International Journal of Institutional Research and Management
International Institute of Applied Informatics
2019, Vol. 3, No. 2, 50 – 61

Concept of a Unified Research Management System and its Application to Data Clustering

Hiroshi Masui [*], Kei Kikuchi [*],
Ren Kurose [*], Xun Shao [*]

Abstract

We propose a concept of a unified research management system (URMS) by utilizing services of cloud computing for the purpose of promoting the open science. In our concept, the important point of URMS is to combine different could services and to provide the services with an optimized combination for researchers according to the concerned scientific activity. We show the total concept and applications of URMS. Further, in order to illustrate our scenario for promoting the open science on URMS, we show an actual method for data clustering and discuss how the result from the clustering gives an impact to the promotion of the open science.

Keywords: Cloud computing, Open Science, Open Research Data, Cloud Orchestration, Data Clustering.

1 Introduction

The open science is one of the important trends for promoting the science and technology [1]. With the help of recent rapid progress of the network and the cloud computing, scientific researchers can share the research activities through the network. The aim of the open science is to disclose the scientific progress by sharing the outcomes from the research activities: experimentally observed data, bibliographical data of the published papers, computational codes, etc. There are two main subjects on the open science, which are called the "open access journal" and the "open research data". The former one is a movement of the distribution of published papers. Basically, the published papers are to be purchased for reading. The open access journal gives an opportunity of reading papers to researches for free. Pre-print article archives and the management of the institutional repository can be considered as one of the activities for the open access journal. The latter one is to disclose and to share the data form scientific researches and the process itself. Researchers who are interested in the subject can use the shared data, and

[*] Kitami Institute of Technology, Hokkaido, Japan

also the process of the research activity can be verified with the annotations and signatures in the shared data.

As an effort of the open science in Japan, JST (Japan Science and Technology Agency) promotes a platform of the open access journal, which is called "J-STAGE [2]." JST provides the experience how to manage the copyright and technical procedure for publish on-line. As another example, Open Science Framework (OSF) [3] is a toolkit for sharing the research activities. The platform is distributed as the open source, and researchers can manage the system by themselves. Recently, based on the OSF, National Institute of Informatics (NII) has developed the open science platform, which is called as "GakuNin RDM [4]." Also, there is another web service named as "Mendeley [5]", which is based on the viewpoint for integrating activities concerned with scientific articles. For this purpose, Mendeley has functions to manage literature information and to form a network of researchers.

Considering the above background, we propose a common research environment for scientific collaborations by combining the services thought the network [6][7]. We define the environment as the "Unified Research Management System (URMS)". In the URMS concept, the research environment is constructed by combining computational services on the premises and/or on the cloud services. The researchers who participate a collaboration on URMS share the activities through the network. URMS is designed to connect the services through the API providing the necessary functions for the scientific activities. Hence, URMS has a function of the portal site for the collaboration and can be considered as a middleware to interconnect the cloud services. In this work, we show a concept of URMS to construct the environment of the scientific collaborations through the API connections. To clarify the implementations of applications on URMS, several examples for the collaboration, e.g. unification of search queries, a file-sharing procedure and a project management will be discussed.

As a specific application on URMS to the open science, we introduce the data compilation activity of nuclear reaction data. We focus on the usage of the complied data of the nuclear reaction database on URMS and show a new kind of data clustering method. The clustering of the data is performed to find a new knowledge form the analysis of the correlation of data. We discuss a possibility for improving URMS with the result of data clustering.

In Section 2, we briefly show the concept of URMS. In Section 3, examples of implementation on URMS are shown. In Section 4, the data clustering using the nuclear reaction data is discussed. The summary and discussion are given in Section 5.

2 Concept of the Unified Research Management System

First of all, we consider here the benefit to use the cloud computing for constructing the research environment. One of the most important feature of the cloud computing is flexibility on system integration. We can modify the capacity of servers and storages even after the services are in-use. From a practical point of view for maintaining systems working on the cloud computing, it is important to consider the installation and management cost of the system under the limitation of the budget of the project. Furthermore, it is also necessary to consider the portability of the system. When the project has been finished and will move to the next stage, sometimes the research system may be required to change the platform. If the system is designed to suit a specific system, to

change the platform might be difficult. Therefore, the portability becomes an important feature of the research management system in the design of the system on the cloud computing.

In this work, we propose a concept for constructing a research environment by combining various services, which are provided as the cloud computing and on the premise systems if necessary. The environment is required to be designed so as to maximize the researcher's demand and to minimize the cost and efforts in order to manage the system. Under this concept, we consider the necessary functions of the environment are categorized as follows:

1) Data sharing
2) Project management
3) Administration of research groups
4) Communication
5) Tools for research

The important point of our concept is to unify the applications and cloud services to provide the above functions instead of using the services individually. Further, the combination of the services is optimized according to the research purpose and is provided to researchers as if the combined service is directly provided form the environment. We call this environment as the "Unified Research Management System (URMS)". Practical implements for unifying the services depend on the network traffic and connections between the servers and clients. We will mention this point later.

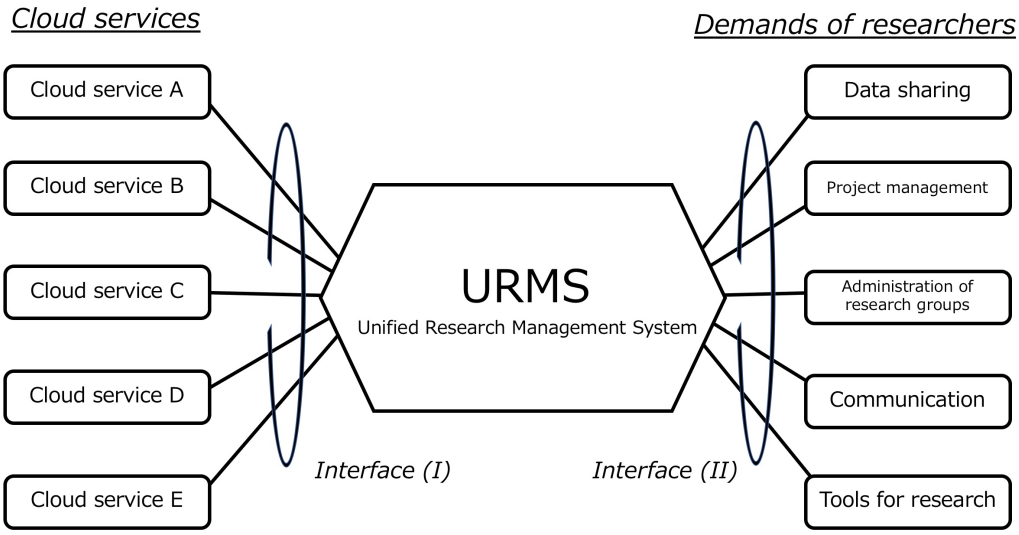

Figure 1: A schematic figure of the Unified Research Management System (URMS)

A schematic figure for the concept of URMS is shown in Figure 1. The interface (I) is the symbolic sum of the connections from URMS to the cloud services. The interface (II)

shows the demands of researchers, which are the necessary functions for the research environment as listed in the previous paragraph. The optimization of the services means that the functions on the interface (II) are provided by combining the cloud services on the interface (I). The weights of the separation of the services are calculated so as to maximize or minimize the gain of the objective function, which is defined on URMS. This concept can be considered to be similar to the "cloud orchestration" and the "cloud brokerage [8]."

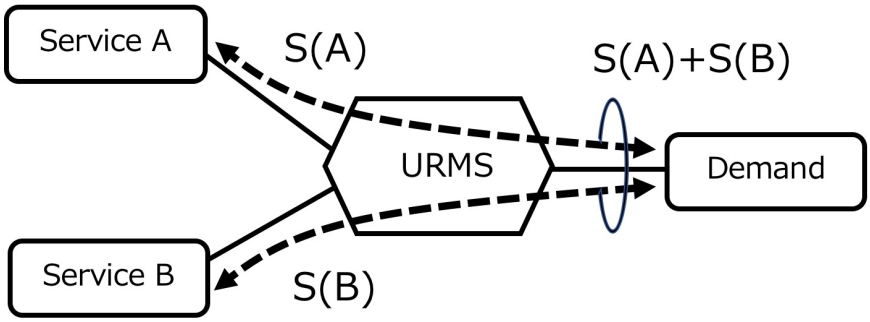

Figure 2: Optimization of services for the user demand

An illustrative example of the cloud orchestration is the combination of the cloud storages. We define the objective function by parametrizing the features of the cloud services such as the cost, data transfer time, response and reliability and, URMS maximizes or minimizes the objective function under the user's condition.

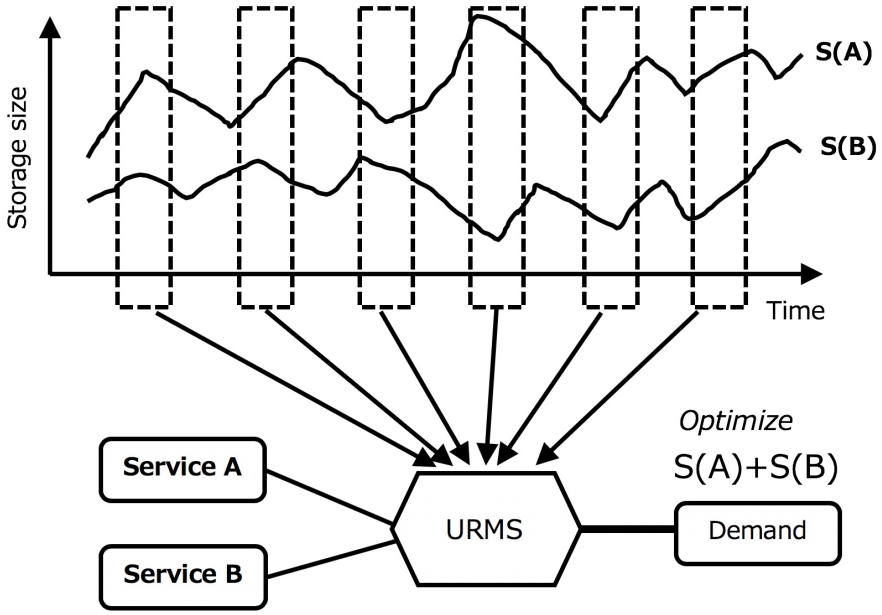

Figure 3: Time dependence of the optimization on URMS

The essential difference between the general cloud orchestration and our URMS concept is the process for defining the objective function. This is because the research demands, for example discussions between researchers and a way of data sharing are not quantifiable. Therefore, we need to "translate" the demands into some quantifiable indicators. Figure 2 shows an idea of the combination of the services S(A) and S(B) to provide a function for the demand. After the translation to the quantifiable values, we can optimize the research demands and define the weights of the separation of the cloud services. Furthermore, the optimization should have the time dependence as the project is ongoing. In Figure 3, we show a schematic figure by focusing on the time-dependent optimization.

Though our concept of URMS is similar to OFS[4] and Mendeley[5] as mentioned in Intro-duction, the characteristic point of this paper is to focus on combining various demands for proceeding the research activities and providing users with an integrated services, which are automatically optimized by the system with defining an objective function for the demands and services. Therefore, the implementation on the service providing to researchers might be similar to that of OSM and Mendeley, but the point that the combination of services is performed by URMS, and the researches are not necessary to optimize the services by themselves, is the essential difference from OSF and Mendeley. To reiterate, the main idea is summarized in Figure 1 as a schematic figure. The concept of URMS is to optimize the combination of cloud services by Interface (I) and the demands in research activities by Interface (II), and this is the main assertion of this work.

In our concept, we consider that URMS can be managed by each research group. Therefore, the management of user ID and the interconnect of different URMS are also an important issues. So far, it is necessary to maintain the user ID as the local ID for each URMS. However, for the authentication of users, to use the Shibboleth [9] would be one of a useful solution. In this respect, the GakuNin RDM realize the user ID management by applying the Shibboleth certification, which is called the GakuNin federation [10].

3 Applications on URMS

In this section, we show a role model of the promotion of the open science on URMS. We assume the role model in the following points. First, we focus on the data sharing of the scientific database, more specifically, as an example, we apply the model to the nuclear reaction database. Second, we show a scenario for using a useful tool on URMS and how the promotion of the open science can be expected.

3.1 Unified HTML Response

In the research activities, the researchers may have a situation that they need to send multiple queries for different scientific databases, and the results are send back to the web browser on different pages. We consider the necessary function of URMS on this situation is to unify these results on different pages into one rearranged page by extracting the essential part in the results concerned with the query.

In the general case, a web site of a data search is constructed by the combination of the front page as the interface to users and the CGI to send a query to the search engine. Hence, the in-terface part can be separated to the CGI part, and the different front pages can be unified to one page with including multiple queries to the different search engines.

Also, the responses from the different search engines are decomposed using an HTML parser and are re-constructed as a unified search result page.

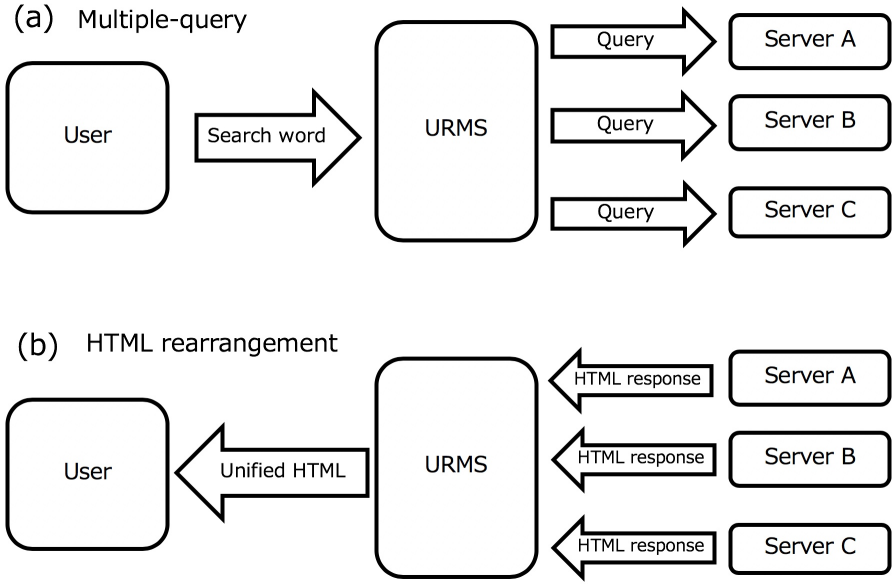

Figure 4: Unified HTML response on URMS: (a) to send multiple-query (b) unification

Therefore, we design the unified HTML response tool as follows. First, we prepare the HTML connection part to the search engines of the scientific databases. Since the form of the query might be different and depend on the databases, the interface should be adjusted to the actual implementation of the search engines. Second, to unify the HTML responses returned form the search engines, we use an application of the HTML parser and extract the necessary information from the HTML response. Last, to provide the unified information from the HTML response, URMS gives the result page, in which all the search results are contained. The general picture of the tool on URMS is shown in Figure 4.

3.2 Data Sharing

For the function of the data sharing on URMS, we consider two parts, the interface (I), URMS to the cloud services and the interface (II). In Figure 5, we show the schematic figure of the data sharing on URMS using different services among researchers: User 1 to 4. For the interface (I), the connections are spanned to various services, such as the public cloud services and also on the premise services. In some cases, a local storage would be connected to URMS. Under the URMS concept, the interface (I) is optimized to maximize or minimize the objective function, which simulates the user demand.

The optimization of the storages is done in the same manner of the combination of services, which is shown in the previous section. For the interface (II), the users are connected to URMS through a file-sharing service. The important point of this connection is that the users do not connect each other through the file-sharing, and only the connection to the URMS should be established.

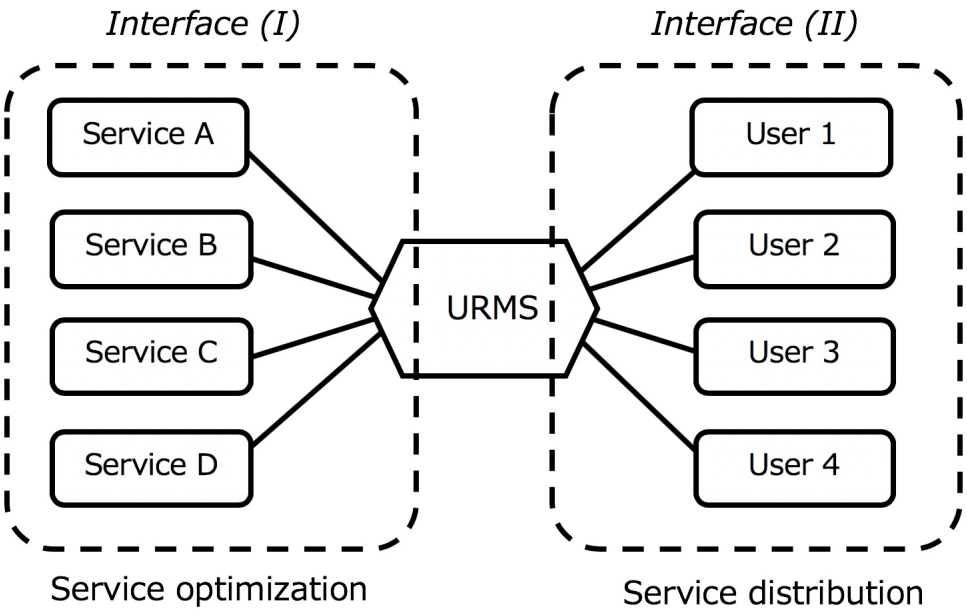

Figure 5: Service optimization and service distribution on URMS

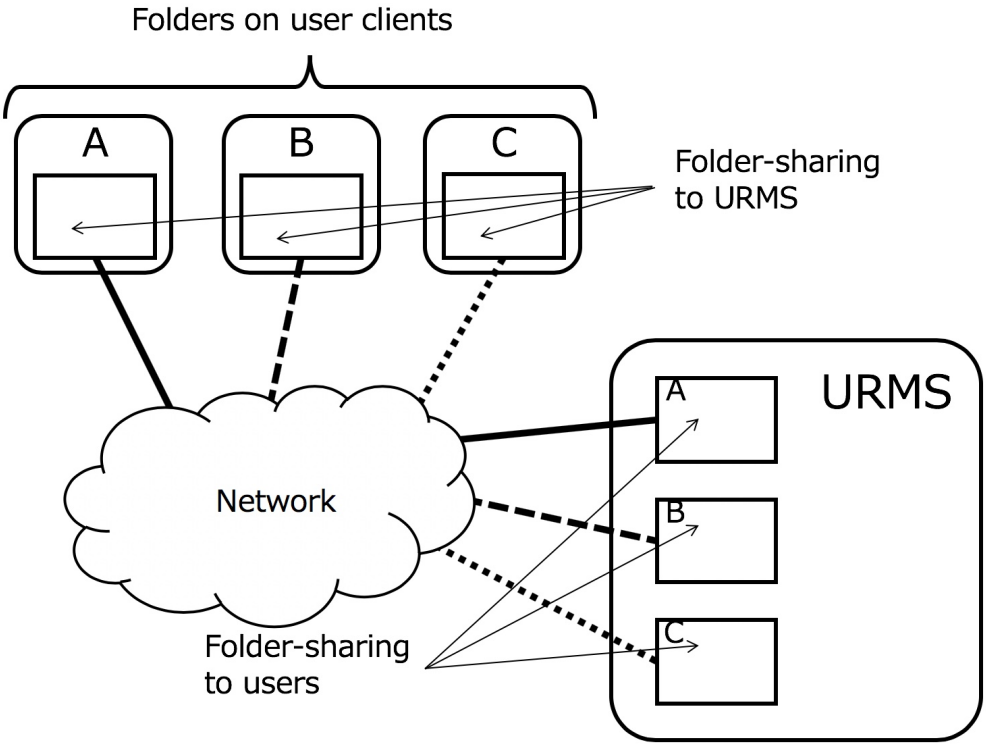

Figure 6: File sharing on URMS through the cloud services

Figure 6 shows the management of the file sharing on URMS. The permissions of the shared files are properly managed on the management section on URMS. For example, the User A can access only to the published data, the data management section on URMS, the lower right part of Figure 6, removes files from the folder of User A in the management section. Furthermore, in the case that the users apply different folder-sharing services, URMS connects folders with users individually and holds the connectivity under such the condition.

3.3 Project Management

For the project management, we consider to combine two different kinds of services, a document storage service such as Evernote and a file storage service such as Dropbox. In the project management, we need to manage the progress of the project and to store the deliverables relating to the progress of the project, see Figure 7. Using the functions on the document storage services, we define the IDs concerned with the progress of the project and put the information to the services through the API. When we change the services or combine other services for this purpose, we modify the API and connect to the services. Information of the connections between the progress of the project and the deliverables can be stored in the document storage services or stored in the local file systems on URMS.

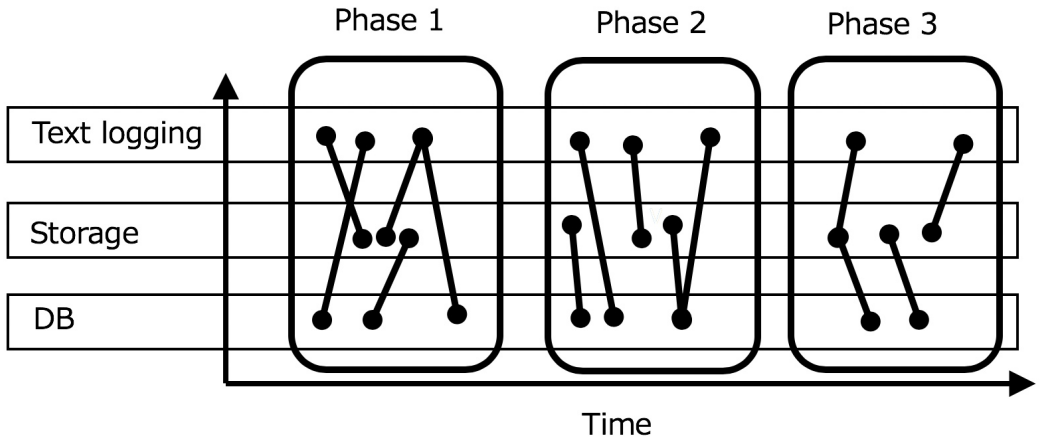

Figure 7: Combination of different services for the project management on URMS

In general, the project management can be classified into two different types. One is the waterfall model and the other is the agile one. The general management model is the waterfall one. The essential point of the waterfall model is that the progress of the development of the project goes the one-way progress from the opening of the project to the closing. The assignment of tasks and persons for each step can be defined and beforehand the project going on. On the other hand, the agile model includes sub-processes, which may repeat several times until the sub-process is completed. Basically, we consider URMS is designed to manage the waterfall model and will extend to apply to the agile model as the future work.

The above three functions are the brief example of the concept of URMS. In the next session, we would like to discuss how the development of URMS gives an impact to the

research field of open science.

4 Data Clustering and Development on URMS

In this section, we show a role model of the promotion of the open science on URMS. We assume the role model in the following points of view. First, we focus on the data sharing of the scientific database, more specifically, as an example, we apply the model to the nuclear reaction database. Second, we show a scenario for using a useful tool on URMS and how the promotion of the open science can be expected.

Before we proceed to show a model, we briefly introduce the nuclear reaction database. The Nuclear Reaction Data Centers (NRDC) in the International Atomic Nuclear Agency (IAEA) has accumulated the nuclear reaction data more than 50 years. The format of the data file is called EXFOR (Exchange FORmat) [11]. As the contribution to EXFOR from Japan, the nuclear reaction data centre (JCPRG) [12] has been compiling nuclear reaction data from published papers, in which experiments are done with the Japanese accelerators. The data files are called as the Nuclear Reaction Data File (NRDF). Most of data files in NRDF are translated to EXFOR. EXFOR and NRDF data files are opened through the website [11][12]. In the data file, the bibliographical data, experimental situation, numerical values of the experimental observations are contained. In the viewpoint of the data distribution and sharing, the nuclear reaction data can be considered as one of the actual application of the open science through the data sharing.

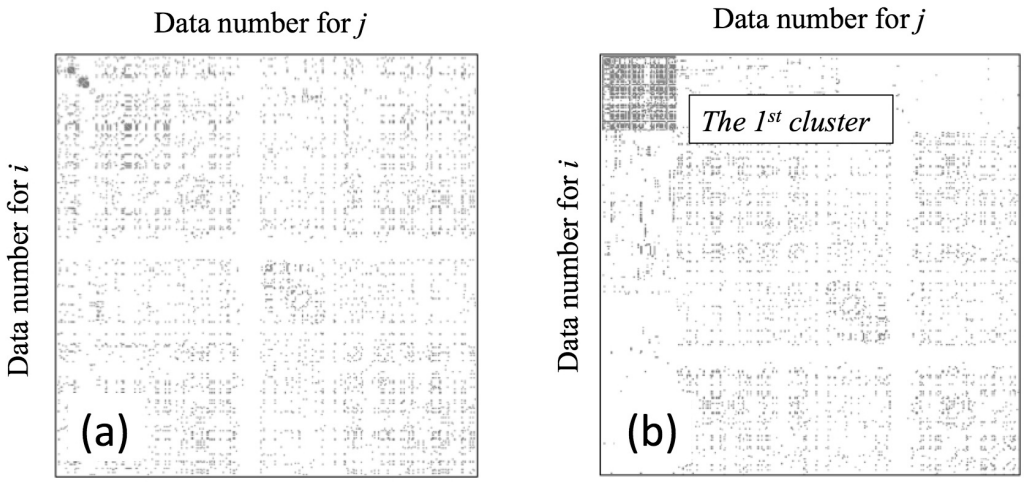

Figure 8: (a) Pair-comparison of the data and (b) the clustering for the first one

For the first point, we consider a situation that researchers share the datafile on URMS. Using the function B explained in the previous section, the nuclear reaction data centers can share the data files on URMS. For the second point, we provide a tool of the scientific analysis for the compiled data, which is related to the function A also in the previous section. We propose a new tool for the scientific analysis, which is concerned with the data clustering. Generally, the data clustering is applied to define groups for the data refereeing the similarity of data. The scenario we consider here is to find a new subject in the

experimental and theoretical research from the knowledge found by the data and tools in URMS.

We show a brief process of our new data clustering, and detailed explanation will be shown in elsewhere. The procedure of our data clustering is summarized as follows:

1) Transform qualitative data into quantitative data

2) Define the correlation of data from the overlap

3) Clustering data from the correlation

In the first step, we transform qualitative data, i.e. bibliographical data in NRDF, into the quantitative data through comparison of data files with defining the overlap between the data. For example, to define the overlap of the data in terms of the authors in each data, we count the total number of authors and the number of identical authors in among the data. The overlap cab be deduced from the numbers of the total and identical authors.

In the second step, we calculate the "similarity" of data. The NRDF data to be clustered here are scientific multi-dimensional data. Authors, author's institutions, accelerator facilities, detectors, incident particle and the mass number, target nucleus and the mass number, incident energy, etc. are included. Data clustering is performed in order to find similar data sets regarding this multidimensional data, which can be considered as vectors in multidimensional space. Therefore, we define the similarity of data using the cosine similarity for the vectors, since the clustering in this case corresponds to grouping the vectors with many attributes in near direction. The clustering is performed for the pair that the similarity is defined as a numerical value. Therefore, the essence of the clustering method is unchanged even if the numerical similarity is expressed by another method. The cosine similarity is the one of the most popular tool to calculate the similarity of the data. The Euclid distance in the multi-dimensional data space is also a tool for estimating the similarity.

The final step is to make clusters according to the similarity. For this purpose, we show a new method for data clustering, which takes following steps:

i) Define the threshold for the similarity

ii) Make a pair-comparison table for all data

iii) Find the most correlated data

iv) Sort the data in pair-comparison table

v) Repeat iii)-iv) until the condition fulfilled

In the step i), we define a threshold value, d_{th}. If the similarity between the data i and j is larger than the threshold, i.e. $d_{ij} > d_{th}$, these two data can be considered as a "correlated" pair. We calculate the mutual similarity for all the pairs and make a pair-comparison table as addressed in the step ii). Figure 8(a) shows an illustrative example for the pair-comparison table taken from NRDF. The dots show the "correlated" pairs, and blanks are the "uncorrelated" pairs. In the step iii), we find the most correlated data, which have the largest number of the correlated pairs in the pair-comparison table under the assumption

that the most correlated data will be the center of a cluster. In the step iv), we put the most correlated data on the first row and column in the pair-comparison table. For the second and after rows and columns, we sort them so that the "correlated" pairs are continued until a blank pair appears as shown in Figure 8(b). We repeat the steps iii) and iv) until the condition for the clustering is satisfied.

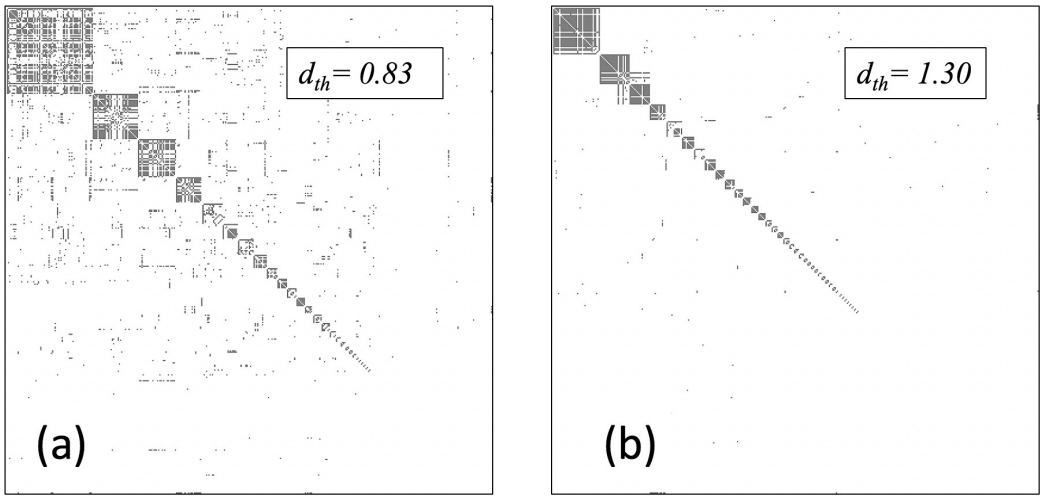

Figure 9: Clustering with different threshold values: (a) d_{th}=*0.83* and (b) d_{th} = *1.30*

Figures 9(a) and 9(b) show the examples of the clustering for the threshold as d_{th} = *0.83* and *1.30*. The off-diagonal elements between the clusters correspond to the overlap of the clusters. As seen from the Figures 9(a) and 9(b), the clusters for the d_{th} = *1.30* case have smaller overlaps than that of the d_{th} = *0.83* case. We can choose the threshold parameter to separate the clusters sufficiently by referring the sorted pair-comparison table.

From the analysis of the data clustering, we expect a feedback to the implementation of URMS for the following points. First, the status of the important data should be added as the property of the data compilation and the management process on URMS. Second, in the data search process on URMS, the new knowledge from the data clustering is shown as a "suggestion" in the search query.

5 Summary and Discussion

We proposed a concept for the research environment with combining the cloud services to promote the open science. We call this environment under the concept as the "Unified Research Management System (URMS)." The essential point of URMS is to connect the several services, for which not only the public cloud services, but also the on the premise implementations and to provide unified functions to the researchers with optimizing the services. The optimization should be defined so as to maximize the researchers demands and also to minimize the other items, e.g. the total cost, the latency of the data storage and so on. Since the functions are implemented as a trial so far, the actual implementation with optimizing the user demands must be necessary.

Under the concept of URMS, we apply a new approach for data clustering to the nuclear reaction data, NRDF and show a role model for promoting the open science. From the analysis of the data clusters on URMS, we expect a useful suggestion for the compilation plan and for the research activities will be given. In this analysis, we restrict ourselves to small number of data, at about 400. Therefore, more comprehensive analysis is to be performed as in the future work.

Acknowledgement

The authors would like to thank the member of the Information Technology for Nuclear Science Laboratory at Kitami Institute of Technology. One to the author, H.M. acknowledges the support by JSPS KAKENHI Grant No.18K03636. This work is partly supported by the collaborative research program 2019 information initiative center, Hokkaido University, Sapporo, Japan.

References

[1] National Institute of Standards and Technologies: US Dept. Comm. (NIST), "The NIST Definition of Cloud Computing", NIST Special Publication 800-145, 2011.

[2] Japan Science and Technology Agency, "J-STAGE", https;//jstage.jst.go.jp.

[3] Open Science Framework, https://osf.io.

[4] Research Center for Open Science and Data Platform (RCOS), "Research Data Management Platform (GakuNin RDM)", https://rcos.nii.ac.jp/service/rdm/

[5] Mendeley Ltd, https://www.mendeley.com/

[6] Hirohi Masui, "Collaboration Environment Using a File-Sharing Procedure on the Cloud Computing", 77th JPSJ meeting, Kyoto, 2015.

[7] Shun Ito and Hiroshi Masui, "A Practical Approach for the Project Management using the Public Cloud", International Workshop on Modern Science and Technology (IWMST) 2016, Taichung, 2016, pp. 182-187.

[8] National Institute of Standards and Technologies: US Dept. Comm. (NIST), "NIST Cloud Computing Standards Roadmap", NIST Special Publication 500-291, 2011.

[9] Shibboleth Consortium, https://shibboleth.net

[10] National Institute of Informatics, "Development of the Academic Access Management Federation in Japan (GakuNin)", https://www.gakunin.jp/

[11] Valentina Semkova, Naohiko Otsuka, Marina Mikhailiukova, Boris Pritychenko and Oscar Cabellos, "EXFOR – a global experimental nuclear reaction data repository: Status and new developments", EPJ Web of Conference, 46, 2017, 07003, and references therein.

[12] Nuclear Reaction Data Centre (JCPRG), http://jcprg.org.