A Study on Improving the EDISON Data Registration Process Based on Dynamic Design

Sunggeun Han
Korea Institute of Science and Technology Information
Daejeon, Korea
sghan@kisti.re.kr

Jeongcheol Lee
Korea Institute of Science and Technology Information
Daejeon, Korea
jclee@kisti.re.kr

Hoon Choi
Chungnam National University
Daejeon, Korea
hc@cnu.ac.kr

***Abstract*—There are numerous data formats used in computational science. In order to effectively integrate and share data, improving the data registration process is crucial. In this study, we transitioned the data registration method within the EDISON platform from a static design to a dynamic design. To enhance data validation, we provided a schema authoring tool allowing data owners to design their own data schema. Additionally, we enabled support for a variety of data formats by allowing custom parsers to be written in a container-based Docker image format. Through these advancements, researchers can register and manage different data formats more efficiently. By implementing a dynamic design, data usability across various domains and environments can be improved, significantly reducing the time and cost associated with data errors while increasing the efficiency of data extraction and processing.**

Keywords—Computational Science, Data Registration, Dynamic Design, EDISON Platform, Data Schema

I. INTRODUCTION

Computational science utilizes computer calculations to analyze mathematical models, offering innovative approaches to research methodologies. This field is characterized by its significant demand for computational resources [1]. Recent advancements in computer performance and cyber infrastructure have ushered in an era of large-scale parallel computations, allowing previously unattainable simulations to be run. These simulations generate vast amounts of data. With the emergence of the data-driven research paradigm, there is a growing emphasis on deriving results from data, as well as an increasing demand for data transparency and sharing [2]. Researchers who hold or seek to utilize this large volume of data prioritize ease of data use. However, challenges persist in data processing and management, highlighting the need for an effective platform to address these issues.

EDISON (EDucation-research Integration through Simulation On the Net) is a web-based online simulation platform that provides software and educational content for computational science research [3]. Users can upload and share their data via the EDISON platform. Currently, EDISON has defined the following seven areas of computational science and provides data services for each area through a hard-coded approach: Computational Thermofluids, Nanophysics, Computational Chemistry, Structural Dynamics, Computational Design, Computational Medicine, and Urban Environment. In addition, efforts are ongoing to expand the scope of the EDISON platform from educational to research purposes. The need to integrate more disciplines beyond the original seven and the need for unified data management has led to arguments for moving away from the current hard-coded approach and introducing customized data registration by individual researchers or research areas.

This study examines the limitations of the current data registration process in the EDISON platform and discusses the design and implementation of a new registration process that allows for customized data entry across various research fields or individual researchers.

II. CONSTRAINTS OF THE CURRENT SYSTEM

The current EDISON platform's data registration process, originally conceived for designated specialized domains, has been built on a static design paradigm, utilizing a hard-coded approach. Once data is registered in the predefined format, a series of data curation activities such as data extraction and validation are carried out in accordance with the predefined data types, subsequently generating shared data. This design paradigm poses significant challenges when supporting a diverse range of data from various domains. Among these challenges, limitations in data validity verification stand out prominently.

A. Constraints in Data Validity Verification

- Static Data Type Verification: The system is confined to validation mechanisms tailored to the previously established data types, limiting its ability to flexibly verify new data types or formats.
- Absence of Data Quality Management: The system lacks the agility to detect or rectify incorrect data or anomalies once they are introduced.
- Limited Individual Data Attribute Verification: The hard-coded approach necessitates predefined validation logic for specific data attributes, negating the possibility of a flexible verification for diverse attributes.

B. Insufficiency in Supporting Varied Data Formats

- Restricted to Designated Formats: The EDISON platform caters exclusively to specific data formats, hampering the integration of source data that might come in a multitude of formats.
- Lack of Data Transformation Tools: The platform lacks automated tools or interfaces for transitioning between varied data formats, compelling users to invest significant time and effort in data conversion tasks.
- Performance Degradation Based on Format: Processing data outside the predetermined formats heightens the risk of performance declines or system errors.

Such constraints could serve as significant roadblocks in the evolution of the EDISON platform to adeptly cater to today's diverse and intricate data landscapes. Consequently, the integration of new design principles, encompassing a user-

centric interface, modularity and extensibility, and comprehensive support for diverse data formats, becomes imperative.

III. Design of the Data Registration Process

Designing a data enrollment process based on dynamic design requires the following elements:

- Data Representation - *Data Collection and Dataset:* In this research, a hierarchical data format, termed as 'data collection' and 'dataset', has been introduced to precisely represent the data owned by researchers. Here, the 'dataset' denotes a set of meaningful data extracted from multiple files, and a collection of such datasets is referred to as the 'data collection'.
- Data Storage - *Repository, DBMS, and Indexing Server:* At the heart of data storage lie the Repository, DBMS, and Indexing Server. The Repository serves as a space to store original and derived files of the researchers [4]. The DBMS plays a central role in structuring and managing this data, while the Indexing Server stores indexing information for efficient data retrieval [5].
- Data Curation: Data curation is defined as the process of managing the quality and validity of research data [6]. During this process, research outcomes are extracted from files, the validity of this data is verified, and then it's transformed as required.

Fig.1 shows the overall flow of the proposed data registration process.

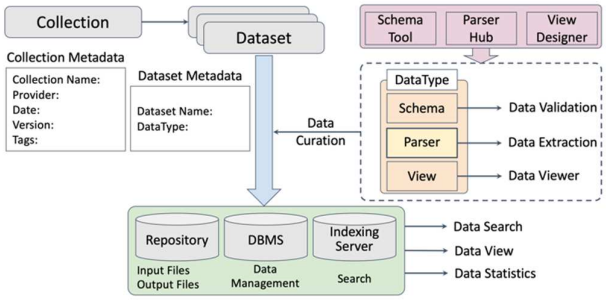

Fig. 1. Overall flow of the data registration process

The basic structure required when a researcher registers data or data files is as follows:

- Data Schema: Defines the structure and validation rules for the data. This research provides a user-friendly schema authoring tool, allowing researchers to easily define their schema.
- Data Parser: Contains the logic to extract necessary data from files. Researchers can craft this in the form of a docker image and share and use it via ParserHub.
- Data View: This determines how data will be visualized based on specific domains. Users can create their own data viewer using the View designer.

During the registration, the researcher provides collection and dataset metadata and uploads the data files. Subsequently, the platform performs curation based on the data type defined by the user and stores it appropriately in the storage system.

IV. Implementation of the Data Registration Process

As illustrated in Fig. 1, the overall flow of the data registration process has been described. In this section, we dive into the details of implementing each feature.

A. Enhancement of Data Validity Verification Function Based on User-Defined Schemas

Unlike the static design of the original EDISON platform, this study allows users to define their own data schema for data they possess. To facilitate this, a schema authoring tool is provided, ensuring user-friendliness. Fig. 2 shows the web-based schema authoring tool. Users can define valuable data from metadata and data files in the schema and impose constraints for data validation on each item. The validation constraints provided are:

- Data Type: Verifies that the data is entered in the correct format (string, number, date, etc.).
- Data Length: Validates the data's length is within an acceptable range.
- Value Range: Ensures numerical values fall between an accepted minimum and maximum value.
- Regular Expressions: Uses regular expressions to determine if the data matches a specific pattern.

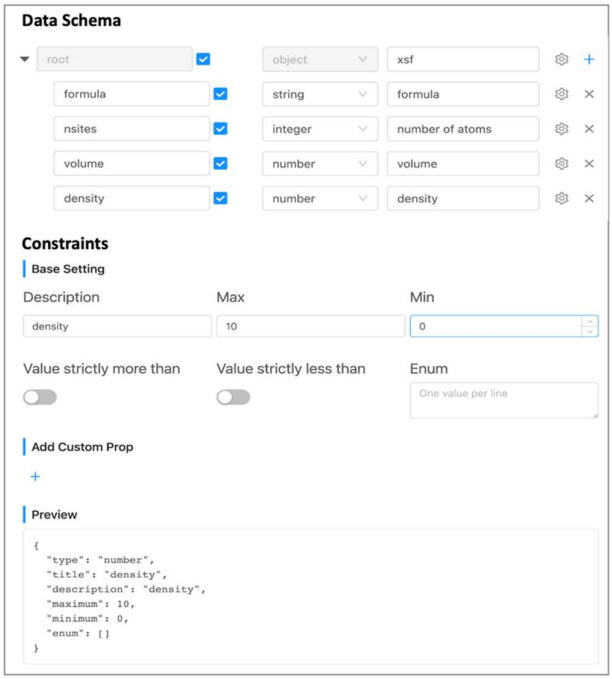

Fig. 2. The web-based schema authoring tool

User-defined schemas can be downloaded in JSON or CSV formats, making them sharable as data templates [7]. This allows you to collaborate seamlessly with other researchers or team members. Using such templates ensures data input consistency and allows for more effective standardization and integration of data. Such schema-based templates significantly reduce errors during data analysis and processing, leading to higher quality results.

The data validity verification function based on user-defined schemas significantly enhances the flexibility and efficiency of data management and validation. Especially

since users can define their own schemas and set validation constraints, the accuracy and quality of the data is elevated. The web-based schema authoring tool simplifies schema creation and management, greatly reducing data management complexity. These features increase data utilization in a variety of fields and situations, reducing the time and cost associated with data errors.

B. *Support for Various Data Formats through User-Defined Data Parsers*

The computational science field supports various data formats. Moreover, each researcher might process and possess data in new, unique formats. To extract valid data from such diverse formats, it is most accurate when researchers define and extract the data themselves. In this study, we developed a ParserHub to manage parsers in the form of container-based docker images [8]. Researchers can write a program for data extraction, convert it into a Docker image, and then register it on ParserHub. Once registered, the platform internally loads it as a container image, and it can be employed as a parser responsible for data extraction during data registration. Fig. 3 displays the operation process of the user-defined data parser.

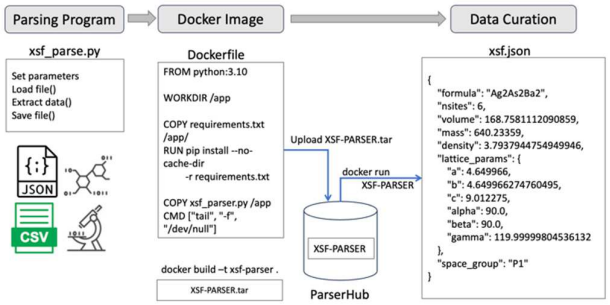

Fig. 3. The operation process of the user-defined data parser

By allowing researchers to write their own parser programs, they can precisely and effectively extract desired information from various data formats. This is instrumental in addressing the common challenge of diverse data formats in computational science. The proposed ParserHub in this study eases the management and execution of these user-defined parsers in Docker image form, greatly enhancing the efficiency of the data extraction and processing phases.

V. Conclusion and Future Research

In this study, we delved into the dynamic design-based data registration process. We conducted an in-depth analysis on the significance of data registration and how this process can be applied in research and various fields. Emphasizing on the integration of user-defined schemas and data parsers, we proposed effective ways to handle a wide range of data formats, elevating the validity and quality of the data. The introduction of data curation processes within the platform, data validity checks, and parsers have been confirmed to significantly enhance the flexibility and efficiency of data management and processing. Such an approach becomes particularly essential as we navigate through a research environment marked by increasing data diversity and complexity. The tools and methodologies proposed in this research are anticipated to aid researchers in managing and utilizing their data with greater precision and efficiency.

Future research should focus on automating data curation using machine learning and AI, and further exploring data security issues through encryption, access control, and data anonymization.

Acknowledgment

This research was supported by the EDISON Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. NRF-2022M3C1A6090416).

References

[1] A. B. Shiflet and G. W. Shiflet, "Introduction to Computational Science: Modeling and Simulation for the Sciences," Princeton University Press, 2014.

[2] C. O. Klingenberg, M. A. V. Borges, and J. A. V. Antunes Jr., "Industry 4.0 as a data-driven paradigm: a systematic literature review on technologies," Journal of Manufacturing Technology Management, vol. 32, no. 3, pp. 570-592, 2021.

[3] D.-S. Jin, Y.-J. Jung, and H.-K. Jung, "EDISON platform to supporting education and integration research in computational science," Journal of the Korea Institute of Information and Communication Engineering, vol. 16, no. 1, pp. 176-182, 2012.

[4] S. Ahn, J. Lee, J.-S. Kim, and J. R. Lee, "EDISON - DATA: A flexible and extensible platform for processing and analysis of computational science data," Software - Practice and Experience, 2019.

[5] H. Akdogan, Elasticsearch Indexing. Packt Publishing Ltd, 2015.

[6] M. McLure, A. V. Level, C. L. Cranston, B. Oehlerts, and M. Culbertson, "Data curation: A study of researcher practices and needs," Portal: Libraries and the Academy, vol. 14, no. 2, pp. 139-164, 2014.

[7] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, "Foundations of JSON schema," in Proceedings of the 25th International Conference on World Wide Web, Apr. 2016, pp. 263-273.

[8] C. Boettiger, "An introduction to Docker for reproducible research," ACM SIGOPS Operating Systems Review, vol. 49, no. 1, pp. 71-79, 2015.