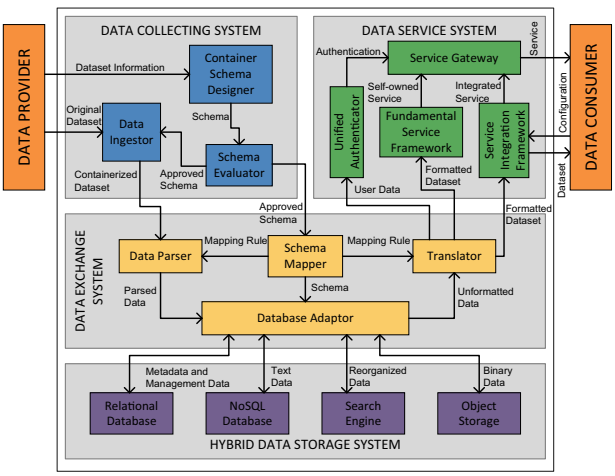

Fig. 1 Overall architecture of MGED. The architecture of MGED consists of four main systems: DCS, DES, HDSS, and DSS. DCS is responsible for collecting original datasets from multiple data providers and normalizing them into standard container format. DES manages data mapping rules and performs data transformation among different formats. HDSS stores each category of data into the corresponding database. DSS provides various data services through the fundamental services framework and the service integration framework.

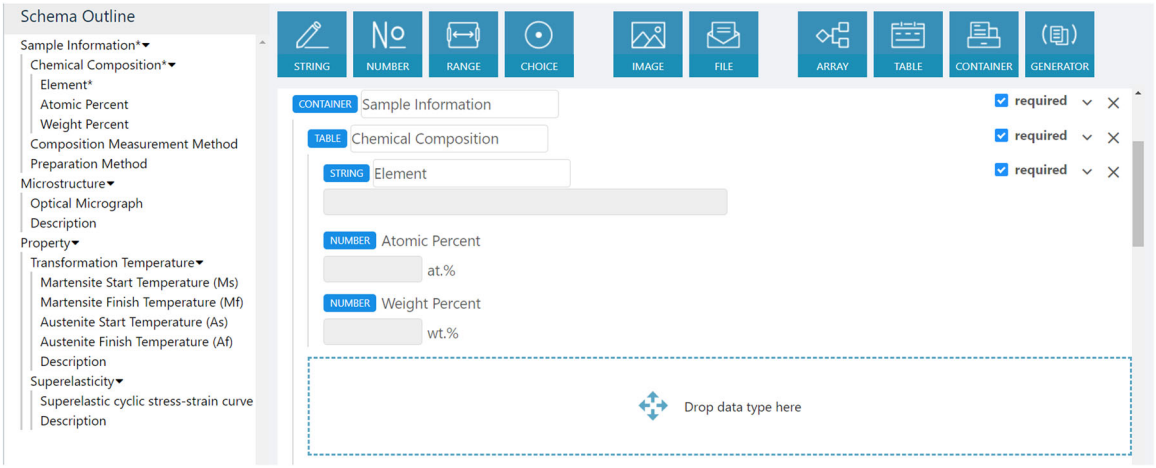

Fig. 2 The graphical user interface of the container schema designer. The designer allows users drag icons of data types and drop them to the dotted box to construct their schemas. The schema structure of the example data of shape memory alloys is shown in the Schema Outline.

Data providers makes data available to themself or to others. Data providers contains end users, researchers, tools, and data platforms that provides diversified data. In the data colleting flow, a data provider customizes schemas to represent the exchange structure of original datasets through the container schema designer. Schemas then are evaluated by the schema evaluator. When approved they will be stored into databases in HDSS through the schema mapper and the database adaptor. Approved schemas will be used by the data ingestor to normalize and transform original datasets from the data provider to containerized datasets, which is the standard data format used in MGED. After normalization, containerized datasets will be parsed as components like metadata, textual materials data and binary files, and these components will be stored separately to appropriate databases by the database adaptor.

Data consumers receive the value output of MGED. Data consumers contains end users, researchers, tools, and data platforms that analyze collected data. A data consumer interacts with MGED through the services provided by the service gateway to get access to the information of interest. The data consumer initiate query commands through the search service in the fundamental service framework to look up datasets. The database adaptor will retrieve the search result from different databases and send it to the translator. The translator performs transformation to present the result in a format that the service expects. The service integration framework receives the formatted result and transmits it to the data consumer for subsequent analysis. The analysis result can be stored to some data provider for later sharing. These two data flows constitute a virtuous circle of data sharing and service sharing.

The data collecting system

In this section we provide an overview of the DCS. We analyzed the factors that need to be considered in the materials data collecting process and proposed the DCM that meets these requirements. Current implementations of DCS are based on DCM and contains the following components: container schema designer, data ingestor, and schema evaluator.

DCM is a user-centered presentation data model that reflects the user's model of the data and allows effective user interaction with the database. Its name comes from the concept of containers in real life. A container in real life generally refers to a device for storing, packaging, and transporting a product. It usually has a fixed internal structure for fixing different types of products, such as a toolbox with corresponding shapes to fix hammers, scissors, and the like. In contrast, abstract containers in DCM are designed to have internal structure constructed dynamically from different types of basic structure. Therefore, DCM provides a way to store, wrap, and exchange data and enables users to customize schemas suitable for the structure of the data. DCM supports customization of attributes and structures. Users can arbitrarily choose attribute names without any restrictions in principle, but practically names in schemas that are publicly available on MGED should follow naming conventions of materials community. Attribute values can be restricted by data types. Structures can be customized by different combinations of data types.

As DCM plays a central role in MGED, its usability largely determines the usability of the platform. To this end, we developed the container schema designer to assist users in visually modifying existing schemas or creating entirely new schemas with built-in types, as illustrated in Fig. 2. We take the data of shape memory alloys as an example and shows how the attributes and structure of the data are described through the graphical user interface (GUI) of the container schema designer.

Container schemas are created by users with great flexibility. Various schemas can be created to describe the same field of materials, which will reduce the quality of data normalization and increase the difficulty for users to discover and use data. Therefore, we have developed the schema evaluator to evaluate the quality of schemas. New schemas in a certain field will be evaluated by experts in that field from the materials expert database. With deep understanding of both materials and schemas, evaluation experts can correct the inappropriate materials terminology and structure in the schema. Approved schemas will be published on the platform.

DCS also contains convenient tools provided by the data ingestor that allow for collecting datasets from data providers and normalizing them into containerized datasets automatically to reduce users' workload.

The data exchange system

In this section we outline the implementation architecture of DES. DES handles two exchange processes: the data persistence process that converts datasets from the container format to the database format, and the data retrieval process that converts datasets from the database format to the service format. DES consists of the following components: data parser, schema mapper, translator, and database adaptor.

In the data persistence process, the datasets are uploaded with information structured in a certain container format, such as XML, JavaScript Object Notation (JSON), or Excel. The container format is a data exchange format that facilitates the data collecting from

data providers to MGED and data sharing from MGED to services. The uploaded containerized datasets then will be converted to the corresponding database format in HDSS to achieve high efficiency for search and retrieval. The schema mapper generates mapping rules from container format to database format. The data parser handles containerized datasets from the data ingestor and breaks down them into parts suitable for different databases. The database adaptor then connects to all databases in HSS and stores each part into proper database.

In the data retrieval process, the database adaptor performs data retrieval operation from databases by generating database query statements satisfying users' search conditions. The retrieved container schema will be sent to the schema mapper for mapping rule generation. The retrieved unformatted data are transferred to the translator. The translator performs reconstruction to present the dataset in a format that the service expects.

The hybrid data storage system

HDSS is responsible for the management of storage technologies used in MGED. Data storage technologies are diverse and optimized for storing different categories of data. Figure 3 shows the categories of data stored in MGED from the perspective of storage. Management data include user account information, access privileges, and other miscellanies required for proper functionality. Scientific data include metadata like authors and digital object identifier (DOI) of the dataset, text data that structure information in text, and binary data that structure information in files. Therefore, HDSS has adopted multiple storage technologies to manage these data.

Figure 3 also shows the storage technologies used in HDSS and the corresponding category of data stored in them. A relational database is used to store metadata and management data that fit into the relational model. A NoSQL database is used to store heterogeneous text data that have no fixed schema. All binary data uploaded to MGED are persisted to an object storage. In addition, metadata and text data on the platform are reorganized and indexed in a search engine to enable complex queries. The current implementation of HDSS has adopted the well-known database systems as its backends. Specifically, PostgreSQL, MongoDB, MongoDB's GridFS, and Elasticsearch are used as the corresponding backends of the relational database, the NoSQL database, the object storage, and the search engine. We are also improving HDSS to support more database systems.

The data service system

In this section we outline the implementation architecture of DSS. DSS consists of the following components: service gateway, unified authenticator, fundamental service framework, and service integration framework.

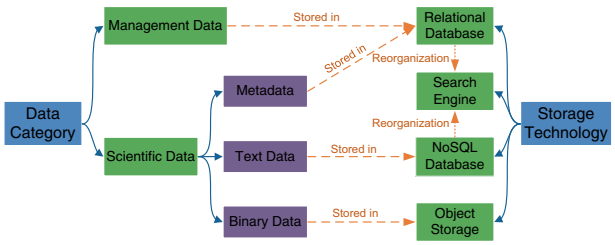

Fig. 3 An overview of data categories and storage technologies used in MGED and their relationships. The data stored in MGED are categorized as management data or scientific data. The scientific data are categorized as metadata, text data, or binary data. Each type of data is stored in the corresponding storage technologies.

The service gateway provides a unified portal for data consumers to access services. It verifies requests from data consumers and distributes them to corresponding services.

The unified authenticator handles user authentication and authorization privilege verification. MGED and third-party tools have user management functions of their own. Because these systems are independent of each other, users need to repeatedly log in to use each of them. The unified authenticator provides an open authorization service that allows secure API authorization in a simple and standard method from third-party services, making it easy for users to use services with the account of MGED.

The fundamental service framework provides services that promotes data discovery and sharing. It mainly includes search and export service, digital identification service (DIS), and classification and statistics service.

The search service provides three kinds of search functions to enable users to make complex queries. The basic search function allows users to quickly locate the required datasets through metadata information like data titles, abstracts, owners, and keywords. The advanced search function based on container schemas allows users to impose constraints on data attributes of interest and accurately access the required datasets; the full-text search function allows users to use multiple keywords to obtain datasets that contain these keywords whether in metadata or attributes. Each piece of data in the search result will be represented in visual interface generated from the corresponding schema. As shown in Fig. 4, the detailed information of the shape memory alloys is represented in the generated interface and the representation structure is the same as the structure described in the schema. Besides, the datasets in search results can be exported to JSON, XML, and EXCEL format for further research. The result datasets can also be exported with filters to select out only concerned attributes. We also provide data export APIs for integrated services.

Datasets are uniquely identified by the DIS with a DOI that contains information about owners and location of the underlying dataset. Association of a digital identifier facilitates discovery and citation of the dataset.

In addition, we also provide a classification and statistics service to help users quickly understand the status of materials data on the platform, as shown in Fig. 5. We divide materials science into different levels of fields and organize them into a category tree. Each piece of materials data on the platform is divided into a field in the tree. Statistics information of each field are shown in various visualization methods. Statistics information includes the total amount of data in MGED and separate amount of data in each field with their respective trends in data volume. Other information like the number of visits and downloads of each piece of data, popular fields, and rankings provides users detailed view to estimate hot data or fields. As of March 2021, there have been 7.3 million pieces of materials data in total collected through two website portals of MGED, including 20 major fields in the category tree. The top five fields with the most data are fields of special alloy, material thermodynamics/kinetics, catalytic materials, the first principle calculation, biomedical materials. We have also developed a simple data evaluation function that allows users to score each piece of data on the platform, which helps others judge data quality.

The service integration framework is responsible for integrating third-party computing and analysis tools for further research. Third-party online services can directly be integrated to MGED with an access portal in the service gateway and a dedicated API to transfer data. The offline service will be provided an introduction portal for users to download and use.

At present, the framework under development has integrated several services developed by cooperative teams in our project, such as MatCloud for HTC[47], OCPMDM for data mining[48], and the Interatomic Potentials Database for atomistic simulations. There have been some studies using data and services provided by

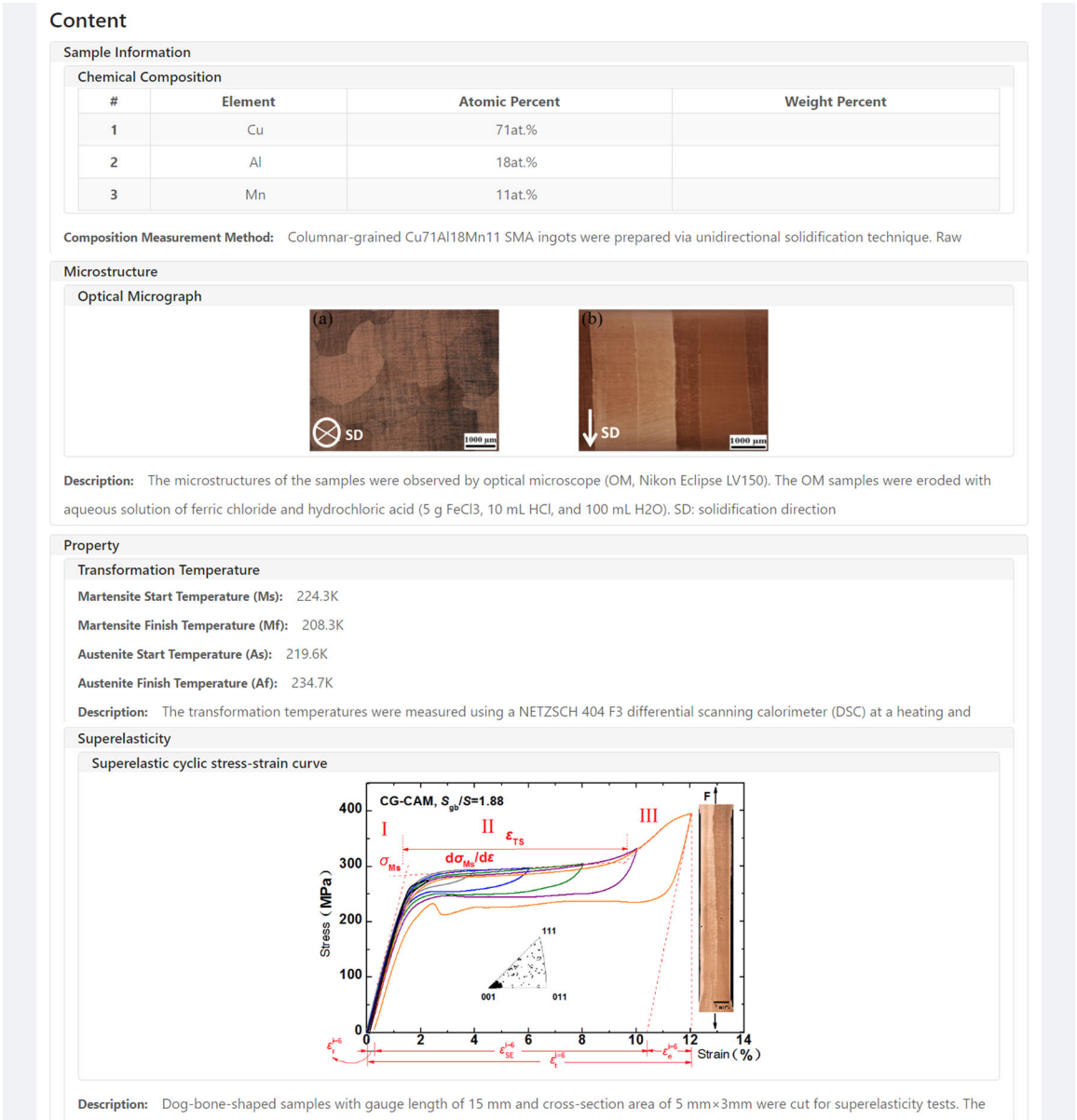

Fig. 4 The data representation interface dynamically generated from a container schema. The data representation interface is generated dynamically according to the schema of the example data of shape memory alloys.

MGED[49–51]. When the framework is fully developed and the integration process standard has been established, MGED will be open to all researchers in materials community and collaborates with them in development and integration of useful tools that improve data utilization, which promotes service sharing process and accelerates materials discovery.

DISCUSSION

To summarize, we have developed a modern materials data infrastructure, MGED, for scalable and robust integrated management of shared data and services for the materials science community. We have concluded from previous work that the development of modern infrastructure for MGE will hinge on two main technical requirements corresponding to integrated management of shared data and services: a user-centered presentation data model for easy and efficient data collecting and normalizing, and a service management framework capable of integrating various tools for analyzing and processing data. To address these requirements, we have developed our emerging architecture of MGED with high usability. In particular, we proposed a user-centered presentation data model, DCM, for materials researchers to get heterogeneous materials data into and out of MGED conveniently. DCM provides schemas to represent data and containers for the content of these data. Schemas consist of standard data types that describe data values and structures, which are designed not only to handle the heterogeneity of the data, but also to provide the convenience of user interaction. We also developed DSS to manage fundamental services for data search and discovery, and enhanced services provided by third-party for data analysis. DSS allows the stored data and integrated tools to be joined into a workflow to make data reuse and analysis more effectively. With the integrated management of shared data and tools, MGED provides researchers with an open and collaborative environment for quickly and conveniently preserving and analyzing data.

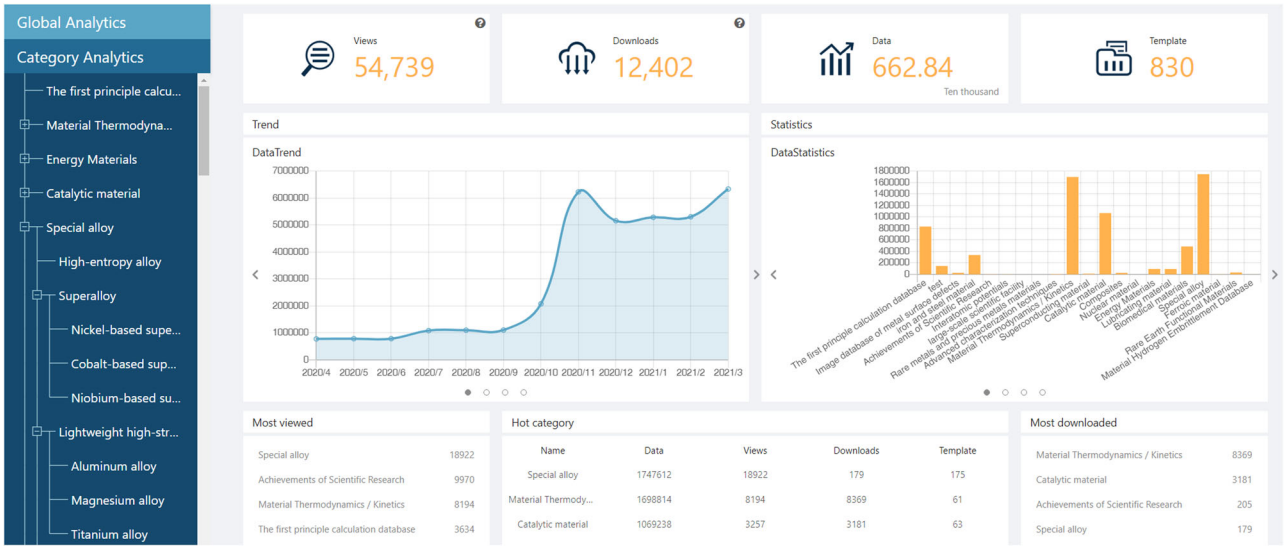

Fig. 5 The data statistics interface of MGED. The fields of materials science are divided into a hierarchy in the category tree shown in the left box. The statistics information of each field is shown in various visualization methods in different white boxes on the right.

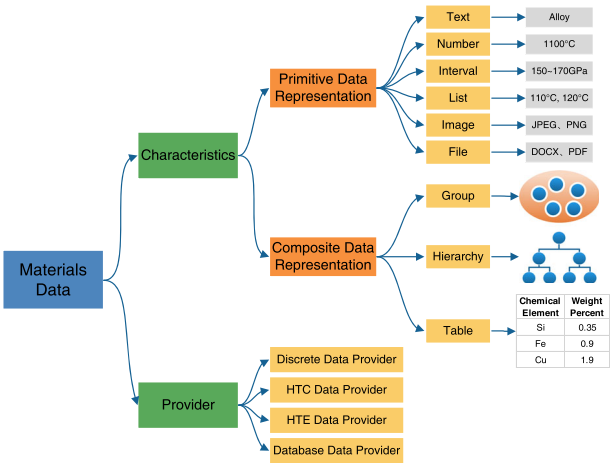

Fig. 6 Characteristics and provider classification of materials data. From the point of view of the characteristics of materials data, they can be characterized by primitive data representation forms and composite data representation forms. From another point of view of providers of materials data, they can be classified into four types based on the regularity of datasets stored in them.

METHODS

Characteristics and classification of materials data

Materials data providers are diverse and fragmented. Datasets they provide are typically heterogeneous and stored in different custom formats. We have developed DCS to communicate with data providers and collect datasets. In the design of DCS, decisions were made around technologies for the improvement of system usability. Data characteristics suitableness and operation convenience were the two mainly considered factors for usability.

Datasets collected from data providers need to be normalized into a common schema to enable accurate search and analysis. Meanwhile, the common schema should be suitable for materials data characteristics to reduce users' cognitive burden and learning cost. We have analyzed a large amount of materials data and summarized their characteristics, as shown in Fig. 6. Materials data are usually composed of a set of attributes with relationships in the abstract. Attributes are identified by their names. Values of attributes can be described in several different forms. These forms are called primitive data representations, such as a paragraph of text, a number, an interval, a list of numbers, or even files. The relationship between attributes is described by composite data representations, such as groups, hierarchy, or tables. Combination of attributes described by different data representations ultimately form a tree-like data structure. Then we developed the DCM to accommodate these characteristics.

At the same time, the data collecting process is time-consuming and laborious when researchers have to manually transfer original datasets to data infrastructure, which will reduce their motivation for sharing data. Therefore, DCS should contains convenient tools that allow for operation automation to reduce users' physical burden. We have classified data providers into four categories based on the regularity of datasets stored in them: discrete data providers, HTC data providers, HTE data providers, and database data providers. A discrete data provider is a materials researcher who organizes materials data with self-defined formats. HTC data providers refer to various materials computing software. HTE data providers refer to experimental equipment such as scientific apparatus. A database data provider is a database that has already stored large volumes of materials data. We provide dedicated data collecting tools for each category with appropriate granularity of operation.

The dynamic container model

DCM contains two main components: container schemas and container instances. A container schema represents the abstract description of attributes and structures of a materials dataset. The colon symbol : is used to represent the relation between the attribute and the type. The type declaration expression $x : T$ indicates that the type of the attribute x is T. A container schema $\mathcal{S}$ is a set of type declaration expressions and defined by $\mathcal{S} = \{x_i : T_i^{i\in 1..n}\} = \{x_1 : T_1, x_2 : T_2, \ldots, x_n : T_n\}$, where x_i is the attribute name and T_i the type name.

A container instance represents the abstract description of a piece of data in the dataset, which is constrained by the schema of the dataset and specifies the value of each attribute of the data. We represent the relation between the attribute and the value by the equal sign $=$. The assignment expression $x = v$ indicates that the attribute x has a value of v at a certain moment. A container instance $\mathcal{C}$ is a set of assignment expressions and defined by $\mathcal{C} = \{x_i = v_i^{i\in 1..n}\} = \{x_1 = v_1, x_2 = v_2, \ldots, x_n = v_n\}$.

A schema together with several instances constrained by it constructs a containerized dataset, which is a normalized description of a materials dataset. A containerized dataset is defined as a pair $(\mathcal{S}, \mathcal{D})$ where $\mathcal{D} = \{C_i^{i\in 1..n}\} = \{C_1, C_2, \ldots, C_n\}$.

The quantity and complexity of data types largely determine the functionality and usability of DCM. With analysis of materials data characteristics above, we define ten kinds of build-in types, including four primitive types, namely String, Number, Image, and File, and six composite types, that is Range, Choice, Array, Table, Container, and Generator.

Primitive types are basic components without internal structures. The type String represents a textual description. The type Number represents a numeric

Table 1. Examples of type declaration forms and attribute assignment forms.

Classification	Type declaration	Assignment
Primitive type		
String	x: String	x = "abc"
Number	x: Number	x = 1
Image	x: Image	x = a.png
File	x: File	x = b.pdf
Composite type		
Range	x: Range	x = (1,2)
Choice	x: Choice {"a", "b", "c"}	x = "a"
Array	x: Array {Number}	x = [1,2,3,4]
Generator	x: Generator { x_1: String, x_2: Number, x_3: Number }	x = {x_1 = "abc"}
Container	x: Container { x_1: String, x_2: Number, x_3: Number }	x = {x_1 = "abc", x_2 = 1, x_3 = 2}
Table	x: Table { x_1: String, x_2: Number, x_3: Number }	x = { {x_1 = "a", x_2 = 1, x_3 = 2}, {x_1 = "b", x_2 = 3, x_3 = 4}, {x_1 = "c", x_2 = 5, x_3 = 6} }

value. The type Image and File represent information in image formats and file formats separately. Considering the popularity of pictures in materials data and the requirement for subsequent image processing, we separate Image from File intentionally as an independent data type for high usability.

Composite types are constructed by combinations of built-in types. The type Range is composed of Number and represents an interval value of two numbers; the type Choice is composed of String and represents the text options that an attribute can take; the type Array is composed of an arbitrary built-in type *T* and indicates that an attribute should take an ordered list of values of *T*; the type Generator, Container, and Table consist of a collection of fields which are labeled build-in types. They differ in the form of values that an attribute can take. An attribute of Generator can only take one value of some field in the collection. An attribute of Container can take one set of values of all fields in the collection. An attribute of Table can take any number of sets of values of all fields in the collection. Table 1 shows some examples of the type declaration form of each build-in type and the corresponding attribute assignment form.

The data collecting tools

The data ingestor is responsible for collecting datasets from data providers and normalizing them into containerized datasets. It contains several dedicated data collecting tools to assist with data collecting process.

For discrete data providers, we provide GUIs and application programming interfaces (APIs) with high usability. The interfaces for data import are generated dynamically from user-designed schemas. The data ingestor also allows for automated curation of multiple datasets via user scripts through the representational state transfer API.

For HTC data providers and HTE data providers, we are developing tools for data extraction and transformation to help users automatically submit data. Datasets generated by calculation software are often in some standard formats. Therefore, extraction rules for them can be obtained easily according to the structure of formats. Datasets from HTC data providers are relatively diverse in formats. Therefore, only necessary metadata information can be extracted, and the original datasets will be submitted together with extracted metadata for future analysis.

For database data providers, we are developing a migration tool to automatically exchange data. Datasets in databases are generally stored in normal forms. Schemas for datasets and mapping rules can be extracted and built by user scripts. With schemas and mapping rules, the migration tool can automatically recognize original datasets, collect, and convert them into containerized datasets.

The digital identification service

In response to the materials community trend toward open data, MGED allows researchers to group data into datasets and publish through the DIS. When a dataset is published, it is given additional descriptive metadata and a digital identifier. The descriptive metadata includes information about titles, authors, and belonged projects to indicate data contribution and promote participation. The identifier is generated automatically by DIS and resolvable to the location of the underlying data and metadata. Association of a digital identifier enables convenient sharing and citation for research data.

In practice, information contained in a dataset may be incomplete or inaccurate, which will reduce confidence and trust in shared datasets. DIS provides internal reviews by experts from our materials expert database to ensure that the dataset being published passes specific quality control checks. DIS currently supports DOI and we are developing a multiple identification framework that integrate other identifiers and provide APIs to meet the diverse needs of users.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 9 September 2020; Accepted: 19 May 2021;
Published online: 08 June 2021

REFERENCES

1. Westbrook, J. H. & Rumble, J. R., Jr. *Computerized Materials Data Systems* (National Bureau of Standards, 1983).
2. Cahn, R. W. *The Coming of Materials Science* (Pergamon, 2001).
3. Kalidindi, S. R. Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. *Int. Mater. Rev.* **60**, 150–168 (2015).
4. Kalidindi, S. R. & De Graef, M. Materials data science: current status and future outlook. *Annu. Rev. Mater. Res.* **45**, 171–193 (2015).
5. Hill, J., Mannodi-Kanakkithodi, A., Ramprasad, R. & Meredig, B. Materials data infrastructure and materials informatics. in *Computational Materials System Design* (eds. Shin, D. & Saal, J.) 193–225 (Springer, 2018).
6. Warren, J. A. & Ward, C. H. Evolution of a materials data infrastructure. *JOM* **70**, 1652–1658 (2018).
7. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. Sect. B* **58**, 364–369 (2002).
8. Hall, S. R., Allen, F. H. & Brown, I. D. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr. Sect. A* **47**, 655–685 (1991).
9. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
10. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
11. Ong, S. P. et al. The materials application programming interface (API): a simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015).
12. Pence, H. E. & Williams, A. ChemSpider: an online chemical information resource. *J. Chem. Educ.* **87**, 1123–1124 (2010).
13. Li, X. Practice analysis about the sharing service of national materials environmental corrosion platform. *China Sci. Technol. Resour. Rev.* **50**, 101–107 (2018).
14. Yin, H., Jaing, X., Zhang, R., Liu, G. & Qu, X. National materials scientific data sharing network and its application to innovative developmentof materials industries. *China Sci. Technol. Resour. Rev.* **48**, 58–65 (2016).
15. Ward, C. H., Warren, J. A. & Hanisch, R. J. Making materials science and engineering data more valuable research products. *Integr. Mater. Manuf. Innov.* **3**, 292–308 (2014).
16. Li, X. et al. Share corrosion data. *Nature* **527**, 441–442 (2015).
17. Jain, A., Persson, K. A. & Ceder, G. Research update: the materials genome initiative: data sharing and the impact of collaborative ab initio databases. *APL Mater.* **4**, 053102 (2016).
18. National Research Council. *Materials Research to Meet 21st-Century Defense Needs* (National Academies Press, 2003).

19. National Research Council. *Accelerating Technology Transition: Bridging the Valley of Death for Materials and Processes in Defense Systems* (National Academies Press, 2004).
20. National Research Council. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security* (National Academies Press, 2008).
21. Holdren, J. P. *Materials Genome Initiative for Global Competitiveness* (National Science and Technology Council, 2011).
22. Jarvis, D. et al. *Metallurgy Europe—A Renaissance Programme for 2012-2022* (European Science Foundation, 2012).
23. Japan Science and Technology Agency. "Materials research by Information Integration" Initiative. http://www.nims.go.jp/MII-I/en/ (2015).
24. Wang, H., Xiang, Y., Xiang, X. & Chen, L. Materials genome enables research and development revolution. *Sci. Technol. Rev.* **33**, 13–19 (2015).
25. Yin, H., Qu, X. & Xie, J. Analysis of the implementation and development of the Material Genome Initiative in Beijing. *Adv. Mater. Ind.* **1**, 27–29 (2014).
26. O'Meara, S. Materials science is helping to transform China into a high-tech economy. *Nature* **567**, S1–S5 (2019).
27. de Pablo, J. J., Jones, B., Kovacs, C. L., Ozolins, V. & Ramirez, A. P. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci.* **18**, 99–117 (2014).
28. Olson, G. B. & Kuehmann, C. J. Materials genomics: from CALPHAD to flight. *Scr. Mater.* **70**, 25–30 (2014).
29. Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. A bridge for accelerating materials by design. *npj Comput. Mater.* **1**, 15008 (2015).
30. Austin, T. Towards a digital infrastructure for engineering materials data. *Mater. Disco.* **3**, 1–12 (2016).
31. Pfeif, E. A. & Kroenlein, K. Perspective: data infrastructure for high throughput materials discovery. *APL Mater.* **4**, 053203 (2016).
32. The Minerals Metals & Materials Society (TMS). *Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering* (TMS, 2017).
33. Hey, T., Tansley, S. & Tolle, K. *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009).
34. Jose, R. & Ramakrishna, S. Materials 4.0: materials big data enabled materials discovery. *Appl. Mater. Today* **10**, 127–132 (2018).
35. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
36. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine learning in materials science: recent progress and emerging applications. *Rev. Comput. Chem.* **29**, 186–273 (2016).
37. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
38. Liu, Y. et al. Materials discovery and design using machine learning. *J. Mater.* **3**, 159–177 (2017).
39. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
40. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).
41. Puchala, B. et al. The materials commons: a collaboration platform and information repository for the global materials community. *JOM* **68**, 2035–2044 (2016).
42. Blaiszik, B. et al. The materials data facility: data services to advance materials science research. *JOM* **68**, 2045–2052 (2016).
43. Material Measurement Laboratory. NIST Materials Data Repository. https://materialsdata.nist.gov/ (2017).
44. Dima, A. et al. Informatics Infrastructure for the Materials Genome Initiative. *JOM* **68**, 2053–2064 (2016).
45. O'Mara, J., Meredig, B. & Michel, K. Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. *JOM* **68**, 2031–2034 (2016).
46. Jagadish, H. V. et al. Making database systems usable. in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. 13–24 (Association for Computing Machinery, 2007).
47. Yang, X. et al. MatCloud: a high-throughput computational infrastructure for integrated management of materials simulation, data and resources. *Comput. Mater. Sci.* **146**, 319–333 (2018).
48. Zhang, Q., Chang, D., Zhai, X. & Lu, W. OCPMDM: online computation platform for materials data mining. *Chemom. Intell. Lab. Syst.* **177**, 26–34 (2018).
49. Zhao, X. P., Huang, H. Y., Wen, C., Su, Y. J. & Qian, P. Accelerating the development of multi-component Cu-Al-based shape memory alloys with high elastocaloric property by machine learning. *Comput. Mater. Sci.* **176**, 109521 (2020).
50. Gao, X., Wang, L. & Yao, L. Porosity prediction of ceramic matrix composites based on random forest. *IOP Conf. Ser. Mater. Sci. Eng.* **768**, 052115 (2020).
51. Ma, B. et al. A fast algorithm for material image sequential stitching. *Comput. Mater. Sci.* **158**, 1–13 (2019).

ACKNOWLEDGEMENTS

The authors thank the Ministry of Science and Technology, the Ministry of Industry and Information Technology, and other relevant departments for their efforts in launching MGE program and their support for our projects and MGED. MGED is a collaboration among the University of Science and Technology Beijing (USTB), the Shanghai University (SHU), Tsinghua University (THU), Sichuan University (SCU), Southwest Jiaotong University (SWJTU), Beijing University of Technology (BJUT), Beijing Institute of Technology (BIT), the Central South University (CSU), Northwestern Polytechnical University (NPU), Shanghai Jiao Tong University (SJTU), the Computer Network Information Center (CNIC) of the Chinese Academy of Sciences (CAS), Ningbo Institute of Materials Technology and Engineering (NIMTE) of CAS, The Academy of Mathematics and Systems Science (AMSS) of CAS, the Central Iron and Steel Research Institute (CISRI), the Institute of Metal Research (IMR) of CAS, the Institute of Chemistry of CAS (ICCAS), the Ningbo Institute of Information Technology Application of CAS, the Institute of High Energy Physics (IHEP) of CAS, Beijing Computing Center, the National Supercomputer Center in Tianjin, Beijing Institute of Aeronautical Materials (BIAM) of the Aero Engine Corporation of China (AECC), and Tianjin Nanda General Data Technology Co., Ltd. This research was supported in part by the National Key Research and Development Program of China under Grant Nos. 2016YFB0700500 and 2018YFB0704300, the Fundamental Research Funds for the University of Science and Technology Beijing under Grant FRF-BD-19-012A, and the National Natural Science Foundation of China under Grant No. 61971031. We would like to thank the users of MGED for their support and feedback in improving the platform.

AUTHOR CONTRIBUTIONS

S.L., Y.S., H.Y., J.H., X.J., X.W., and X.Z. participated in the design and discussion of the infrastructure architecture. S.L., X.W., H.G., Z.L., H.X., and J.W. developed and implemented the architecture. S.L formulated the problem and participated in the development of the data model with J.H. and X.Z. H.H. prepared the data used for the demonstration. S.L. prepared the initial draft of the paper. All authors contributed to the discussions and revisions of the paper.

COMPETING INTERESTS

A patent application (201710975996.1) on the storage method of early version of MGED has been submitted by University of Science and Technology Beijing with X.Z., S.L., Y.S., H.Y., X.J., and X.W. as inventors.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to X.Z.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.