International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12570-12576
© Research India Publications. http://www.ripublication.com

The Development of General-Purpose Scientific Data Repository for EDISON Platform

[1]Jae-Sung Kim, [2]Jeong-Cheol Lee, [3]Sun-Il Ahn* and [4]Sik Lee, and [5]Kum-Won Cho

[1,2,3,4,5]Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea.

**Corresponding Author*
[1,3]Orcid: 0000-0003-4191-3474, 0000-0003-0840-8783

Abstract

In the age of Big Data, we have faced on the challenges such as diversity, data reliability, and scalability issues in order to find new knowledge from data. Recent advancement in computational science leads to investigate Data-driven research methodologies as well as well-designed data repository for big data and platform, so we have developed EDISON-Scientific Data Repository (SDR) platform to store various and heterogeneous data from a variety of data sources. In this paper, we propose a database model to store heterogeneous and diverse data from various sources in one general-purpose repository and an indexing method to support further analysis. To evaluate the usefulness and functionality of EDISON-SDR platform, as a pilot study, we have stored the simulation data in the Materials field on EDISON-SDR platform and we have applied to three applications developed for searching and analyzing using stored data.

Keywords: Data Repository, Data Analysis, Computational Science, Simulation

INTRODUCTION

EDISON platform [1] is an educational research environment in which students can learn the computational science methodologies by supporting the web-based computational science engineering software. Recently, there is an increasing demand for Open Science that can increase the efficiency of research and education by opening up various data such as experiments and simulations and sharing with the communities. Open Science aims to make scientific research, data, and other products accessible to everyone. Therefore, we have been developing EDISON-SDR (Scientific Data Repository) platform to store simulation data derived from EDISON platform.

EDISON-SDR platform is a repository that can easily publish, preserve, and analyze simulation data. Sharing and reusing simulation data not only avoid duplication of calculation time and cost but also can provide a new research method to extract meaningful information through the analysis of accumulated data [2].

The characteristics of the simulation data generated by EDISON platform are as follows. First, the types of simulation data are diverse and heterogeneous. EDISON platform provides simulation software for a variety of fields including Computational Fluid Dynamics, Nano-Physics, Computational Chemistry, and Computational Medicine. Each type of data which is generated in different fields is different. In addition, each type of data generated by different simulation software in the same field is different and even in the case of same software, the type of data may also be different depending on the software version. Second, in order to analyze accumulated simulation data, additional information is needed to summarize and represent the simulation. When the data is stored in the database, most of the data only needs general information such as who the owner of this data is, when the data is published, what the title of this data is, where the data is located, and so on. However, when the user wants to search the simulation data from database, it needs more information that is specific to simulation software such as what the input parameter of this simulation is, what the output of this simulation is, and so on.

Reflecting these characteristics, EDISON-SDR platform is designed to store heterogeneous and diverse data from various sources in one general-purpose repository and to support functions for further analysis. To evaluate the usefulness and functionality of EDISON-SDR platform, as a pilot study, we have stored the simulation data in the Materials field on EDISON-SDR platform and we have applied to three applications for search and analysis. Therefore, in this paper, we describe i) how we have designed database models of EDISON-SDR platform, ii) an indexing method for further analysis and iii) three applications developed for search and analysis using accumulated data in EDISON-SDR platform.

The rest of this paper is organized as follows. Section 2 explains the related research of data repositories. Section 3 explains database models for storing simulation data and a method for indexing data stored in the database. Section 4 describes three search and analysis applications using stored data in EDISON-SDR platform. Finally, Section 5 concludes this study.

RELATED RESEARCH

There are many general-purpose data repositories including Ckan [3], Dataverse [4], and Dspace [5]. However, they have limitations in providing advanced analysis that is specific to the field. The repositories that is specific to one field include NoMad [6] and Materials Project [7] in Materials field and HepSim [8] in High Energy Physics field. However, they are less flexible and scalable to support various types of data. The sweet spot of EDISON-SDR platform is that it can i) store various types of simulation data, ii) have a flexible preprocessing framework, and iii) provide applications for further analysis.

DESIGN AND IMPLEMENTATION

EDISON-SDR platform was designed to store the data derived from EDISON platform and EDISON platform has used Liferay portal framework [9] and MySQL DBMS. Thus, EDISON-SDR platform also has used the same framework and database for integration with EDISON platform.

Database Model

There are four database models to store simulation data: Dataset, Collection, DataType, and DataTypeSchema.

Dataset model, the most important model in EDISON-SDR platform is a model for storing simulation results. In order to store each simulation result and to identify each simulation, Dublin Core [10] and Descriptive Metadata [11] information are required. Dublin Core is metadata that describes what the title of this simulation is (i.e., Title), who executed the simulation (i.e., Author), when the simulation was run (i.e., Date), and so on. Dublin core information is a common property of all simulation results and is stored in each field in Dataset model as shown in figure 1.

Descriptive Metadata refers to describing and identifying information resources. In EDISON-SDR platform, Descriptive Metadata describes the simulation result. It is extracted from the result files and stored in the database. For example, in the field of Materials, information on volume, density, number of elements, and coordinate of the material is Descriptive Metadata that can describe this simulation. In the field of Computational Fluid Dynamics, information on thickness, Umach, Cl, and Cdp can be Descriptive Metadata. As you can see in this example, Descriptive Metadata can have various names and data types depending on the field. Even in the same field, Descriptive Metadata may vary depending on the simulation software. This type of data is called unstructured data. Storing such unstructured data in the relational database is too inefficient because the attribute names are not fixed so that database model must be created everytime the new Descriptive Metadata is generated. In the example above, we have to create the model with attributes volume, density, number of elements, and coordinates for storing Materials data and to create another model with attributes thickness, Umach, Cl, and Cdp for storing CFD data. This approach is against the general-purpose nature of the EDISON-SDR platform. Therefore, as can be seen in figure 1, we designed to store the whole Descriptive Metadata in one attribute using JSON format, which is evaluated to best represent unstructured data.

Collection model is a model for representing a set of related Datasets. EDISON users could run simulations multiple times and generate many simulation results to solve one problem like when we research a certain topic. In other words, it is a model for bundling several related simulation results. EDISON-SDR platform sets the access control and license policy to the Collection so that we can manage the data in Collection unit.

DataType model is a model for classifying the types of simulation results and setting the corresponding preprocessing modules separately. Just as data types such as Integer, Double, and String in Computer Science have different processing modules, simulation results on EDISON-SDR platform also require different preprocessing modules. The preprocessing module refers to a method of extracting Descriptive Metadata from simulation results. In general, preprocessing module is mapped to the simulation software one by one because simulation results from the same software have same Descriptive Metadata.

The preprocessing modules can be defined using a code or script provided by an expert who knows the simulation software. However, the provided code could have a security concern and often has dependencies on a specific execution environment. To solve these issues, as shown in figure 2, preprocessing code and scripts are developed and executed through the docker [12]. The name of the docker image has the same name as the DataType, and the preprocessing is omitted if there is no docker image for the corresponding DataType [13].

DataTypeSchema model is a model for storing information of Descriptive Metadata for each DataType. This model stores the name of Descriptive Metadata, the description of Descriptive Metadata, data type of metadata value, maximum value, minimum value, unit, and whether it is mandatory. For example, in the field of Materials, the simulation software called VASP [14] can be a DataType, and after the preprocessing step, volume, density, number of elements, and coordinates are extracted from the simulation results as Descriptive Metadata. These metadata are stored in the DataTypeSchema model.

Descriptive Metadata Indexing

In order to retrieve the information of the simulation data through EDISON-SDR platform, it is necessary to index the data stored in the database to the search engine. The most challenging part when we designed the indexing system was that EDISON-SDR platform has had to support individual retrieval of each key-value pair of Descriptive Metadata stored in JSON form. An example of an individual retrieval is as follows. As shown in figure 1, the keys of DM 1 are 'material', 'spacegroupsymbol', 'nelements' and the values are 'Li2O', 'Fm-3m', '2.' Individual retrieval of key-value pair means that the user can search for the specific key and value. For example, when the user searches the data with the number of elements is two (i.e., nelements:2), the first record is retrieved and the second and third records are not retrieved because even though the second record has 'nelement' as a key, its value is three and the third record doesn't have 'nelement' as a key.

Liferay framework provides the automatic indexing methods using Lucene [15] library. The one thing that we have to do is decide the data type of each attribute in the database and choose the method according to that data type. For example, as shown in figure 3, Dataset model has attribute 'Title' and its value is 'Vasp Sim 1.' We know this attribute is String type so that index it with addText method. And for Number type, addNumber method is used and for Date type, addDate method is used. In the database models of EDISON-SDR platform, most attributes have only one value. Thus, we can use automatic indexing method immediately for each attribute.

However, we cannot immediately use the indexing method to Descriptive Metadata because multiple values are stored in one attribute in the form of JSON and the types of these values are various such as Number, String, and Date. Thus, it is impossible to determine which method to use among addText, addNumber, or addDate. In addition, the first parameter of three indexing methods is the name of attribute. However, unlike fixed attribute name such as title and author, various attribute names are generated depending on the simulation software so that we cannot determine the first parameter of indexing method. In summary, we do not know the value of the first parameter of addText, addNumber, and addDate methods, and we do not know which of the three methods to use.

In order to solve these issues, EDISON-SDR platform performs two stage of processing to index Descriptive Metadata. First step is JSON key-value parsing process. As can be seen in figure 3, after we retrieve the whole data from Descriptive Metadata attribute, parse each key-value pairs. In this example, parsed key-value pairs are '<key: material, value: Li2O>', '<key: spacegroupsymbol, value: Fm-3m>', '<key: nelement, value: 2>', '<key: elements, value: [Li, O]>', '<key: lattice, value: [3.25, 3.25, 3.25]>', and '<key: coordinate, value: [{…}] >.' Second step is data type detection process. We implemented the data type detection method so that when the parsed pairs are submitted, the method checks the values and find the data type dynamically. In this example, 'material', 'spacegroupsymbol', 'elements', and 'coordinate' are determined as String type so that they are indexed with addText method. 'nelements' is determined as Number type so that it is indexed with addNumber method. 'coordinate' is actually JSON object type but Liferay does not provide the method for JSON object type. Thus, we treated it as String type and used addText method.

USECASE

To test the platform works successfully, we have stored about 130,000 simulation data from four data sources in the Materials field. The data has been indexed in the way described in the above section. It makes a variety of analyzes become possible. We have developed three applications that utilize the indexed data. The first application is Data search application that allows the users to view Collection information and Datasets belonging to the Collection. The second application is Advanced search application that performs searches such as equal search and range search using Descriptive Metadata. The last application is 3-D visualization chart application that analyzes the overall tendency in the form of a three-dimensional chart. In the following sections, we describe each application in detail.

Data Search Application

It is the most basic application that can search and view data stored in EDISON-SDR platform. This application is divided into a Collection tab and a Dataset tab. In the Collection tab, the user can see the list of Collections and the user can search Collections by the title or description. Figure 4 shows the data from four data sources stored successfully. In the Dataset tab, the user can see the list of Datasets regardless of Collections. The user can also search Datasets by the title or description.

Advanced Search Application

Advanced search is an application that is used to search Datasets by using Descriptive Metadata. This application consists of three parts. The first one is the part that selects DataType. The user can search Datasets only by specific DataType. The user can also search by selecting multiple DataTypes. The second one is the query part. The user can perform equal search, range search, and/or search using Descriptive Metadata that they want to search. The third one is the result-of part that selects Descriptive Metadata that wants to see as a result.

When the user selects a DataType, the list of Descriptive Metadata belonging to that DataType is listed up in the selection box of the query part. Now that the user knows Descriptive Metadata that he/she can use, he/she can write the query according to the Lucene syntax. For example, as shown in figure 5, when the user selects a DataType named ‘vasp’, the user can see the corresponding Descriptive Metadata such as 'density', 'crystal system', 'bandgap', 'nelements', 'nsites', etc. in the selection box. When the user wants to search Datasets that have the density between 0 and 10 and crystal system of cubic, the user has to enter ‘density: [0.0 TO 10.0] AND crystalsystem: Cubic’ in the query part. If the user wants to see only density and crystal system used in the query, select the density and crystal system in the result-of part. When the user hits ‘search’, Datasets corresponding to the query are returned in table form.

3-D Visualization Chart Application

We have developed a general-purpose analysis tool which can make a 3-D visualization chart by using the Descriptive Metadata in order to help the user to recognize data trends very fast and intuitively. In addition, we also have provided an advanced search function for the Descriptive Metadata which is not a numeric data that cannot be used as parameters [16].

The use cases of 3-D visualization chart are as follows. When the user selects a DataType, a list of numeric Descriptive Metadata that belongs to that DataType is listed in the “Parameters” section as you can see in figure 6. When the user selects Descriptive Metadata as x, y, and radius, and click the "Make Chart" button, the application extracts the corresponding values from the indexed Lucene documents and draw it as a chart on the right side of the screen. At the same time, the upper side of the search bar lists a list of additional searchable Descriptive Metadata. When the user writes a query according to the Lucene query syntax, the result corresponding to the query is reflected in the chart for further analysis. For example, when the user wants to extract only element values that contain Li, the user can search for ‘elements: Li.’ When the user wants to analyze with more than two conditions, such as extracting datasets that contain Li and a range of density values from 1.0 to 5.0 as shown in figure 6, the user can search for ‘elements: Li AND density: [1.0 TO 5.0].’

CONCLUSIONS

We have developed EDISON-SDR platform for storing simulation data generated from EDISON platform, which provides simulations in various fields. We have designed general-purpose database models for various and heterogeneous simulation data. For providing integrated analysis, we had to index Descriptive Metadata to search engine. However, we could not use Liferay indexing method immediately because Descriptive Metadata was stored in JSON form. Thus, we have developed key-value parsing and data type detection method that parse the key-value of Descriptive Metadata and detect the data type of it. To evaluate the functionality of EDISON-SDR platform, we have stored about 130,000 simulation data in the Materials field. After the data was stored and indexed successfully, we have developed three applications that utilize the indexed data: Data search, Advanced search, and 3-D visualization chart.

ACKNOWLEDGEMENTS

This research was supported by both the EDISON Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. NRF-2011-0020576) and the KISTI Program (No. K-17-L01-C02).

REFERENCES

[1] J. Yu, et al., 2013, “EDISON Platform: A Software Infrastructure for Application-Domain Neutral Computational Science Simulations. Future Information Communication Technology and Applications,” Springer Netherlands, pp. 283-291.

[2] W. Joo, et al., 2016, “A Trend of Data-driven Approach for Computer Simulation,” ICONI

[3] J. Winn, 2013, “Open data and the academy: An evaluation of CKAN for research data management.”

[4] G. King, 2007, “An introduction to the Dataverse Network as an infrastructure for data sharing,” Sociological Methods & Research, Vol.36, No.2, pp.173-199.

[5] M. Smith, et al., 2003, “DSpace: An open source dynamic digital repository,” https://dspace.mit.edu/handle/1721.1/29465.

[6] N. Zacharias, et al., 2005, “The Naval Observatory Merged Astrometric Dataset (NOMAD),” AAS, 205, 4815.

[7] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek and K. A. Persson, K. A., 2013, “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation,” Apl Materials, Vol.1, No.1.

[8] S. V. Chekanov, 2015, “HepSim: a repository with predictions for high-energy physics experiments," Advances in High Energy Physics.

[9] Liferay Inc, http://www.liferay.com/digital-experience-platform

[10] S. Weibel, J. Kunze, C. Lagoze, M. Wolf, 1998, “Dublin Core Metadata for Resource Discovery,” No. RFC 2413.

[11] J. R. Ahronheim, 1998, “Descriptive metadata: Emerging standards,” The journal of Academic

Librarianship, Vol.24, Issue 5, pp.395-403.

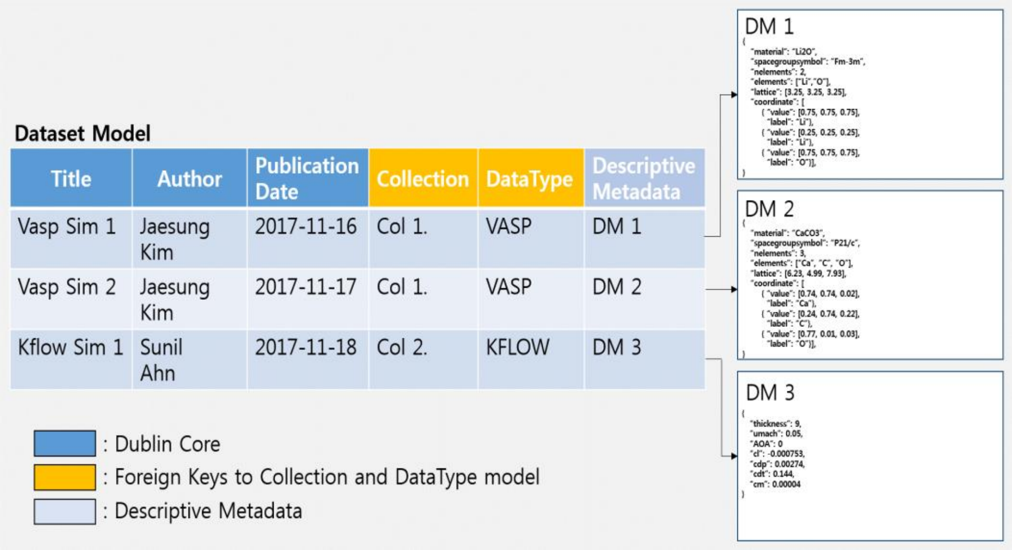
[12] D. Merkel, 2014, “Docker: lightweight linux containers for consistent development and deployment,” Linux Journal Vol. 239, No. 2.

[13] Sunil Ahn, 2017, “Data Quality Assurance for the Simulation Data Analysis in the EDISON-SDR,” The Convergent Research Society Among Humanities, Sociology, Science, and Technology.

[14] G. Kresse and D. Joubert, 1999, Phys. Rev. 59 , 1758.

[15] Lucene, https://lucene.apache.org/

[16] J. Kim, et al., 2017, “The Development of General-Purpose 3-D Visualization Analyzer for Big Data Repository,” The 9th International Conference on Computer Science and its Applications.



Dataset Model

Title	Author	Publication Date	Collection	DataType	Descriptive Metadata
Vasp Sim 1	Jaesung Kim	2017-11-16	Col 1.	VASP	DM 1
Vasp Sim 2	Jaesung Kim	2017-11-17	Col 1.	VASP	DM 2
Kflow Sim 1	Sunil Ahn	2017-11-18	Col 2.	KFLOW	DM 3

🟦 : Dublin Core
🟨 : Foreign Keys to Collection and DataType model
▭ : Descriptive Metadata

DM 1

```
{
  "material": "Li2O",
  "spacegroupsymbol": "Fm-3m",
  "nelements": 2,
  "elements": ["Li","O"],
  "lattice": [3.25, 3.25, 3.25],
  "coordinate": [
    { "value": [0.75, 0.75, 0.75],
      "label": "Li"},
    { "value": [0.25, 0.25, 0.25],
      "label": "Li"},
    { "value": [0.75, 0.75, 0.75],
      "label": "O"}],
}
```

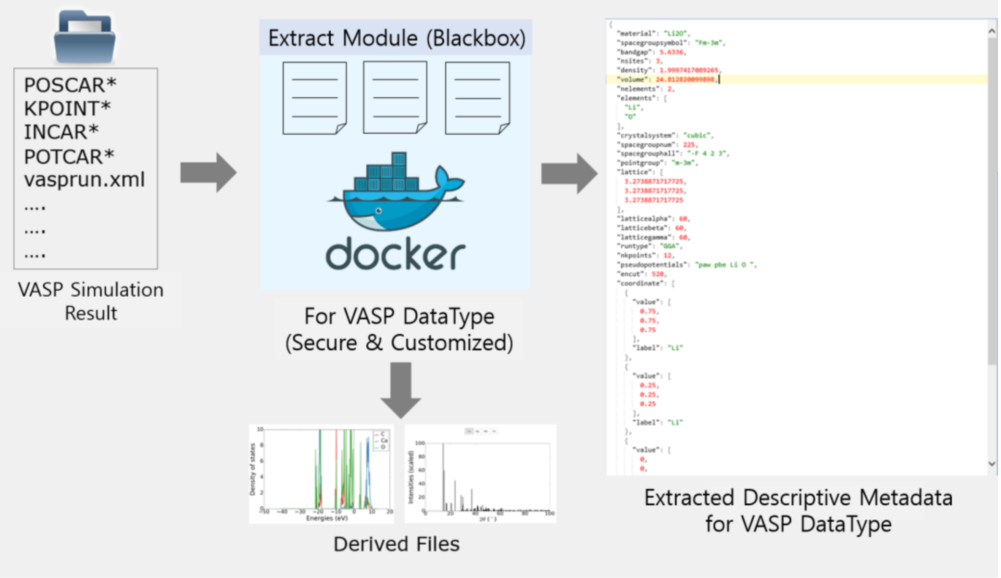
DM 2

```
{
  "material": "CaCO3",
  "spacegroupsymbol": "P21/c",
  "nelements": 3,
  "elements": ["Ca", "C", "O"],
  "lattice": [6.23, 4.99, 7.93],
  "coordinate": [
    { "value": [0.74, 0.74, 0.02],
      "label": "Ca"},
    { "value": [0.24, 0.74, 0.22],
      "label": "C"},
    { "value": [0.77, 0.01, 0.03],
      "label": "O"}],
}
```

DM 3

```
{
  "thickness": 9,
  "umach": 0.05,
  "AOA": 0
  "cl": -0.000753,
  "cdp": 0.00274,
  "cdt": 0.144,
  "cm": 0.00004
}
```

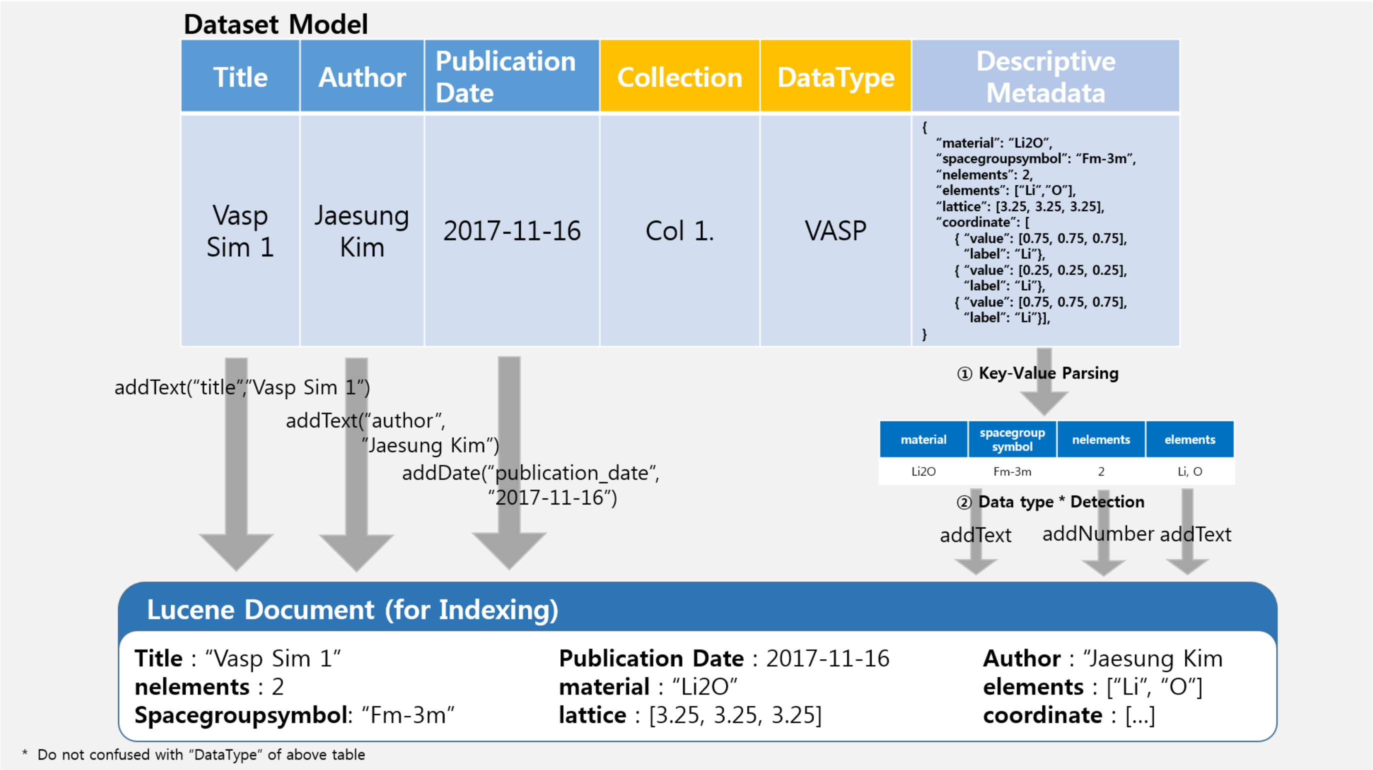
Figure 1: The example records of Dataset Model



POSCAR*
KPOINT*
INCAR*
POTCAR*
vasprun.xml
....
....
....
VASP Simulation Result
Extract Module (Blackbox)
docker
For VASP DataType (Secure & Customized)
Derived Files
Extracted Descriptive Metadata for VASP DataType

```
{
  "material": "Li2O",
  "spacegroupsymbol": "Fm-3m",
  "bandgap": 5.6336,
  "nsites": 3,
  "density": 1.9997417089265,
  "volume": 24.812820099898,
  "nelements": 2,
  "elements": [
    "Li",
    "O"
  ],
  "crystalsystem": "cubic",
  "spacegroupnum": 225,
  "spacegrouphall": "-F 4 2 3",
  "pointgroup": "m-3m",
  "lattice": [
    3.2738871717725,
    3.2738871717725,
    3.2738871717725
  ],
  "latticealpha": 60,
  "latticebeta": 60,
  "latticegamma": 60,
  "runtype": "GGA",
  "nkpoints": 12,
  "pseudopotentials": "paw pbe Li O ",
  "encut": 520,
  "coordinate": [
    {
      "value": [
        0.75,
        0.75,
        0.75
      ],
      "label": "Li"
    },
    {
      "value": [
        0.25,
        0.25,
        0.25
      ],
      "label": "Li"
    },
    {
      "value": [
        0,
        0,
```

Figure 2: The preprocessing example for 'VASP' DataType [13]



Dataset Model

Title	Author	Publication Date	Collection	DataType	Descriptive Metadata
Vasp Sim 1	Jaesung Kim	2017-11-16	Col 1.	VASP	{ “material”: “Li2O”, “spacegroupsymbol”: “Fm-3m”, “nelements”: 2, “elements”: [“Li”,”O”], “lattice”: [3.25, 3.25, 3.25], “coordinate”: [{ “value”: [0.75, 0.75, 0.75], “label”: “Li”}, { “value”: [0.25, 0.25, 0.25], “label”: “Li”}, { “value”: [0.75, 0.75, 0.75], “label”: “Li”}], }

addText(“title”,”Vasp Sim 1”)

addText(“author”, ”Jaesung Kim”)

addDate(“publication_date”, “2017-11-16”)

① Key-Value Parsing

material	spacegroup symbol	nelements	elements
Li2O	Fm-3m	2	Li, O

② Data type * Detection

addText addNumber addText

Lucene Document (for Indexing)

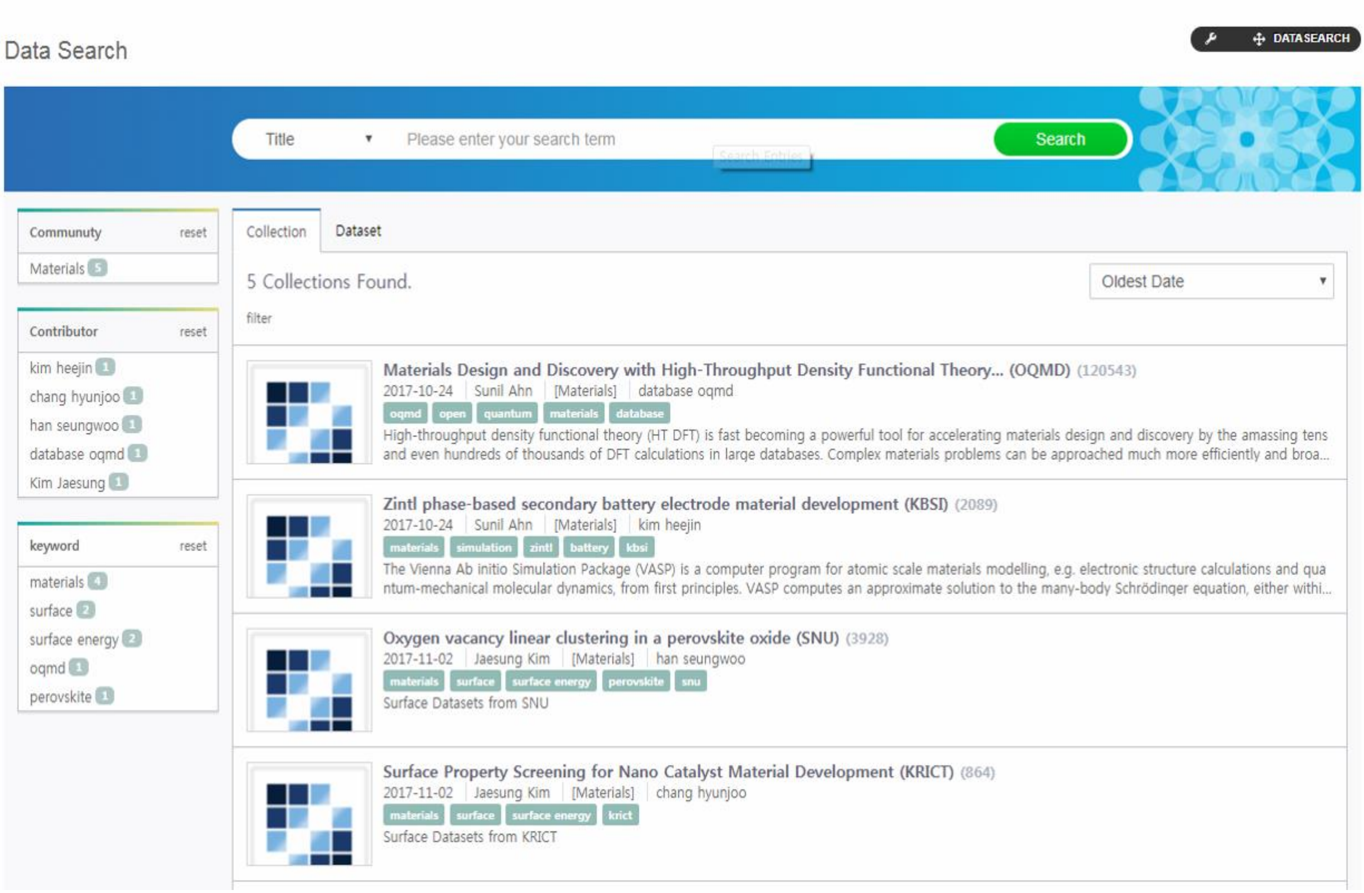
Title : “Vasp Sim 1”
nelements : 2
Spacegroupsymbol: “Fm-3m”

Publication Date : 2017-11-16
material : “Li2O”
lattice : [3.25, 3.25, 3.25]

Author : ”Jaesung Kim
elements : [“Li”, “O”]
coordinate : [...]

* Do not confused with “DataType” of above table

Figure 3: The indexing method for Dublin Core and Descriptive Metadata



5 Collections Found.

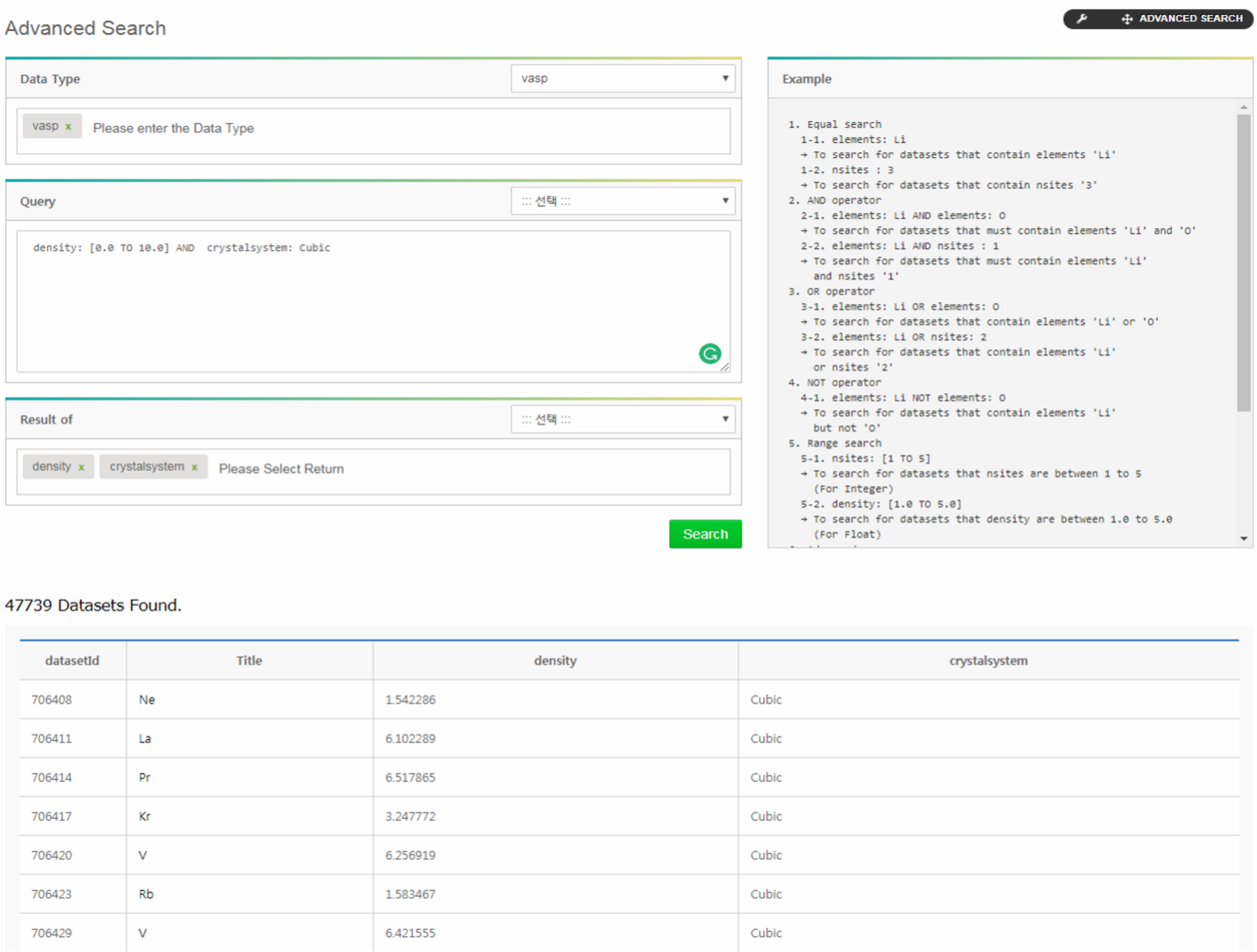
Materials Design and Discovery with High-Throughput Density Functional Theory... (OQMD) (120543)
2017-10-24 | Sunil Ahn | [Materials] | database oqmd
oqmd open quantum materials database
High-throughput density functional theory (HT DFT) is fast becoming a powerful tool for accelerating materials design and discovery by the amassing tens and even hundreds of thousands of DFT calculations in large databases. Complex materials problems can be approached much more efficiently and broa...

Zintl phase-based secondary battery electrode material development (KBSI) (2089)
2017-10-24 | Sunil Ahn | [Materials] | kim heejin
materials simulation zintl battery kbsi
The Vienna Ab initio Simulation Package (VASP) is a computer program for atomic scale materials modelling, e.g. electronic structure calculations and qua ntum-mechanical molecular dynamics, from first principles. VASP computes an approximate solution to the many-body Schrödinger equation, either withi...

Oxygen vacancy linear clustering in a perovskite oxide (SNU) (3928)
2017-11-02 | Jaesung Kim | [Materials] | han seungwoo
materials surface surface energy perovskite snu
Surface Datasets from SNU

Surface Property Screening for Nano Catalyst Material Development (KRICT) (864)
2017-11-02 | Jaesung Kim | [Materials] | chang hyunjoo
materials surface surface energy krict
Surface Datasets from KRICT

Figure 4: The example of Data search application for Materials Simulation Data



Advanced Search

Data Type: vasp

vasp x Please enter the Data Type

Query: ::: 선택 :::

```
density: [0.0 TO 10.0] AND  crystalsystem: Cubic
```

Result of: ::: 선택 :::

density x crystalsystem x Please Select Return

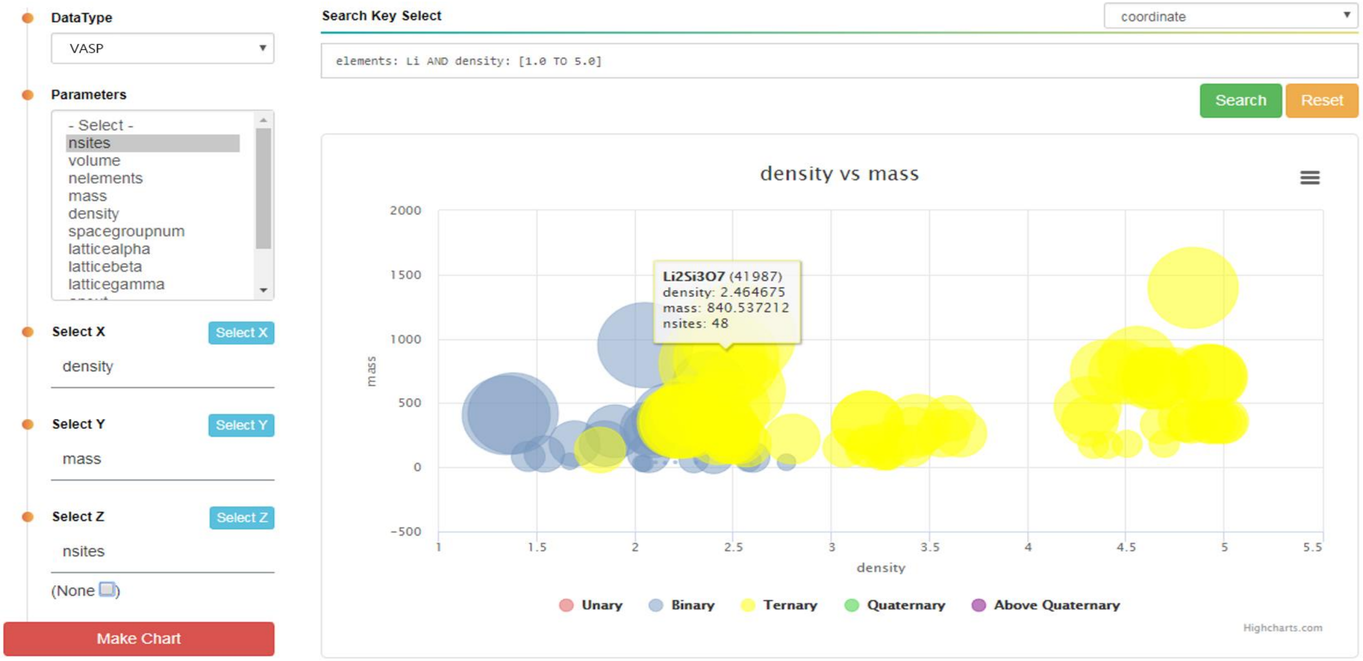
Example

```
1. Equal search
  1-1. elements: Li
  → To search for datasets that contain elements 'Li'
  1-2. nsites : 3
  → To search for datasets that contain nsites '3'
2. AND operator
  2-1. elements: Li AND elements: O
  → To search for datasets that must contain elements 'Li' and 'O'
  2-2. elements: Li AND nsites : 1
  → To search for datasets that must contain elements 'Li'
    and nsites '1'
3. OR operator
  3-1. elements: Li OR elements: O
  → To search for datasets that contain elements 'Li' or 'O'
  3-2. elements: Li OR nsites: 2
  → To search for datasets that contain elements 'Li'
    or nsites '2'
4. NOT operator
  4-1. elements: Li NOT elements: O
  → To search for datasets that contain elements 'Li'
    but not 'O'
5. Range search
  5-1. nsites: [1 TO 5]
  → To search for datasets that nsites are between 1 to 5
    (For Integer)
  5-2. density: [1.0 TO 5.0]
  → To search for datasets that density are between 1.0 to 5.0
    (For Float)
```

47739 Datasets Found.

datasetId	Title	density	crystalsystem
706408	Ne	1.542286	Cubic
706411	La	6.102289	Cubic
706414	Pr	6.517865	Cubic
706417	Kr	3.247772	Cubic
706420	V	6.256919	Cubic
706423	Rb	1.583467	Cubic
706429	V	6.421555	Cubic

Figure 5: The example of Advanced search application for Materials Simulation Data



DataType: VASP

Parameters

- Select -
nsites
volume
nelements
mass
density
spacegroupnum
latticealpha
latticebeta
latticegamma

Select X: density

Select Y: mass

Select Z: nsites

(None ☐)

Search Key Select: coordinate

```
elements: Li AND density: [1.0 TO 5.0]
```

Figure 6: The example of 3-D Visualization chart for Materials Simulation Data