Experimental Data Management Platform for Materials Informatics

Hideya Yoshiuchi*, Hiroki Miyamoto, Sayaka Tanimoto
Hitachi, Ltd., Research & Development Group, Kokubunji, Tokyo, Japan.

* Corresponding author. Tel.: +81-80-9874-6036; email: hideya.yoshiuchi.fq@hitachi.com
Manuscript submitted August 3, 2022; revised October 12, 2022; accepted December 30, 2022.
doi: 10.17706/jsw.18.2.99-105

Abstract: We are promoting the research and development of materials informatics, which realizes the efficiency and sophistication of material research by utilizing a huge amount of experimental data. In order to accelerate new material research and development, it is important to predict the results of new ma-terial development by unimplemented experiments from past experimental data, and for that purpose, a large amount of experimental data generated through various material research processes is used for business. The challenge of material research is to reduce the time for search and acquire data from enormous experiment data. In this paper, we show the experimental da-ta management platform by describing the research and development work-flow with a unified data model. Through prototyping and evaluating the pro-posed technique, we confirmed that the time required for system construction and data acquisition of required experimental data can be reduced by an average of 94.4% in the target work.

Key words: Materials informatics, data management platform.

1. Introduction

In recent years, in response to changes in social trends such as a shortage of working population mainly in developed countries, globalization, and changes in consumer needs, expectation of Internet of Things (IoT) [1] and Artificial Intelligence (AI) [2] technology is rising. The scope of application of these technologies is wide-ranging, such as improving the efficiency of management operations of buildings for commercial and office, improving the production efficiency of factories that produce various products, and forecasting trends in financial products. Materials Informatics (MI) [3], which accelerates material research by utilizing information science [4, 5], is attracting attention.

MI is a technique aimed at extracting knowledge from data in material science [5, 6]. For example, MI can develop a material with better physical properties by creating a prediction model of the physical properties from the experimental data of the physical properties of the substance created in the past experiment and applying the model to unknown data. It is one of the purposes. For this purpose, a technique for collecting experimental data conducted in the past and efficiently searching for it and a technique for predicting the characteristics of a substance by combining and learning the collected data are required.

In materials development, available data includes public information published in the form of papers and patents in the past, evaluation results of physical properties of materials collected in R&D activities conducted inside company, and various workflows and working conditions under which the materials were created. Since these data are recorded and collected by the individual developer who collected the data, there is a problem that the format and structure of the data has many variations. Therefore, it is necessary to develop a data management platform that enables the acquisition of useful data for future research and development from the vast amount of experimental data that has been collected in the past in a form that can be understood by those involved. In order to solve these problems, this study examines and evaluates the requirements, system architecture, system construction, and reduction of management man-hours for an experimental data management infrastructure that enables MI to collect and man-age data in a unified manner.

2. Materials Informatics

2.1. Overview of Materials Informatics

The conventional materials development process has relied heavily on the knowledge, experience, and abilities of engineers. The general flow of research and development is to conduct theoretical calculations to meet the needs, search for existing research, repeat experiments to fabricate prototypes of materials, and proceed with the evaluation of physical properties.

MI is an attempt to speed up materials development through the power of computational science and information science. MI aims to accelerate materials research by having computers calculate physical properties such as atomic arrangement, or by analyzing past simulation data and publication data through machine learning.

MI collects experimental data by synthesizing materials, which is a fundamental activity in materials research, preparing test samples, and evaluating various physical properties of the materials, and then accumulates the collected data and the results of analyzing the data. MI uses the accumulated data to predict the results of experiments that have not yet been conducted through machine learning and simulations. By iterating this series of processes, it is expected to significant-ly reduce the man-hours required for materials research. From cutting-edge technologies and research, such as digital technology and biotechnology, to materials that support our daily lives, materials research is considered to be the corner-stone of industry and innovation. In order to efficiently apply and operate MI, it is essential to establish a system that enables data sharing beyond the boundaries of individual companies.

2.2. Issues for Applying Materials Informatics

Since the objective of MI is to develop new materials in a short time by ma-chine learning and simulation of accumulated data, it is important to collect a large amount of data from actual experiments and manage them in an easy-to-use format.

1) The key to an MI system is the design of the data model. It is important to incorporate the requests of system users to the greatest extent possible and to realize the management and sharing of data required by the users. Data management in MI systems involves the following issues:
2) There is no common understanding of the material development process, and there are discrepancies in process recognition among developers
3) Prior to the introduction of MI, experimental data was managed individually by each developer, and the data was in different formats, making it time-consuming to search for the desired data.
4) Interfaces for efficient data retrieval are not in place

To solve these issues, it is necessary to manage and visualize the material development process in a unified manner for all parties involved, and to obtain the actual experimental data conducted in accordance with the aforementioned unified process.

3. Experimental Data Management Platform for Materials Informatics

3.1. Requirements for Experimental Data Management Infrastructure

In the management of experimental data, which plays an important role in MI, it is necessary for engineers involved in materials development to be able to find the desired data quickly. For this purpose, it is essential to define a unified data format to represent all experimental data, and to provide functions for easy retrieval, visualization, and acquisition of data through a data management infra-structure. Based on this perspective, the requirements for a data management infrastructure for MI are defined as follows:

3.1.1 Workflow management of R&D process

To centrally manage information related to the workflow in materials re-search and development, such as the sequence of work, composition of materials, and equipment used in processing, as a data model, so that all par-ties involved can share the information.

3.1.2. Design of experimental data management database

Design a database to manage data from experiments, including those con-ducted in the past, so that it can be searched and retrieved with traceability in a format consistent with the research and development flow.

3.1.3. Advanced data search and retrieval

In addition to the normal data retrieval application interface (API), it is possible to retrieve experimental data related to newly developed materials by setting appropriate conditions.

3.2. System Architecture of Experimental Data Management Platform

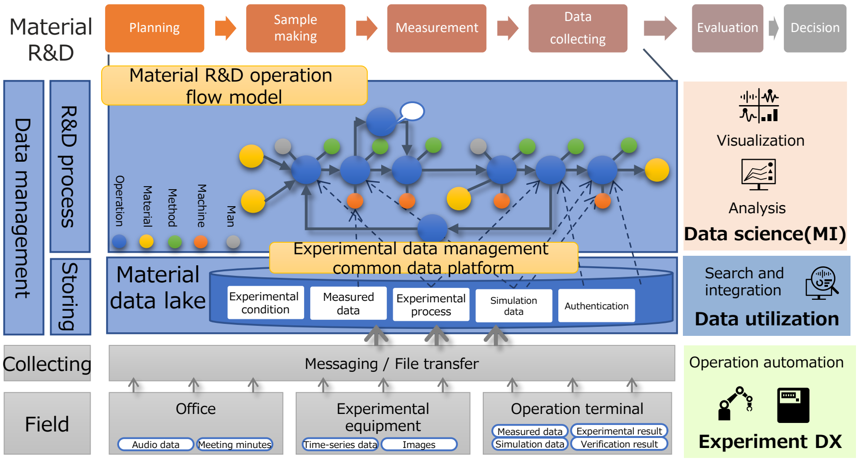


Fig. 1. System architecture of experimental data management platform for MI.

Fig. 1 shows the system architecture of the experimental data management platform. The purpose of the experimental data management platform is to man-age various data generated in the field during a series of material research and development processes from the planning stage to the decision to commercialize a product. For example, in the process of actual R&D after the plan is approved, various experimental conditions and simple measurement results related to raw material processing are generated as data. At the stage where test samples are completed and evaluated, measurement data obtained by various measuring instruments and evaluation results obtained by analyzing the measurement data are accumulated. These experimental data are generated by various operations and stored in a data lake. The data models map the experimental data to the tasks described in the materials research business process model. By using these data models, developers can easily obtain the actual data of experiments conducted in the past in the work related to their own research and development work.

3.3. Workflow Management of Material Research and Development Processes

In materials research, various combinations of raw materials and processing methods are considered to produce the material to be developed. For example, for the combination of raw materials, the types of raw materials, amounts of raw materials, and timing of raw material input are considered in the planning stage of experiments. Regarding the processing methods of raw materials, there are tasks such as polymerization and heating, which are always necessary when develop-ing any kind of material, and tasks such as drying, agitation, and aging processing, which are selectively incorporated into experiments with the aim of bringing about changes in the properties of the generated sample depending on whether it is processed or not. By varying the sample preparation conditions, samples of different compositions are prepared, and decisions on product commercialization are made by accumulating evaluation results for the prepared samples from various viewpoints. In order to smoothly carry out these series of operations in materials research, a method to model the workflow that comprehensively covers all operations that may occur in the research and development industry in a unified format and to describe the process flow by means of the model is considered.

In this study, we solve this problem by using modeling tool to describe the workflow of R&D work using operation and 4M (Man, Machine, Method, Material) information. In the workflow of materials research and development work, it is possible for all members to share a model of the workflow that comprehensive-ly covers all tasks that may occur in the work, so that they can coordinate their understanding of the workflow and accumulate data from experiments that each researcher is in charge of.

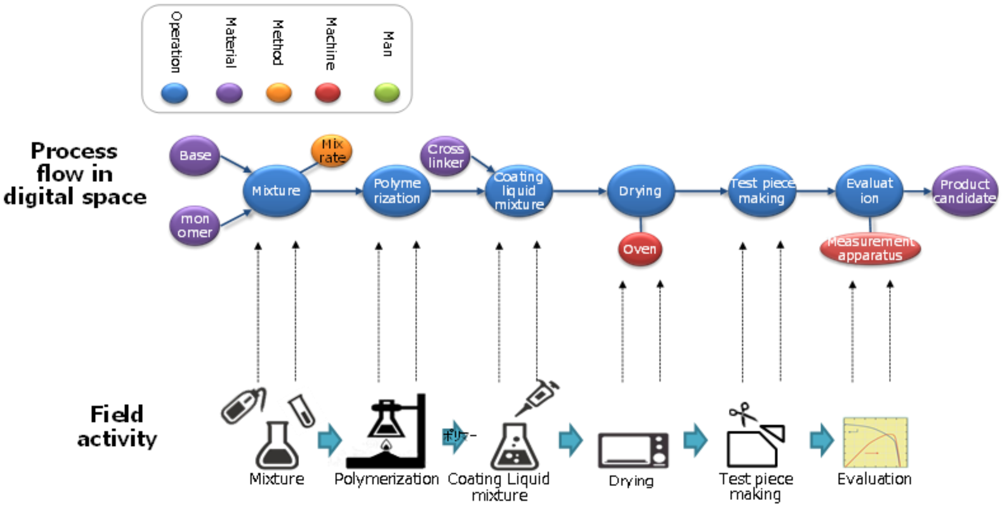


Fig. 2. Example of process flow model.

Fig. 2 shows an example of a workflow data model. The workflow is a digital space that represents the activities conducted in the field by using tasks and 4M items.

3.4. Design of Experimental Data Management Database

In materials research and development, multiple raw materials are repeatedly combined to form compounds, which are then combined to form intermediate products. In the processing of compounds, the characteristics of the products are changed by changing the working conditions, and through evaluation of the final products, the raw materials and processing methods suitable for commercialization are narrowed down. Sample IDs are used to distinguish products, but product characteristics can be changed to different samples by adding or removing new materials or compounds to the product, or by changing working conditions such as temperature, humidity, or processing speed in a given operation. In order to manage samples that change depending on the work, in the database that manages experimental data, the table that manages sample IDs and the table that manages experimental data for each work should be managed separately. The sample ID management table should be prepared for each work whose sample ID may change. The experimental data management table for each operation is referred by the sample ID in the sample ID management table as a foreign key. This structure allows us to identify the products used in each job and express the connection between jobs by products and materials, while guaranteeing the unique-ness of the sample IDs.

3.5. Advanced Data Search and Retrieval

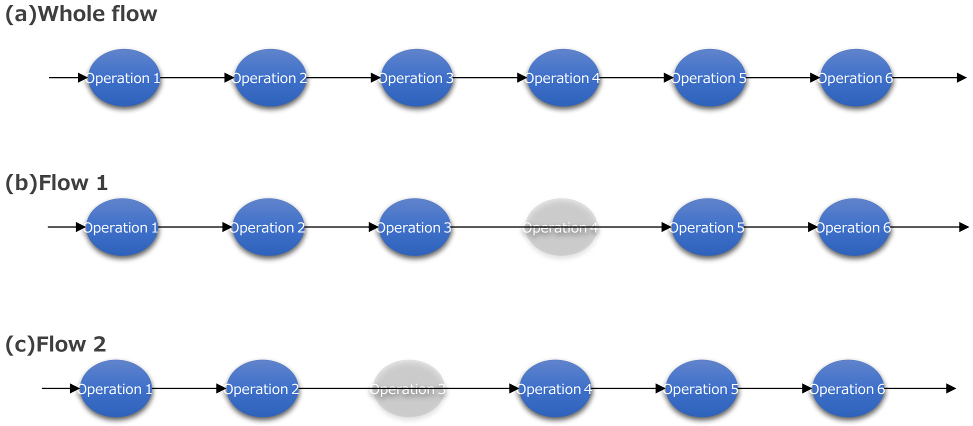


Fig. 3. Example of flexible process flow model.

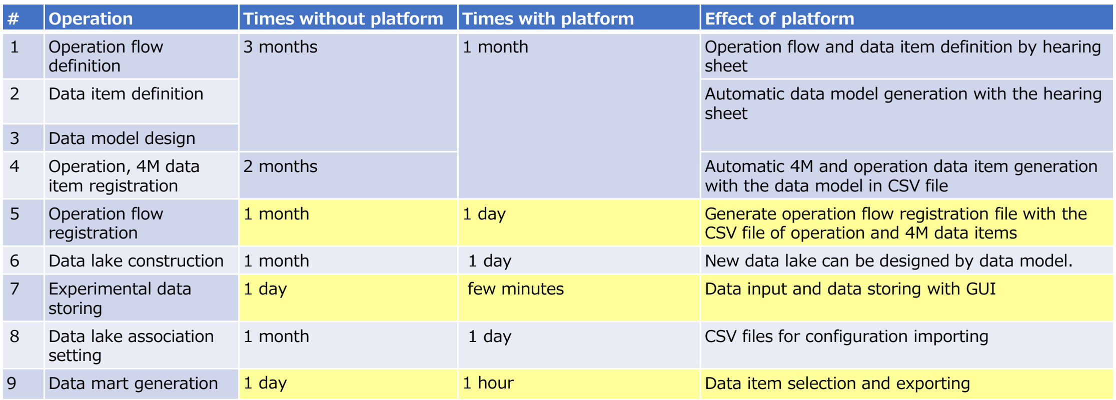
Fig. 3 shows an example of a materials research and development flow with various patterns. In this example, the list of processes required to prepare samples for material processing is fixed, and it is assumed that the composition of the final product is changed by selecting the process to be performed, and that characterization is performed by selecting the process to be performed. In Fig. 3, (a) is a model that includes all possible processing operations in R&D, while the fourth operation (Operation 4) and the third operation (Operation 3) are not performed in Flow 1 and Flow 2 in (b) and (c), respectively. By creating a business flow model for all processing operations that may be implemented in R&D work in the overall flow in (a), and managing experimental data corresponding to (b) or (c) or other experimental operations depending on the actual operations, it is possible to manage experimental data corresponding to all possible manufacturing flows that may be assumed in R&D. In the flow shown in Fig. 3, we consider retrieving the experimental data according to the business flow (b). In this case, experimental data corresponding to Operations 1, 2, 3, 5, and 6 are to be extracted, and such data may include experimental data obtained through Business Flow (a) as well as Business Flow (b). However, there is a difference between business flow (b) and (a) in terms of the presence or absence of Operation 4, and these data cannot be compared in the same line. Therefore, when retrieving data, the similarity of business flow is taken into account.

4. Prototype Development and Evaluation

We designed and developed an experimental data management platform through collaborative creation activities with our customer. Customer’s operation flow consists of 28 operations with related material and method items. The following procedure was used to construct the MI system with experimental data management platform.

1) Operation Flow Definition
2) Data item definition
3) Data model design

4) Operation, 4M data item registration
5) Operation flow registration
6) Data lake construction
7) Experimental data storing
8) Data lake association setting
9) Data mart generation



#	Operation	Times without platform	Times with platform	Effect of platform
1	Operation flow definition	3 months	1 month	Operation flow and data item definition by hearing sheet
2	Data item definition			Automatic data model generation with the hearing sheet
3	Data model design			
4	Operation, 4M data item registration	2 months		Automatic 4M and operation data item generation with the data model in CSV file
5	Operation flow registration	1 month	1 day	Generate operation flow registration file with the CSV file of operation and 4M data items
6	Data lake construction	1 month	1 day	New data lake can be designed by data model.
7	Experimental data storing	1 day	few minutes	Data input and data storing with GUI
8	Data lake association setting	1 month	1 day	CSV files for configuration importing
9	Data mart generation	1 day	1 hour	Data item selection and exporting

Fig. 4. Effect of operation cost reduction with experimental data management platform.

Fig. 4 shows the development items of the experimental data management infra-structure for each task, and the effects of system construction and reduction of operation man-hours realized by the experimental data management infrastructure. In this study, we developed the experimental data management infrastructure in the following three areas.

1) Operation flow Registration: Outputs business flow information based on CSV files for batch import of business and 4M data items that constitute the data model.
2) Experimental data registration: A tool for registering experimental data in the data lake using a unified data entry form.
3) Data mart generation for data analysis: Select and export data items required for data analysis using the basic functions of experimental data management platform.

With experimental data management platform, the following reductions in manhours for system construction and operation can be achieved.

1) Operation flow Registration: By preparing CSV files of operation and 4M data items, a configuration file for importing operation flow into platform can be output based on the CSV files. The manhours required to set up the system can be reduced from one month to one day.
2) Experimental data registration: If there is no input form, data registration must be requested to the system administrator, and it takes about one day. By using a data registration tool, users can register their own experimental data directly into the data lake, which takes only a few minutes.
3) Data mart generation for data analysis: By selecting the data items necessary for data analysis and executing export by platform GUI, user can select and retrieve the necessary data in about an hour's work.

The cumulative effect of these tools in reducing man-hours for system construction and operation is as follows. Operation flow Registration: Reduced one month to one day. Shortening effect 96.7%. Experimental data registration: Reduces a day (8 hours) to a few minutes (5 minutes). Shortening effect 99.0%. Data mart generation for data analysis: Reduced 1 day (8 hours) to 1 hour. Shortening effect 87.5%. The cumulative time saved is 29 days, or about 15 hours, and the average effect of the reduction is 94.4%.

5. Conclusion

In this report, we examined the experimental data management infrastructure by describing research and development work in a unified data model. We created a data model of the workflow covering all possible materials development tasks, and designed an experimental data management infrastructure to manage all possible combinations of workflows and associated experimental data in the R&D of target materials by managing the data corresponding to the actual work per-formed. We designed an experimental data management infrastructure to realize the management of all combinations of workflows and associated experimental data in the research and development of target materials. Through prototyping and evaluation of the proposed technology, it was confirmed that the MI system construction and operation man-hours could be reduced by 94.4% in the targeted tasks.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Hideya Yoshiuchi proposed Experimental Data Management Platform for Materials Informatics with advanced data search function. He developed prototype and evaluated its performance from the viewpoint of usability. Hiroki Miyamoto supervised Yoshiuchi's work with hearing customers' requirement. Sayaka Tanimoto has much knowledge about material research and gave Yoshiuchi valuable advices.

References

[1] Ramson, S. R. J., Vishnu, S., & Shanmugam, M. (2020). Applications of internet of things (IoT) — An overview. *Proceedings of the 2020 5th International Conference on Devices, Circuits and Systems (ICDCS)* (pp. 92-95).

[2] Ong, Y. S., & Gupta A. (2019). AIR5: Five pillars of artificial intelligence research. *Proceedings of the IEEE Transactions on Emerging Topics in Computational Intelligence*, *3(5)*, 411-415

[3] Rajan, K. (2005). Materials informatics. *Materials Today, 8(10)*, 38-45

[4] Tanaka, F., Sato, H., Yoshii, N., & Matsui, H. (2018). Materials informatics for process and material co-optimization. *Proceedings of the 2018 International Symposium on Semiconductor Manufacturing (ISSM)* (pp. 1-3).

[5] Yu, G., Chen, J., & Zhu, L. (2009). Data mining techniques for materials informatics: Datasets preparing and applications. *Proceedings of the 2009 Second International Symposium on Knowledge Acquisition and Modeling* (pp. 189-192).

[6] Orii, Y., Hirose, S., Toda, H., & Kobayashi, M. (2020). Development of materials informatics platform. *Proceedings of the 2020 Pan Pacific Microelectronics Symposium (Pan Pacific)* (pp. 1-5).