

## Research data express: a data accumulation and sharing system for digital transformation in materials research

Jun Fujima, Hideki Yoshikawa, Hiroko Nagao, Hiroaki Tosaka, Takuya Kadohira & Masahiko Demura

**To cite this article:** Jun Fujima, Hideki Yoshikawa, Hiroko Nagao, Hiroaki Tosaka, Takuya Kadohira & Masahiko Demura (2025) Research data express: a data accumulation and sharing system for digital transformation in materials research, Science and Technology of Advanced Materials: Methods, 5:1, 2597702, DOI: [10.1080/27660400.2025.2597702](https://doi.org/10.1080/27660400.2025.2597702)

**To link to this article:** <https://doi.org/10.1080/27660400.2025.2597702>



© 2025 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group



Published online: 15 Dec 2025.



Submit your article to this journal [↗](#)



Article views: 106



View related articles [↗](#)



View Crossmark data [↗](#)

# Research data express: a data accumulation and sharing system for digital transformation in materials research

Jun Fujima , Hideki Yoshikawa , Hiroko Nagao, Hiroaki Tosaka, Takuya Kadohira and Masahiko Demura 

Materials Data Platform, National Institute for Materials Science, Tsukuba, Japan

## ABSTRACT

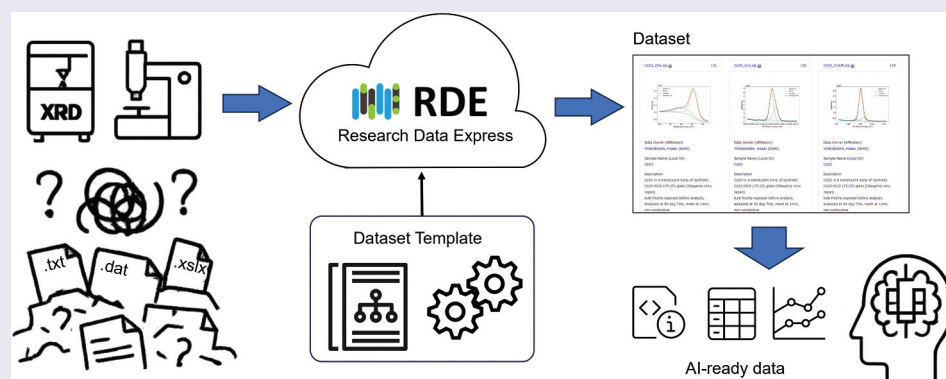
To accelerate the digital transformation (DX) in materials science, foundational technology for effectively managing vast amounts of data from diverse experiments and calculations and sharing them in a reusable format is essential. However, a significant barrier is that data formats, description styles, and terminologies differ among research fields, instrument manufacturers, and models, and are often not publicly available as readily usable schemas, making them unreadable. Research Data Express (RDE), introduced in this paper, is a highly flexible and scalable data accumulation and sharing system developed to solve these challenges. The core feature of RDE is the 'Dataset Template', which defines the format for data description and translation. Users can flexibly define data structures tailored to their research content using these templates. During data registration, raw files from experimental instruments and manually entered experimental conditions are automatically interpreted, integrated, and structured by a processing program defined in the template. This automated process includes metadata extraction, translation into common domain-specific terms, data visualization, and even the calculation of feature values for machine learning applications. In this paper, we detail the basic design and system architecture of RDE and explain the data management methodology based on Dataset Templates. RDE significantly reduces the burden of routine data processing for researchers and enhances data findability, interoperability, reusability (the FAIR principles), and traceability, thereby strongly promoting data-driven materials research.

## ARTICLE HISTORY

Received 2 October 2025  
Revised 11 November 2025  
Accepted 27 November 2025

## KEYWORDS

Materials informatics; data management system; digital transformation (DX); Dataset Template; workflow automation



## IMPACT STATEMENT

Research Data Express (RDE) uses 'dataset templates' to achieve both high flexibility for heterogeneous experimental data and the scalability needed to accelerate materials research through automated workflows.

## 1. Introduction

The digital transformation (DX) of materials research is an initiative to streamline and innovate materials research by leveraging data analysis, AI, simulation, and automation technologies. Within this context, there is a growing need for a common platform system to effectively manage the vast amounts of materials data generated through research activities and to provide standardized data access [1]. This data includes

experimental data such as material fabrication or synthesis process conditions, results from various measurements and characterizations, computational data from simulations, and their associated metadata. Properly recording and preserving this data in a reusable format is key to improving research efficiency and fostering new discoveries [2]. Furthermore, the development of a common data access infrastructure not only facilitates collaboration among different

**CONTACT** Jun Fujima  [fujima.jun@nims.go.jp](mailto:fujima.jun@nims.go.jp)  Materials Data Platform, National Institute for Materials Science, 1-1, Namiki, Tsukuba, Ibaraki 305-0044, Japan

© 2025 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

research groups but also enables knowledge sharing across various fields.

A major challenge in the DX of materials research is that data generated from diverse experimental instruments and simulations exist in different manufacturer-specific formats, often accompanied by inconsistent metadata [3]. This makes data aggregation, comparison, and reuse remarkably difficult. Traditionally, researchers have had to spend considerable time and effort on tasks such as data cleaning, format conversion, and documentation. This burden has been a factor in researchers' reluctance to share data, hindering the advancement of data-driven research.

Moreover, materials research covers a wide range of application fields, including energy materials, environmental materials, structural materials, and electronic materials. Since processes and analytical methods differ in each field, it is difficult to represent data descriptions in a unified format [4]. Even if a data description method is defined for one field, it needs to be revised whenever new materials or technologies emerge. Therefore, new data management must be able to quickly respond to the revisions.

Since 2017, the National Institute for Materials Science (NIMS) has been conducting interviews with researchers in various fields to study methods for recording research data at different granularities, from individual research themes to broader research domains [5–7]. As a result of trial and error, we have developed a system for recording research data called RDE (Research Data Express) [8], which is introduced in this paper. RDE allows you to register data files obtained from experiments and simulations as they are, automatically interprets them to generate structured files and stores all files in an organized state. This reduces the effort required to register data in a format suitable for DX, lowering the barrier for researchers to register data without missing information immediately after an experiment before forgetting

detailed conditions. Furthermore, this system does not unify the data description methods themselves, but rather standardizes the way data description methods are defined according to various types of data. As long as a data description method is defined according to this unified approach, it can be handled by the system, allowing for flexible recording of data for various materials. The definition of a data description method is called a Dataset Template in RDE, and data recorded using the same Dataset Template will have the same data structure. This means that by analyzing the Dataset Template, one can understand what kind of data is included without looking at the data directly. By using RDE, it becomes possible to digitize research data from its point of creation and record it in a way that enhances its reusability.

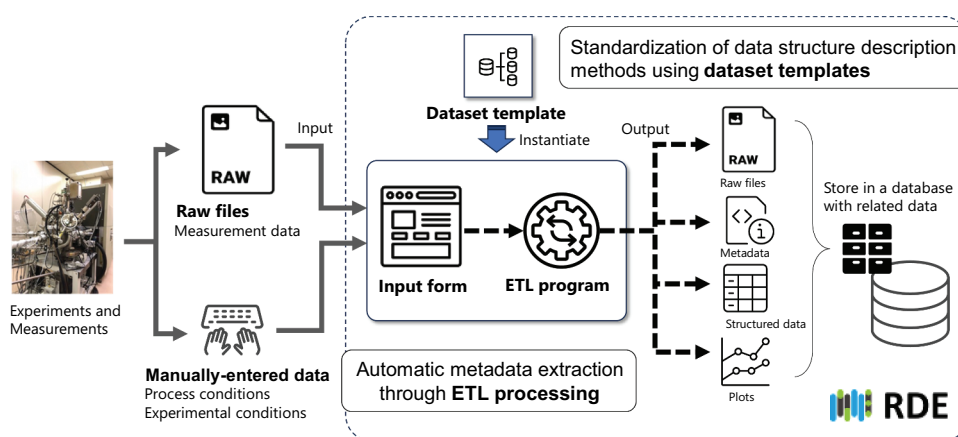
In this paper, we first explain the basic design of RDE as a system, and then detail the method for defining data description formats based on Dataset Templates, the method for structuring research data, and provide utilization examples. Furthermore, we compare our efforts with initiatives for data accumulation and sharing in other projects and systems.

## 2. Basic design of the RDE system

### 2.1. Core concepts of RDE

To comprehensively collect, structure, manage, and share the diverse data generated in materials research, we have developed the Research Data Express (RDE) system. Figure 1 shows the basic concept of RDE. RDE handles data generated in the research field by categorizing it into two types based on its nature: 'raw files' and 'manually entered data'.

The first data source consists of 'raw files' directly output from experimental instruments or simulation software. These files often contain not only numerical data of experimental results but also detailed metadata such as instrument model information, basic settings,



**Figure 1.** Data processing flow in RDE enabling immediate sharing, structuring, and reuse of experimental materials data.

and experimental conditions in their header sections. RDE emphasizes uploading these raw files to the system as is, and automatically extracting as much information as possible from the files as metadata. In actual research settings, sample information and some experimental conditions are often managed digitally by researchers as ‘electronic lab notebooks’ in user-defined formats using general-purpose software like Microsoft Excel. RDE treats these existing user-specific digital documents as a type of raw file, allowing the information described in electronic lab notebooks to be automatically extracted as metadata by uploading them to the system, similar to instrument output files.

The second data source is information that is not included in either raw files from instrument outputs or electronic lab notebooks. This includes information recorded on paper or information that is digitally recorded but not aggregated or standardized in format. Examples include sample synthesis processes described in a section of a report file or environmental conditions such as room temperature and humidity recorded on paper during a measurement. Since this second type of data source is often managed separately from the first, linking it with the first data source becomes difficult over time, significantly reducing the reusability of the data. To simplify the task for researchers of registering this second data source simultaneously with the first, RDE is equipped with a function to register it as ‘manually entered data’ through input forms customizable for each research theme. However, an increase in the number of manual input items can be a burden on the researcher and may hinder the smooth progress of experiments. Therefore, manual input items should be limited to what is truly necessary for later data analysis (usually no more than 10–15 items considering human concentration) and presented in a GUI that allows for intuitive input.

The raw files and manually entered data input into the system are automatically interpreted, integrated, and structured by a processing program defined in the ‘Dataset Template’. The purpose of this program is not only to organize and register miscellaneous information but also to add new value to the data through the ETL (Extract, Transform, and Load) processing, maximizing its FAIR principles of findability, interoperability, reusability, and traceability. The ETL processing involves multiple steps. A basic process is to automatically read metadata such as instrument model information, basic settings, and experimental conditions from the raw files and translate manufacturer-specific terms (including abbreviations) into common domain-specific expressions. This improves the findability and interoperability of the data. In addition to these basic processes, it is also possible to

incorporate processes to enhance readability. This includes converting the original data into a structured format with high machine-readability, and visualizing the data by generating graphs to help people intuitively grasp the data’s overview. This improves the reusability of the data. Furthermore, it is possible to automatically calculate feature values necessary for later machine learning from the original data such as spectra and images, and even perform advanced analyses like regression analysis using machine learning models. This series of processes can be executed automatically by the system by pre-defining the process details in the ‘Dataset Template’. This mechanism frees researchers from routine data processing and enables them to perform reproducible analyses efficiently. Finally, the series of information generated through this ETL processing, such as ‘raw files, metadata, structured data, visual data, and analysis results’, is systematically registered into the database in a standardized format. These processes are automatically executed according to definitions in the Dataset Template. As a result, researchers are relieved from routine data handling and can conduct reproducible analyses more efficiently. Ultimately, information including Raw Files, Metadata, Structured Data, Visualization Data, and Analysis Results are systematically stored in the database in a standardized format per template.

The RDE system is also flexibly designed to enable efficient operations, such as registering multiple files at once, depending on how the processing is defined in the Dataset Template. For example, it can be configured to automatically organize and register multiple raw files uploaded at once as individual data records. Similarly, when a tabular file like a Microsoft Excel sheet is provided, it can be configured to interpret each row as a separate piece of data and register them into the database in a batch. By defining such data processing in a template, it becomes possible to manage the correspondence between numerous synthesis conditions and measurement results or large volumes of data from continuous measurements, systematically with a single operation.

In materials research, the points of interest in material creation processes, material measurements, and material property evaluations vary greatly by research field, so the handling and description of data must be adapted for each study. To address this, RDE adopts a method of preparing multiple different Dataset Templates for various research themes, which can be switched as needed according to the user’s interests. This allows for appropriate data description even in different research areas and realizes flexible data management. Additionally, data described using the same Dataset Template can be handled consistently as

a unified format, which is expected to promote data sharing and reuse among different research groups.

## 2.2. Logical structure of RDE

Figure 2 shows the relationships among the basic components of RDE. Accumulated data is managed in units called ‘Datasets’, with a research team consisting of members with access rights acting as the owner. It is possible to set the scope of sharing with other research teams and a public release embargo period. In a dataset, multiple data files are stored in logical units called ‘Data’, which have a structure defined by a ‘Dataset Template’ according to a certain data description format. A single ‘Data’ represents a set of files related to one experiment or calculation. For example, a dataset for accumulating X-ray diffraction (XRD) measurement data would include not only the profile data of the measurement results but also metadata related to the measured sample, such as chemical composition, sample name, environmental conditions like temperature and pressure, and measurement date and time. Another dataset might contain information about SEM images and segmentation results of the sample’s internal structure obtained from image analysis. ‘Data’ can be considered the smallest unit of information that has practical value for materials researchers as a data registrant.

The Dataset Template defines necessary manual input items for data registration, metadata item definitions, and the processing programs for converting data into a common storage format. It is possible to create multiple datasets from the same template, thereby efficiently creating and managing multiple datasets with a unified data structure.

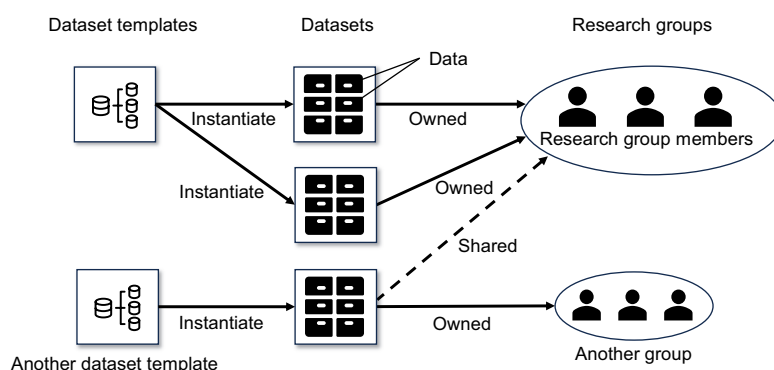
The use of RDE is conducted on a virtual ‘Research Group’ basis. A single research group can create and own multiple datasets, which are shared among the members of the group. Each member can be assigned a role such as ‘Administrator’ or ‘Data Registrant’, and the access rights to the owned datasets change according to the assigned role. Furthermore, individual

datasets can have sharing permissions set for other research groups, enabling dataset-level data sharing between different research groups.

RDE has a function to register and reference only ‘Sample’ and ‘Instrument’ metadata as master information in the system. It should be noted that the ‘Sample’ master is for reference within the same research team, while the ‘Instrument’ master is for reference by all RDE users, with different reference scopes. By using this master function, each data can be linked to a ‘Sample’ stored in the ‘Sample Master’ managed within the research team, making it possible to link different experimental results for the same sample. The ‘Instrument Master’ enables management of measurement and material synthesis instruments across research teams and allows for the management of instruments and Dataset Templates in an associated manner. This function makes it possible to identify which instrument’s data a Dataset Template is intended for, or to retrieve data obtained from the same instrument in a linked manner. It is also possible to register sample and instrument information as general metadata instead of as master information.

## 2.3. System architecture

The RDE system is built on the Microsoft Azure public cloud and is provided as a web application that users can access via a web browser. The RDE system employs a scalable architecture utilizing Microsoft Azure managed services to accommodate future increases in data volume and users. First, the computationally intensive ETL processing that occurs during data registration is packaged as independent containers and executed on a batch processing infrastructure on Azure. This infrastructure can dynamically allocate computational resources according to the number of registration requests and execute numerous processes in parallel, making it possible to maintain high throughput. Although individual ETL processes may require a certain amount of processing time, especially when complex analyses are involved (e.g. TEM image



**Figure 2.** Logical relationship among templates, datasets, and research groups in RDE.

segmentation described in Section 4), this processing is executed asynchronously in the background. Therefore, the design minimizes the impact on the system's overall data retrieval performance and responsiveness. Second, the web application layer, which handles interactive requests from users such as data search and browsing, is placed on an application hosting platform capable of load balancing and automatic scaling. This ensures that a responsive user experience can be provided to each user, even as the number of concurrent users increases. Furthermore, the generated data is permanently stored in highly scalable cloud object storage, which is not constrained by capacity limits.

As shown in Figure 3, the RDE system architecture consists of a group of backend components that manage data processing (center of the figure), a storage component that permanently stores data (bottom right), and a group of frontend applications that users directly interact with (top). There are two types of applications: one for general users and one for administrators.

User management in RDE utilizes the authentication function of DICE ID Management, an ID management infrastructure provided by NIMS. Within RDE, an Authorization service manages the permissions of authenticated users, controlling which research team a user belongs to and what access rights they have to which datasets.

The data registration process by a user is initiated through the Dataset registration app. The user fills in a form presented on the screen with basic information, sample information, and template-defined input items, and can upload multiple related raw files in a single operation. This action triggers the ETL (Extract, Transform, and Load) engine. The ETL engine reads the appropriate Dataset Template from the Dataset template database based on the registration content, and further retrieves the necessary data processing container from the Container registry.

Subsequently, the ETL processing of the data is performed using the retrieved processing container. The file group generated by the process is saved in the Dataset DB.

The stored data can be accessed across multiple applications, including the Dataset viewer app and the Sample management app, via the Data Access service. The Dataset viewer app allows for searching and viewing detailed information of datasets, while the Sample management app centrally manages sample information. These applications cooperate with the Dataset template database and the Dataset DB to support quick access to the data users need. The details of a selected dataset display basic dataset information, a data catalog, and a list of metadata items defined in the associated Dataset Template. Individual data within a dataset can be viewed visually in a tile format (Figure 4(a)) or a classification tree format based on metadata (Figure 4(b)). This classification tree display format is useful for users to intuitively understand the contents of a dataset and navigate to the required data. The metadata that determines the structure of the classification tree is taxonomic metadata selected from all the metadata handled within the target dataset, and the selection and order of the taxonomic metadata can be freely set by users. The data viewing application allows for the download of the entire set of files stored in a dataset or on an individual data basis.

As applications for administrators, the Instrument management app and the Dataset template app are provided. These applications are for managing the instrument master of the RDE system and the Dataset Templates installed in the system. Basically, system administrators perform tasks such as registering new instrument information, registering new Dataset Templates, and editing them upon request from users.

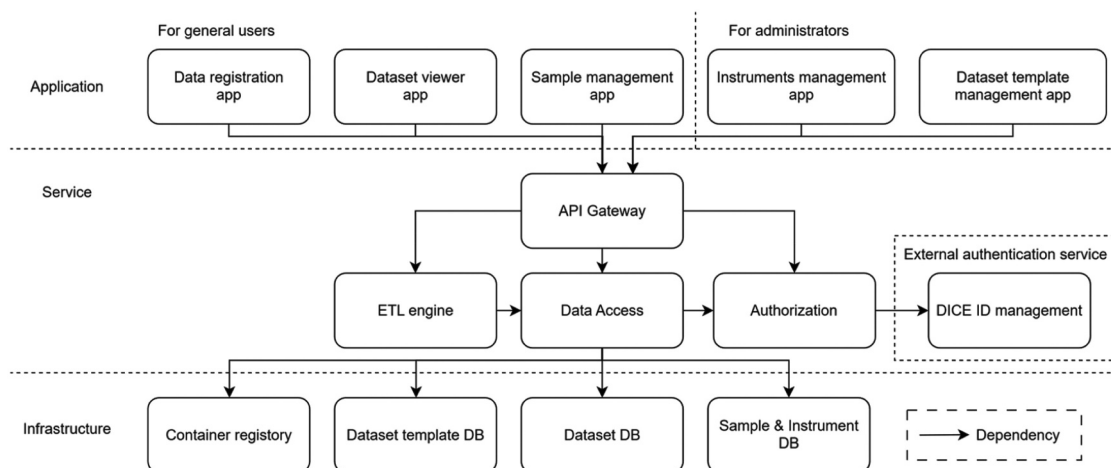


Figure 3. System architecture of RDE.



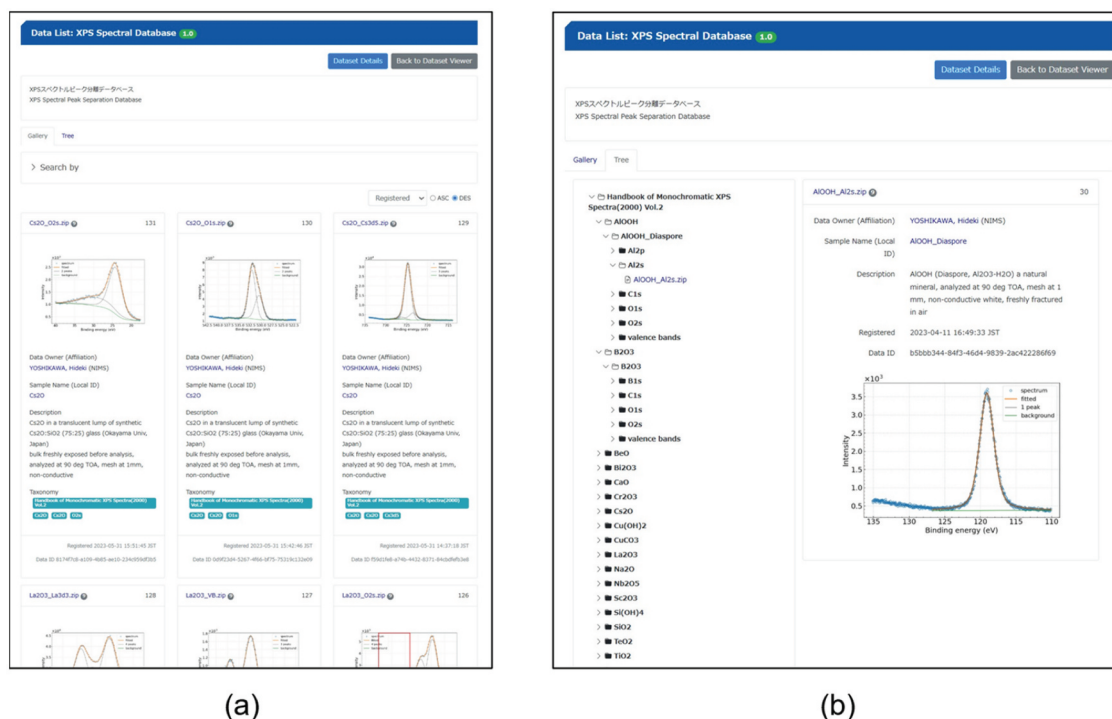


Figure 4. Example views of dataset contents in RDE.

### 3. Data structuring with dataset templates

The core of data management in RDE is defined by ‘Dataset Templates’. This section details how these templates realize everything from defining data structures to automating the registration process, explaining their configuration and functions.

#### 3.1. Structure of dataset templates

The structure of a dataset and the data stored within it are centrally defined by a ‘Dataset Template’. This serves as a blueprint to achieve both the sharing of data description and the efficiency of the registration process. RDE is built utilizing standard open-source technologies, such as PostgreSQL as its database infrastructure and Docker container technology for packaging and executing the structuring (ETL) processes. Regarding the template definition files, invoice.schema.json and catalog.schema.json comply with the common specification (standard) of JSON Schema. On the other hand, components that fulfill RDE’s unique requirements, such as metadata-def.json and the specific control logic for ETL processing, are custom-implemented. These definition files allow materials researchers and data structure designers, who are data users, to flexibly define data structures by editing text or to create new templates by modifying existing ones. The definition files created by the data structure designer can also be obtained by data users as part of the

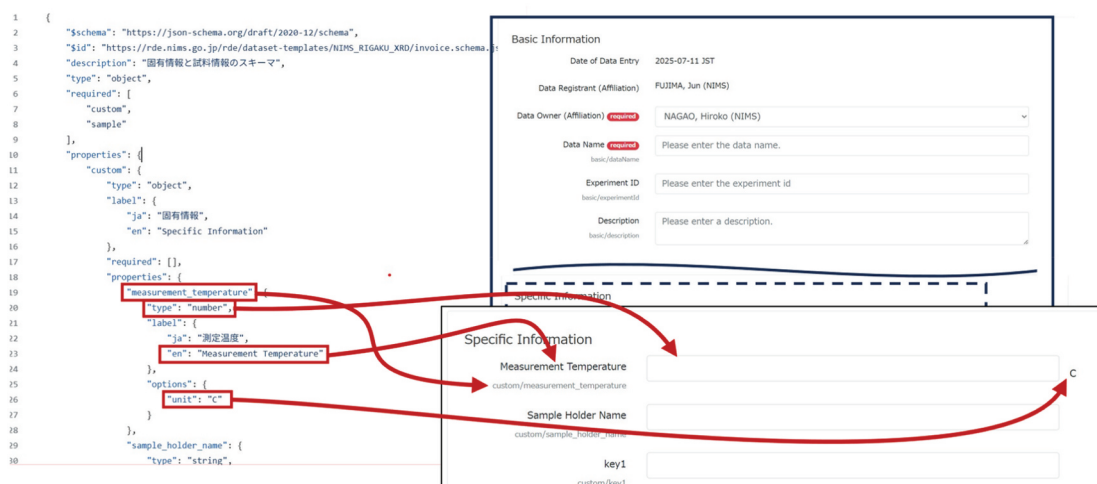
structured data. This ensures the readability of structured data, which varies for each template, for users over the long term.

When a newly created template is registered in the system by an RDE operator, it can be linked to individual instruments registered in the instrument master. This makes it possible to identify which instrument’s output data the dataset is intended to store.

Each Dataset Template is, in principle, to be shared among RDE users, but it is possible to restrict the use of a Dataset Template to specific research groups if a particular data description format or ETL processing needs to be kept confidential.

##### 3.1.1. Invoice schema: defining the data registration interface

invoice.schema.json (Figure 5) defines the configuration of the screen for users to enter information during data registration (the ‘invoice’ screen). In addition to the standard RDE items such as registration date and registrant information, it is possible to add template-specific input items as a ‘custom’ element. Here, items are set for entering information that is not present in the raw files but is necessary for data reuse, such as sample synthesis conditions or environmental information during the experiment. The schema can define an identifier (key) for each input item, its display name on the screen (label), data type (text, number, date, etc.), unit of value, and a description. The RDE system automatically generates an input form with appropriate UI components such as text boxes and



**Figure 5.** Mapping invoice.schema.json to the actual invoice user interface.

calendars based on this definition file. Also, sample information items can be selected from a list of items predefined in the RDE system's sample master. This allows for the sharing of sample information and serves as a clue for selecting and integrating data of the same sample from many types of datasets created for each experimental instrument.

Furthermore, because the input form is automatically generated by the system based on this schema definition, the required metadata items that researchers must manually enter are clarified, which effectively prevents input omissions (data incompleteness) during registration.

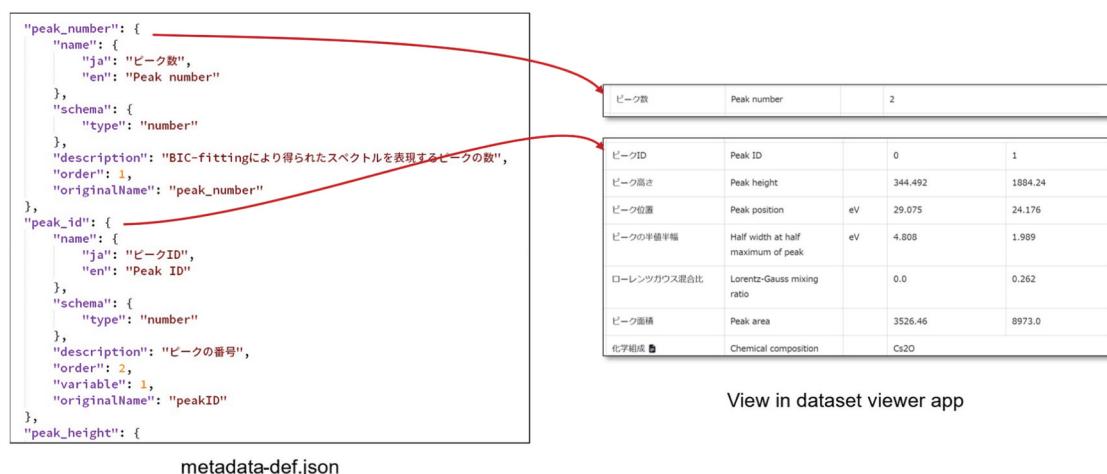
### 3.1.2. Metadata definition: defining metadata fields

metadata-def.json (Figure 6) is a core element of data management that defines the metadata items associated with individual data. This definition greatly influences the findability and reusability of the data.

Metadata items are classified into two types depending on whether the value in a key-value set is

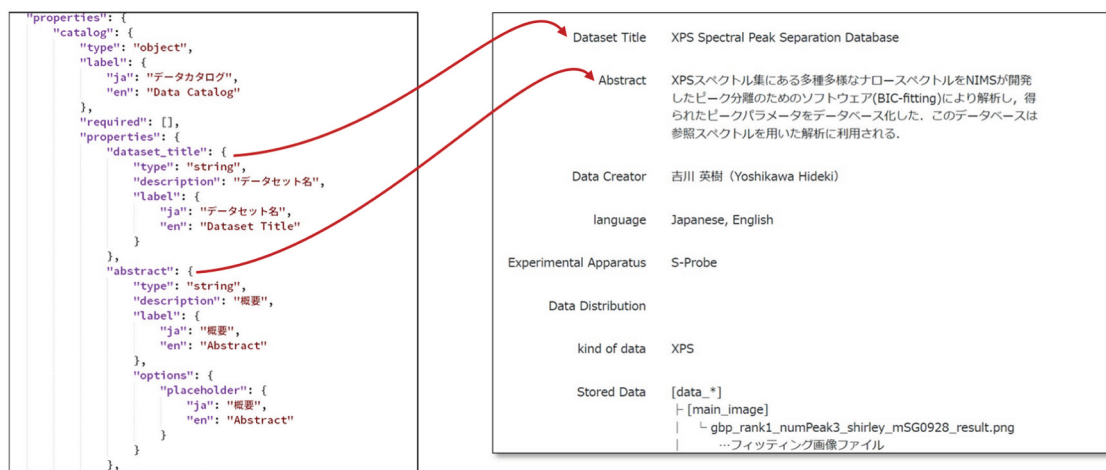
fixed to a single value or not. One is constant, which holds a single value for an item within the data, such as measurement date and sample information. The other is variable, which holds multiple values for an item within the data, such as when storing experimental conditions for multi-point measurements where temperature or pressure is varied. The group of items defined as variable organizes multidimensional experimental data as structured tuples. When viewing the data, it is automatically displayed in a tabular format based on this structure, allowing for an intuitive grasp of complex data relationships.

Depending on the design of the template, in addition to basic metadata for classifying and searching data, such as data descriptions and experimental conditions, it is also possible to store feature values calculated by the processing program from spectral data or image data in the raw files. Feature values are important elements in the initial stages of data analysis, and by calculating and storing them as metadata in advance, users can immediately find the high-value-added data, start analysis or build machine learning



**Figure 6.** Constant and variable items in metadata definition file.





**Figure 7.** Catalog.schema.json defines the format of bibliographic information in the dataset.

models without performing additional preprocessing in later analytical steps.

### 3.1.3. Catalog schema: defining bibliographic information of the dataset

catalog.schema.json (Figure 7) defines items for describing the catalog information of a dataset, that is, comprehensive information such as an overview of the dataset, related project names, and contact information, as well as data format and classification information. This clarifies the purpose and background of the research that created the dataset and the specifications of the data, promoting the discovery of the dataset's value by other researchers and helping them to properly understand how to use the data.

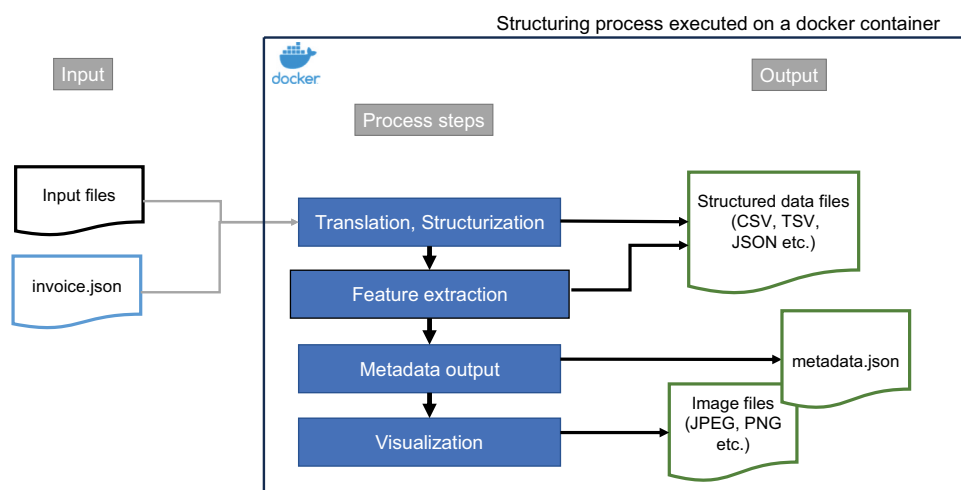
### 3.1.4. Job template and processing programs: automating structuring

The automatic execution of the ETL process based on this template is extremely important for ensuring data quality. Since a series of processes, such as metadata extraction from raw data, unit conversion, and feature

value calculation, are executed with high reproducibility by the program, human errors associated with data organization and transcription conventionally performed manually by researchers are eliminated. This ensures data accuracy and consistency.

jobs.template.yml (Job template) is a configuration file that controls the series of ETL processes executed during data registration. It describes the command of the processing program to be executed, the name of the container image containing the program, and the dependencies between multiple processes.

Figure 8 shows the processing flow executed by the processing program (mainly written in Python), which converts input raw files and manually entered data into a prescribed data structure based on the items defined in the metadata definition. A typical processing flow consists of the following steps: ① decoding the instrument-specific file format, ② merging with manually entered data, ③ making manufacturer-specific terms, including abbreviations, readable and not manufacturer-specific, ④ converting numerical data into a highly readable tabular format (such as



**Figure 8.** Data processing flow of the structuring program in RDE.

CSV with a simple header), ⑤ calculating feature values, and ⑥ visualizing the results (generating graph images). The feature values calculated here are stored as metadata according to the Metadata Definition, accelerating subsequent data searches and applications in machine learning.

Finally, all the files generated through these processes are arranged according to the standard folder structure defined by RDE (Table 1). For example, raw files are stored in the raw folder, and structured metadata is stored in metadata/metadata.json. This standardized folder structure enables consistent data access control for any dataset.

The previously mentioned one-shot registration of multiple data is realized by an extension of this common folder structure. Specifically, the ETL processing program generates a folder named divided in the root of the output directory. Inside this divided folder, sequential subfolders, such as 001, 002, etc., are created, and within each of these subfolders, the common folder structure shown in Table 1 (e.g. raw, metadata, structured) is placed. When the RDE system detects this divided folder in the processing results, it interprets the group of subfolders underneath it as independent data (tiles) and registers them into the dataset in a batch.

### 3.2. Packaging of the structuring process

In RDE, data structuring is not merely about storing data – it is an active process that transforms raw data into valuable information assets. To ensure reproducibility and extensibility, the structuring process is packaged using container technology such as Docker. Here, we explain a typical workflow using the example of handling XRD measurement data.

- (1) **Data registration:** A researcher accesses RDE through a web browser and selects the target dataset. An ‘invoice’ screen, automatically generated based on the Invoice Schema, is

displayed. The researcher manually enters the sample ID and supplementary information about the measurement (e.g. measurement atmosphere) and uploads the raw data file output from the XRD instrument (e.g. measurement.txt).

- (2) **Process initiation:** Data registration triggers the RDE’s Data-structuring processor. The processor reads the Job Template associated with the dataset, retrieves the specified Docker container image (e.g. nims/xrd-analysis:1.0) from the container registry, and executes it as an independent container instance.
- (3) **ETL processing within the container:** In the container, a predefined Python script is executed. This script performs processing according to the common file system structure provided by RDE (Table 1).

- **Input:** The script reads measurement.txt from the inputdata folder and invoice.json from the invoice folder. measurement.txt is copied to the raw folder.
- **Analysis and feature extraction:** Using libraries such as scipy and pymatgen, the script extracts the 2 $\theta$  angle, intensity, and full width at half maximum (FWHM) of numerous diffraction peaks from the numerical data series of the XRD profile recorded in measurement.txt through an automatic peak search program represented by a model function. Furthermore, it calculates the average crystallite size using the Scherrer equation from the FWHM of the main peaks.
- **Visualization:** Using the matplotlib library, a graph plotting the XRD profile is generated and saved as xrd\_plot.png in the main\_image folder, and its thumbnail as thumbnail.png in the thumbnail folder.
- **Generation of readable numerical data:** The numerical data series of the XRD profile is converted into a simple, highly readable tabular format independent of the XRD

**Table 1.** List of input and output files and their stored folders on the RDE structuring process.

Folder Name	Description	Example Files
invoice	Stores manually entered information from the invoice screen.	invoice.json
raw	Shareable raw files used as input for structuring.	Output files from measurement devices, Excel files in which the experimenter describes experimental conditions, etc.
nonshared_raw	Raw files not intended for sharing. Some raw files are used as input for structuring.	Internal notes, intermediate files, raw files with manufacturer copyright in the description format etc.
metadata	Extracted or generated metadata stored in JSON format.	metadata.json
structured	Structured data in machine-readable formats (CSV, HDF5, etc.).	Analyzed tabular data, feature data, etc.
main_image	Primary image displayed on the data detail screen.	Representative plots, microscope images, etc.
other_image	Supplementary images.	Additional graphs, photos, etc.
thumbnail	Thumbnail images displayed on the data list screen.	thumbnail.png
filemeta.json (root)	Metadata about each file within the dataset.	File size, creation date, file format, etc.

instrument manufacturer and written out as `xrd_profile.csv` in the structured folder.

- **Metadata generation:** All information is aggregated, and a `metadata.json` file is generated in the metadata folder. The XRD instrument settings and measurement conditions recorded in `measurement.txt`, as well as the manually entered sample ID and measurement atmosphere, are recorded as constant metadata. The parameters of the multiple extracted peaks ( $2\theta$  angle, intensity, FWHM) are recorded as variable metadata, and the calculated average crystallite size is recorded as a new metadata item. When recording the XRD instrument settings and measurement conditions from `measurement.txt`, if unique abbreviations or terms specific to the XRD instrument manufacturer are used, they are translated into more readable, generic terms before being recorded.

(4) **Registration completion:** Once the processing is complete, all files generated within the container (e.g. `metadata.json`, `metadata-def.json`, `xrd_profile.csv`, `xrd_plot.png`) are permanently saved in the RDE's database and file storage, and the user can view the results through the Dataset viewer app.

Figure 9 shows the flow of the ETL processing execution. Developers create a Docker image that includes all necessary libraries and tools for data processing. This image is registered in the RDE's container registry and associated with a specific dataset via a Dataset Template. Triggered by a user's data registration operation, the system launches a container from the corresponding image and executes the ETL processing in a completely isolated environment.

This architecture offers two major advantages. First, it eliminates 'environment-dependent issues' arising from differences between the development

and execution environments, ensuring consistently stable processing results. Second, it achieves high scalability. Even when numerous data registrations occur simultaneously, the system can distribute the load by launching container instances in parallel according to the number of requests, enabling processing without delay. Thus, the packaging of processes is a foundational technology that supports both the robustness and scalability of the system.

#### 4. RDE utilization example

RDE is utilized as a core data infrastructure in the Materials DX Platform initiative promoted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (see 'Strategy for Strengthening Material Innovation Power' [9]). Full-scale operation began in January 2023, and it is used as a data management infrastructure for various programs and research projects, such as the Advanced Research Infrastructure for Materials and Nanotechnology (ARIM) in Japan [10] and the Data creation and utilization-type MaTerial R&D project (DxMT) [11]. Data generated using shared equipment for synthesis, processing, and measurement at 25 institutions in Japan participating in ARIM, as well as data from the five materials fields of structural materials, magnetic materials, battery and water electrolysis materials, ceramic semiconductor materials, and polymer/biomaterials created in DxMT, is accumulated and managed in RDE, promoting the DX of materials research as a whole. To date, the number of RDE users has exceeded 3000, over 1000 Dataset Templates have been implemented, more than 7000 datasets have been created, and the total number of accumulated data files has surpassed one million. Behind this rapid adoption and high utilization record in such a short period

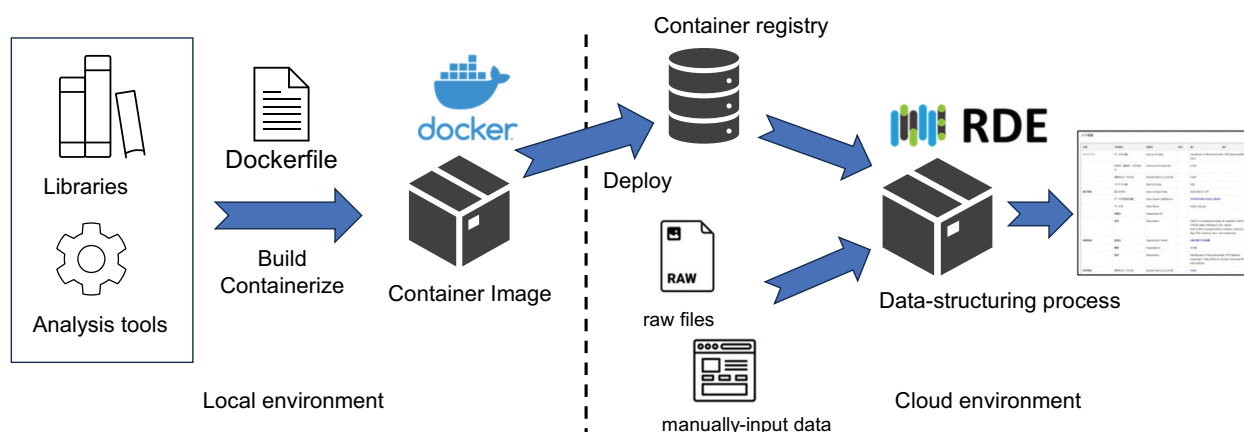


Figure 9. Workflow from development to application of a structuring program in RDE.

lies not only the system's functionality but also a robust support system for users. We continuously provide detailed documentation covering everything from basic RDE usage to methods for developing dataset templates, which are a core function. Furthermore, for research groups wishing to start using the system, our development team provides close support, such as hands-on development assistance, to develop templates tailored to their research content. The use of RDE as such a large-scale data management and sharing infrastructure strongly supports the digital transformation of materials research in Japan.

As a specific example of a dataset configuration, we introduce a case where Dr. Yukiko Takahashi of the NIMS Research Center for Magnetic and Spintronic Materials consolidated magnetic material synthesis experiment recipes and various measurement data into an RDE dataset (Figure 10). In her research, for magnetic thin films produced under various synthesis conditions, the synthesis conditions, characterization, and property information are linked and stored in a dataset.

Synthesis conditions are summarized in a Microsoft Excel spreadsheet, and by registering this in RDE, each row of the table is automatically saved as one data tile. At this time, the one-shot registration mechanism of RDE, as mentioned earlier in this paper, is utilized. That is, by uploading the Excel file describing the synthesis conditions and the group of related raw files output from multiple different measurement instruments such as XRD, superconducting quantum interference device (SQUID), and vibrating sample magnetometer (VSM) in a single operation, each is registered as an independent data tile. The synthesized

thin films are characterized by various measurement techniques such as XRD, SQUID, and VSM. The results of these measurements are also saved as individual data tiles on RDE, with each condition item stored as metadata. For transmission electron microscope (TEM) data, a segmentation program developed by Dr. Takahashi's research group [12] is incorporated as an ETL process in RDE, and automatic segmentation is performed at the time of data registration to save the grain size distribution. The measurement results from the VSM are visualized as hysteresis curves using a tool developed at NIMS-MDPF, and the coercivity, remanent magnetization, and maximum magnetic flux density are automatically calculated and saved as metadata. By storing everything from synthesis conditions to measurement results and post-process results together in this way, it becomes easier to apply further data analysis and machine learning, thereby enhancing the reusability of the data.

## 5. Comparison with existing initiatives

Data management platforms in the materials science field share the common goal of realizing the FAIR principles, but they differ significantly in their core approach to data description. This difference in approach directly determines their ability to respond to diverse research needs (flexibility) and their capacity to handle large volumes of data (scalability). In this section, we compare RDE with other major platforms from the perspective of 'how data structures are defined and managed', as shown in Table 2.

The Materials Project [13] and OQMD [14], widely used in computational materials science, primarily

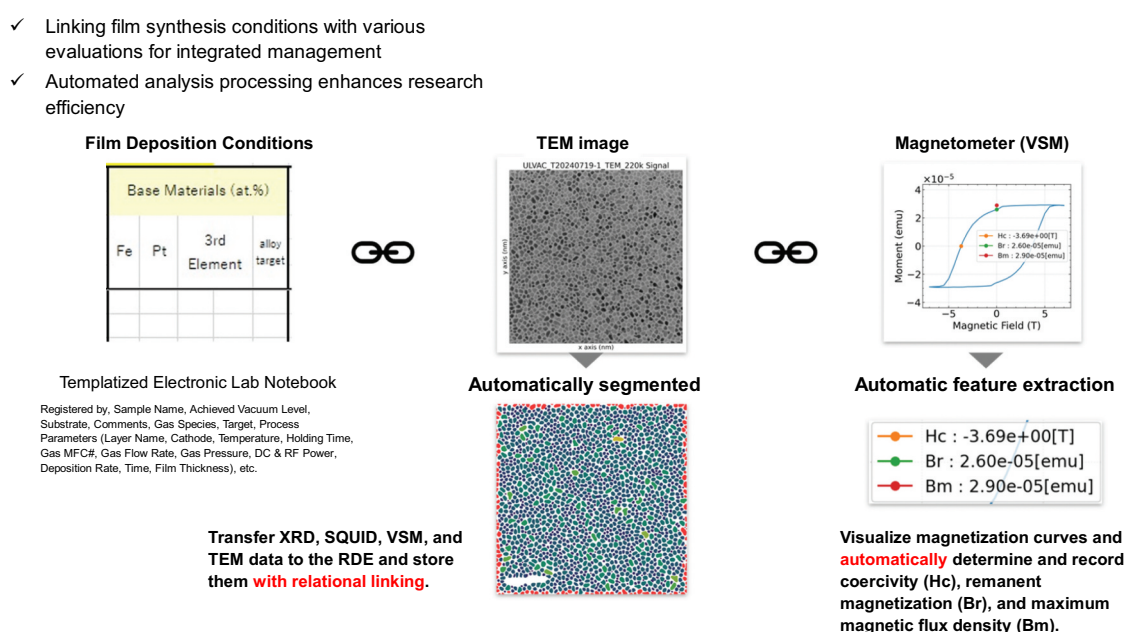


Figure 10. Application of RDE in magnetic materials research (Yukiko Takahashi, NIMS).



**Table 2.** Functional comparison between RDE and other materials data platforms.

Data Platform	Data Description Paradigm	Computational Data	Experimental Data	Flexibility	Scalability
RDE	Defines data structure via templates	✓	✓	***	***
Materials Project/OQMD	Applies a single, fixed schema	✓		*	***
NOMAD/FAIRmat	Extends a common base structure	✓	✓	**	***
Citrine (GEMD)	Describes via a specific process model		✓	*	**
Materials Commons	Describes via a specific workflow model	✓	✓	*	*
Materials Data Facility (MDF)	User-defined structure	✓	✓	***	*
MaterialDigital	Defines meaning with an ontology	✓	✓	**	**

\*\*\* high/comprehensive, \*\* medium, \* limited.

handle computational data and thus apply a single, strictly defined data format (schema). This approach guarantees data consistency and streamlines the search and use of large datasets, making it excellent in terms of scalability. However, this strict standardization is a trade-off with the flexibility required to handle experimental data from diverse methods and instruments. Citrine [15] and Materials Commons [16], which specialize in managing the provenance of material processes and properties, similarly adopt a method of recording data in the form of a predefined, specific workflow. This also means that while data conforming to the defined model can be handled efficiently, there are constraints on the flexibility to handle data outside that framework.

In contrast, the Materials Data Facility (MDF) [17], which aims to promote data publication, allows for highly flexible data sharing by essentially leaving the data structure up to the user. In this respect, it can be said to have the highest flexibility, as it can register any type of data. However, because the structures are not standardized, it is difficult to process and analyze the entire registered data landscape comprehensively or automatically, posing challenges in terms of data processing scalability.

NOMAD/FAIRmat [18,19], a large-scale data infrastructure initiative in Europe, and MaterialDigital [20,21], promoted as a national project in Germany, take an intermediate approach. NOMAD provides a common ‘grammar’ for data description, which is extended for specific applications through plugins. MaterialDigital strictly defines the relationships between data using a specialized dictionary (ontology). These approaches maintain data consistency and achieve good scalability by adhering to common rules or dictionaries while retaining a degree of flexibility. However, since all data must comply with central rules, there can be constraints in rapidly introducing entirely new data formats.

In contrast, RDE does not standardize the data format itself but rather standardizes the ‘method of defining the data format’ using self-contained packages called ‘Dataset Templates’. Each template

bundles the data structure definition required for a specific experiment or calculation, an input form, and a data processing program. This allows researchers to freely create new templates for new experimental methods, providing extremely high flexibility. Furthermore, the processing for each data registration is executed in parallel in an independent container environment for each template, distributing the overall system load and achieving high scalability to handle numerous registration requests. The RDE approach is a practical solution that balances the flexibility to meet the diverse needs of the research field with the scalability required of a large-scale data infrastructure.

## 6. Conclusion

In this paper, we have demonstrated that the Research Data Express (RDE) system for accumulating and sharing materials research data is an effective tool for strongly promoting the digital transformation (DX) of materials research.

The core concept of RDE, the ‘Dataset Template’, adopts an approach of standardizing the ‘method for defining data formats’ rather than the data formats themselves. This makes it possible to simultaneously satisfy the often-competing requirements of flexibility to accommodate diverse experimental data, and scalability and reproducibility through automated ETL processing using container technology.

In practice, RDE is widely used as the core data infrastructure of the Materials DX Platform, and its track record – with over 3000 users and more than 7000 datasets created – confirms the system’s effectiveness within the Japanese materials research community. Whereas many existing data infrastructures are centered on computational data, RDE possesses a distinct uniqueness in its strength for managing and structuring diverse experimental data, thereby meeting the practical needs of the research front.

Looking ahead, linking the high-quality structured data accumulated in RDE with AI analysis



systems under development at the National Institute for Materials Science (NIMS) is expected to accelerate the design and discovery of new materials through a data-driven approach. Furthermore, a key challenge is to ensure interoperability with international data-sharing infrastructures and to develop RDE into a more open platform for scientific discovery. As a specific approach for this, we plan to map the structured metadata accumulated in RDE to established common vocabularies, such as schema.org [22] and Data Catalog Vocabulary (DCAT) [23], using JSON for Linking Data (JSON-LD) [24]. This will further enhance the FAIR principles, especially Interoperability. Furthermore, since the metadata structure of each data is explicitly defined by dataset templates (e.g. JSON Schema) in this system, it is anticipated that Large Language Models (LLMs) will interpret these definitions in the future to automate the mapping between datasets with different definitions and support data integration.

In addition, to promote reproducibility and community use of the system, we have released an open-source software toolkit (RDEToolKit) [25] for developing and executing the core RDE data structuring process in a local environment, along with specific examples of templates for major measurement techniques. This allows the research community to implement and test the RDE data management methodology in their own environment and efficiently develop their own templates.

## Acknowledgements

In this research, we extend our deepest gratitude to Dr. Yukiko Takahashi of the Research Center for Magnetic and Spintronic Materials at the National Institute for Materials Science for providing an important research case study that demonstrated the effectiveness of RDE. Her cooperation and valuable data were crucial elements that supported this study. We sincerely thank her.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Jun Fujima  <http://orcid.org/0000-0002-0858-8035>  
 Hideki Yoshikawa  <http://orcid.org/0000-0002-7389-8865>  
 Masahiko Demura  <http://orcid.org/0000-0002-7308-3041>

## References

- [1] Kimmig J, Zechel S, Schubert US. Digital transformation in materials science: a paradigm change in

- material's development. *Adv Mater* [Internet]. 2021;33(8):2004940. doi: [10.1002/adma.202004940](https://doi.org/10.1002/adma.202004940)
- [2] Ghiringhelli LM, Baldauf C, Bereau T, et al. Shared metadata for data-centric materials science. *Sci Data*. 2023;10(1). doi: [10.1038/s41597-023-02501-8](https://doi.org/10.1038/s41597-023-02501-8)
- [3] Ghiringhelli LM, Carbogno C, Levchenko S, et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *NPJ Comput Mater*. 2017;3(1). doi: [10.1038/s41524-017-0048-5](https://doi.org/10.1038/s41524-017-0048-5)
- [4] Brinson LC, Bartolo LM, Blaiszik B, et al. Community action on fair data will fuel a revolution in materials research. *MRS Bull* [Internet]. 2024;49(1):12–16. doi: [10.1557/s43577-023-00498-4](https://doi.org/10.1557/s43577-023-00498-4)
- [5] Tanifuji M, Matsuda A, Yoshikawa H. Materials data platform - a fair system for data-driven materials science. In: 2019 8th international congress on advanced applied informatics (IIAI-AAI); Toyama, Japan; 2019. p. 1021–1022. <http://ieeexplore.ieee.org/document/8992690>
- [6] Tanifuji M, Yoshikawa H. Research data flow in the materials data platform system DICE — practice and future visions. *IPSI Trans Digit Pract*. 2021;2:57–63.
- [7] Matsunami S, Matsuda A, Chikyow T, et al. Data architecture for IoT data collection system. *IPSI Trans Digit Pract*. 2021;2:80–89.
- [8] RDE(RDE) - dice: national Institute for Materials Science (NIMS) [Internet]. Available from: <https://dice.nims.go.jp/services/RDE/>
- [9] Ministry of Education, Culture, Sports, Science and Technology (MEXT). Sosharu ojekuto toshite no zairyō dejitaru purattofōmu [materials digital platform as a social object] [Internet]. 2023. Available from: [https://www.mext.go.jp/content/20231115-mxt-mxt\\_nanozai-000032704\\_8.pdf](https://www.mext.go.jp/content/20231115-mxt-mxt_nanozai-000032704_8.pdf)
- [10] Arim Japan official homepage [Internet]. Available from: <https://nanonet.go.jp/page/dir000011.html>
- [11] Dxmt portal [Internet]. Available from: <https://dxmt.mext.go.jp/en/>
- [12] Kulesh N, Bolyachkin A, Suzuki I, et al. Data-driven optimization of FePt heat-assisted magnetic recording media accelerated by deep learning TEM image segmentation. *Acta Mater* [Internet]. 2023 [cited 2025 Jul 11];255:119039. doi: [10.1016/j.actamat.2023.119039](https://doi.org/10.1016/j.actamat.2023.119039)
- [13] Jain A, Montoya J, Dwaraknath S, et al. The materials project: accelerating materials design through theory-driven data and tools. In: Andreoni W, Yip S, editors. *Handbook of materials modeling: methods: theory and modeling* [Internet]. Cham: Springer International Publishing; 2018 [cited 2024 Nov 25]. p. 1–34. doi: [10.1007/978-3-319-42913-7\\_60-1](https://doi.org/10.1007/978-3-319-42913-7_60-1)
- [14] Kirklin S, Saal JE, Meredig B, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *NPJ Comput Mater* [Internet]. 2015;1(1):15010. doi: [10.1038/npjcompumats.2015.10](https://doi.org/10.1038/npjcompumats.2015.10)
- [15] 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* [Internet]. 2016 [cited 2025 Jul 11];68(8):2031–2034. doi: [10.1007/s11837-016-1984-0](https://doi.org/10.1007/s11837-016-1984-0)
- [16] Puchala B, Tarcea G, Marquis EA, et al. The materials commons: a collaboration platform and information repository for the global materials community. *JOM* [Internet]. 2016 [cited 2024 Dec 23];68(8):2035–2044. doi: [10.1007/s11837-016-1998-7](https://doi.org/10.1007/s11837-016-1998-7)

- [17] Blaiszik B, Chard K, Pruyne J, et al. The materials data facility: data services to advance materials science research. JOM [Internet]. 2016 [cited 2024 Oct 31];68(8):2045–2052. doi: [10.1007/s11837-016-2001-3](https://doi.org/10.1007/s11837-016-2001-3)
- [18] Scheidgen M, Himanen L, Ladines AN, et al. Nomad: a distributed web-based platform for managing materials science research data. J Open Source Softw [Internet]. 2023 [cited 2024 Nov 25];8(90):5388. doi: [10.21105/joss.05388](https://doi.org/10.21105/joss.05388)
- [19] Nationale Forschungsdateninfrastruktur (NFDI). Consortia FAIRmat | NFDI [Internet]. Available from: <https://www.nfdi.de/consortia-fairmat/?lang=en>
- [20] MaterialDigital. The material digitalization platform [Internet]. Available from: <https://www.materialdigital.de/>
- [21] Bekemeier S, Caldeira Rêgo CR, Mai HL, et al. Advancing digital transformation in material science: the role of workflows within the MaterialDigital initiative. Adv Eng Mater [Internet]. 2025;27(8):2402149. doi: [10.1002/adem.202402149](https://doi.org/10.1002/adem.202402149)
- [22] Schema.org [internet]. Available from: <https://schema.org/>
- [23] Data catalog vocabulary (DCAT) [Internet]. Available from: <https://www.w3.org/TR/vocab-dcat-3/>
- [24] Json for linking data [Internet]. Available from: <https://json-ld.org/>
- [25] Rdetoolkit [internet]. Available from: <https://github.com/nims-mdpf/rdetoolkit>