ISSN 1598-0170 (Print)
ISSN 2287-1136 (Onine)
http://www.jksii.or.kr

A Data-driven Approach for Computational Simulation: Trend, Requirement and Technology

Sunghee Lee[1] Sunil Ahn[1*] Wonkyun Joo[1] Myungseok Yang[1] Eunji Yu[1]

ABSTRACT

With the emergence of a new paradigm called Open Science and Big Data, the need for data sharing and collaboration is also emerging in the computational science field. This paper, we analyzed data-driven research cases for computational science by field; material design, bioinformatics, high energy physics. We also studied the characteristics of the computational science data and the data management issues. To manage computational science data effectively it is required to have data quality management, increased data reliability, flexibility to support a variety of data types, and tools for analysis and linkage to the computing infrastructure. In addition, we analyzed trends of platform technology for efficient sharing and management of computational science data. The main contribution of this paper is to review the various computational science data repositories and related platform technologies to analyze the characteristics of computational science data and the problems of data management, and to present design considerations for building a future computational science data platform.

☞ Keyword : data-driven, computational science, trend, platform

1. Introduction

There is a growing demand for open science that can open up diverse data such as research, experiment, observation, simulation, and improving the efficiency of research and education. Open science was mentioned as a way to solve current global and social problems such as climate changes through international cooperation at the World Science & Technology Forum which was held in the Republic of Korea in October 2015. Open science [1] aims to make scientific research, data, and other artifacts accessible to everyone. At present, the world trend is to maximize research efficiency by opening and sharing research outcome.

In addition, with the explosion of interest in big data in recent years, a data-driven research methodology that retrieves meaningful information from big data has attracted attention [2-5]. The data-driven research methodology utilizes theories and techniques of various fields such as machine learning, statistics, data mining and artificial intelligence.

With the emergence of a new paradigm called Open Science and Big Data, the need for data sharing and collaboration is also emerging in the computational science field [6]. Research data through computational science is explosively generated, but research on the analysis of these big data is rare. Sharing and reusing data derived through computational science not only avoids duplication of computations that are time-consuming, but also saves costs, as well as a new research method that extracts meaningful information through the analysis of accumulated computational science data can be provided.

These computational science data sharing is still in its infancy, but it has already been applied in several fields. Typical fields include new material development, bioinformatics, high energy physics and so on.

The main contribution of this paper is to benchmark the various computational science data repositories and related platform technologies to analyze the characteristics of computational science data and the problems of data management, and to suggest design considerations for building a future computational science data platform. Section

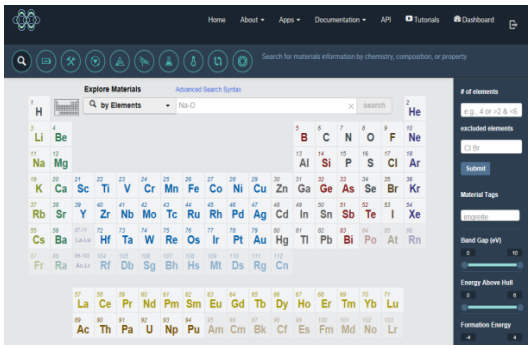
[1] KISTI, 245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea
* Corresponding author (siahn@kisti.re.kr)
[Received 29 August 2017, Reviewed 6 September 2017(R2 25 October 2017), Accepted 27 November 2017]
☆ This research was supported by "Establishment and management of national research outcome utilization system" project of Korea Institute of Science and Technology Information(KISTI).
☆ A preliminary version of this paper was presented at ICONI 2016 and was selected as an outstanding paper.

2 introduces data-driven research cases for computational science by field. In Section 3, we analyze the characteristics of the computational science data and the data management issues. In Section 4, trends of platform technology for efficient sharing and management of computational science data. In Section 5, we conclude this paper and introduce the direction of future research.

2. Data-driven research cases for Computational Science

2.1 Material Design

For the new material design, computational science methods have been used to shorten development time through existing experimental methods. In recent years, some researches have been using data mining, machine learning, genetic algorithms, and other techniques to build material physics databases based on computational science and to analyze and predict material big data to further shorten material development time. Material properties calculations are based on the first principles calculation algorithm and utilize software such as Quantum Espresso [7], VASP [8] and WIEN2K [9].



(Figure 1) The portal of Materials Project [11]

Since 2011, the United States has been promoting material designing using data through the MGI (Material Genome Initiative) [10], and the core program Material Project [11] has been building a material property database through computational science using supercomputers. The current Material Project includes computational simulation data about more than 60,000 inorganic compounds and more than 500,000 nanoporous materials. Japan is building the MatNavi [12] material science database, and NoMaD [13] Project in Europe also collects results of simulating material properties from researchers.

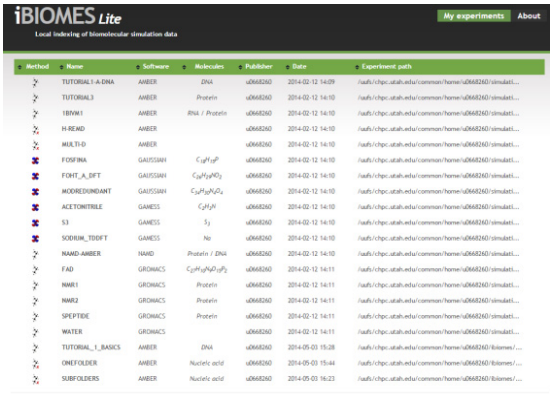
2.2 BioInformatics

Bioinformatics uses molecular dynamics simulations to solve protein folding problems. Molecular dynamics is a method of analyzing the motion of molecules by solving Newton equations through computational simulation. Protein folding and other computational simulation results can be reused for future analysis, but they are not preserved in practice and are often discarded after use. This leads to problems such as redundancy in a long time consuming calculation of molecular dynamics and a lack of reference data for benchmarking.

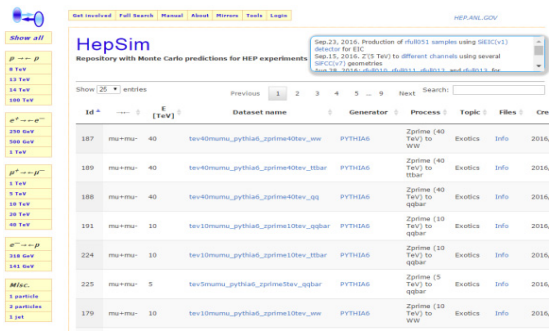
To solve these problems, there have been various attempts to construct the calculated simulation results as a database in the field of bioinformatics. Dynameomics [14] Provides databases and services on the results of approximately 7,000 molecular dynamics simulations for studies on protein folding and stability. MoDEL (Molecular dynamics extended library) [15] constructed a database of molecular dynamics simulation results for 1,700 monomeric soluble structures of PDB [16]. Similar projects include BIGNASim [17], IBIOMES [18], And DCMS [19].

2.3 High-energy physics

In the high energy physics field, Monte Carlo computing simulation method is used to generate and utilize tens of times of virtual data in addition to the data generated by the accelerator. It is statistically difficult to prove theories and models with the data generated by the accelerator only due to lack of the number of samples. Therefore, it is usual to generate additional Monte Carlo data, that is shared and analyzed by researchers. Belle, CMS, ATLAS, ALICE, etc. has built their own customized database for Monte Carlo simulation data, and databases such as HepSim [20] are also in service.



(Figure 2) The portal of iBiomes [18]



(Figure 3) The HepSim service [20]

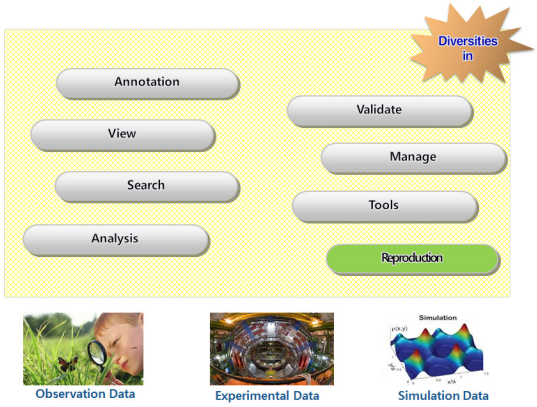
3. The characteristics and issues in the management of computational science data

The primary goal of developing and sharing computational science simulation database is to avoid calculation redundancy and reduce costs. In many cases, in addition, they are performed for data analysis which extracts meaningful information by analyzing the accumulated computational science data. The general data which have been managed through conventional database have requested a quick search, using keywords. In contrast, computational science data often asks for the integrated analysis of related data as well as certain metadata-based search.

First, for effective data analysis, data quality control needed for high-level analysis is critical. Hence, a high level of quality control functions such as curation, validation and workflow is required. To accumulate and search community knowledge in an efficient and productive manner, controlled vocabulary and utilization of ontology are very important.

Second, unlike experimental data, data reliability often becomes an issue in computational science data. Therefore, it is crucial to increase the reliability of simulation data by opening data provenance and revealing a possibility reproduction. It is also needed to provide a method with which the level of simulation data accuracy can be checked through comparison with experimental data. The provenance of the simulation data becomes not only a means of reliability guarantee and a target of search as metadata. Therefore, there should be support on the efficient storage, management, search and analysis of the provenance of the simulation data.

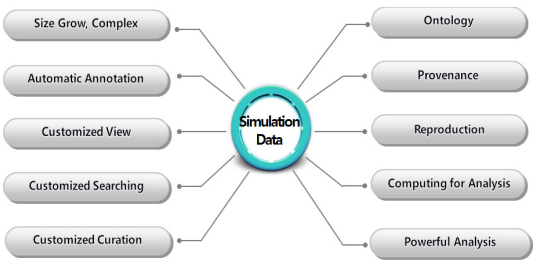
Third, there should be flexibility to support the diverse types of computational science data. Because metadata which can be extracted from the computational science simulation results are very diverse and complicated, there should be a method to automate the extraction of metadata from the simulation data. In addition, several software programs could be used for analysis even under the same purpose, and metadata extraction methods differ. In addition to the metadata extraction method, a data validation method is also different. Depending on data types, data expression, search and analysis methods and utilization tools change. For the management of diverse data in extensible fashion, there should be a flexible framework which can support customized curation, view, search, analysis and other functions.



(Figure 4) diversities in management of scientific data

Fourth, there should be a connection between analysis tools and computing infrastructure. The computational science data are expressed in diverse formats including texts. Therefore, the tools which cover post-processing such as visualization as well as a simple text viewer are often used. Therefore, there should be an option which can connect several tools depending on the type of simulation data or file type. As mentioned above, it is needed to reveal a possibility of reproduction to increase data reliability. Furthermore, it is pretty common to analyze the data through large computing resources for the stored data. To support this, there should be a connection between the simulation data and computing resources.

Figure 5 summarizes the features and management issues of the computational science data described in a single picture.



(Figure 5) characterisitcs and requirements of computational science data

4. Computational Science Data Platform Technology Trends

Sharing and managing computational science data requires a platform to efficiently collect, preprocess, store, access, search and analyze data. The computational science data platform is a data management system that facilitates the collection, selection, preservation, long-term availability, dissemination and access of research data.

The well known platforms in computational science are the Nano Hub [21] and Hub Zero [22]. These platforms provide a foundation for easily leveraging computational science software on the Web and support the development of computing software collaboratively and dissemination. It provides the features to upload and share educational content and simple data, but it does not yet support sharing and managing the results of computational science simulations. NeesHub [23] has developed a platform to share and manage seismic experiment data by extending the hub zero platform.

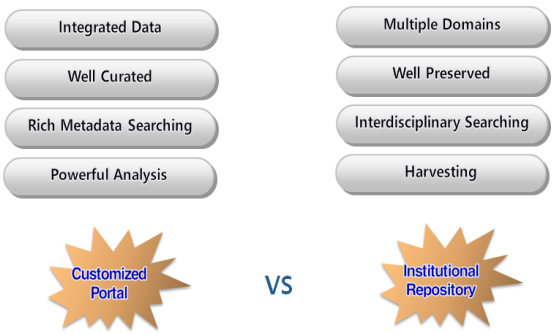
AIIDA (Automated Interactive Infrastructure and Database for Computer Science) [24] was developed with the goal of a flexible and scalable platform for managing, preserving and distributing computational science software, data and workflow. It provides an environment for abstracting various computing infrastructures to execute simulations as well as supporting the repository of simulation result data. In order to increase the reliability of the simulation data, AIIDA manages the provenance of data generation as well as the computational science data itself and provides a reproduction feature. However, it lacks a GUI or Web-based interface, making it difficult for researchers to access. It is currently used to provide an efficient environment to run the Quantum Expresso software in the new material development field and to manage the simulation result data.

Data management platforms are often devided into tow categories: general purpose data platforms and specialized purpose data platforms. General purpose data management platforms include DSpace [25], Fedora Commons [26], Islandora [27], Ckan [28], and Dataverse [29]. Dspace is the digital data repository developed by MIT and HP in 2002 and is currently the most widely used software. It stores and manages various digital contents such as PDF, Word, JPG, MPEG, and etc, and it was developed based on Java and relational database. The biggest feature is the turnkey method, which is very easy to install and use. It supports metadata collection based on the OMI_PMH protocol [30] and may also collect data based on the SWORD protocol [31]. It is possible to customize it to fit in a specific environment, but it is not sufficient to provide the flexibility to support various computational science data. Islandora is a general purpose data repository developed by extending the Fedora Commons. Islandora was developed using the Drupal content management system to support flexible data processing and web-based customized data representation. The preprocessing functions according to the data type can be customized by using the Drupal hook, and the developed modules can be provided by the plug-in method. However, because it is a general-purpose data store, it is not suitable for storing

complex and large-scale computational science data files and does not provide a computing infrastructure necessary for preprocessing or analysis. These general purpose data platforms have limitations in searching for community specific data or providing a high level of analysis.

In addition, there are many special purpose platforms by field. Typical examples are Materials Project, NoMads in materials field, LTER [32] in long-term ecology research field, DataOne [33] in global environment field. These platforms are community- specific data repositories and have the limitations of being less scalable to other areas.

Figure 6 shows cons and procs between special purpose data platform and general purpose data platform. And Figure 7-9 compares and summerizes features of various data platforms. DSPACE, Islandora, CKAN, PURR platforms are in the category of common purpose data platform, and Materials project is in the category of special purpose data platform.



(Figure 6) cons and procs between special purpose data platform and general purpose data platform

	Features	DSPACE	Islandora /Fedora	CKAN	PURR/ HubZero	Material Project
Submission	Large Data	O	O	O	O	
	Bulk Import	O	O	O	O	
Preprocess/ Curation	Customized Ingestion	O	O	O		
	Various Data Ingestion	O	O	O		
	Customized Workflow	O	O	O		O
	Customized Validation		O	O		O
	Validation for Various Data Type		O			

(Figure 7) comparison of various data repositories : data submission

	Features	DSPACE	Islandora /Fedora	CKAN	PURR/ HubZero	Material Project
Metadata	Persistent Identifier	O	O	O		O
	Customized schema	O	O	O	O	O
	Schema for Various Data Type	O	O		O	
	Data Preservation	△	O	O	△	
	Relation Management		O	O	△	△
	Data Provenance	△				O
	Customized Data View	O	O	O	O	O
	Data View for Various Data Type		O		O	
	Automated Annotation		O			O
	Taxonomy, Keyword	O	O	O		

(Figure 8) comparison of various data repositories : Metadata, and Data management

	Features	DSPACE	Islandora /Fedora	CKAN	PURR/ HubZero	Material Project
Security	Single Sign On	O	O	O	?	X
	Access Control	O	O	O	O	X
	Embargo	O	O	X	O	X
Search/ Analysis	Customized Search	O	O	O	O	O
	Flexible Descriptive Metadata Search					
	Data Analysis					O
	Tool Integration		O	O		O
Interface	Open API	O	O	O	O	O
	Command Line Interface	O	O	O	O	O
	Metadata Harvesting	O	O	O	O	
Computing	Reproduction					

(Figure 9) comparison of various data repository : Interface, Security, Search and Analysis

5. Conclusion

This paper analyzed data-driven research cases for computational science by field. And we analyzed the characteristics and management issues for sharing and reusing computational science data. To manage computational science data effectively it is required to have data quality management, increased data reliability, flexibility to support a variety of data types, and tools for analysis and linkage to the computing infrastructure.

Most computational science data platforms are developed specifically for the community, so that they are not scalable to other areas. Conventional general purpose data repositories have limitations in providing advanced search or high-level analysis for community- specific data. Although some researches are aiming at a general purpose computational science platform, they have limitations in their user interfaces and preprocessing features.

References

[1] A. B. Nosek, et al. "Promoting an open research culture," Science, Vol. 348, No. 6242, pp. 1422-1425., 2015.
http://dx.doi.org/10.1126/science.aab2374

[2] Lee, Ki Yong, et al. "Design and implementation of a data-driven simulation service system," Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory. ACM, 2016.
http://dx.doi.org/10.1145/3007818.3007826

[3] S. R. Jeong, and G. Imran, "Semantic Computing for Big Data: Approaches, Tools and Emerging Directions (2011-2014)," KSII Transactions on Internet & Information Systems, Vol. 8, No. 6, 2014.
http://dx.doi.org/10.3837/tiis.2014.06.012

[4] K. Y. Kim, "Business Intelligence and Marketing Insights in an Era of Big Data: The Q-sorting Approach," KSII Transactions on Internet & Information Systems, Vol. 8, No. 2, 2014.
http://dx.doi.org/10.3837/tiis.2014.02.014

[5] M. Chung, J. Kim. "The Internet Information and Technology Research Directions based on the Fourth Industrial Revolution," KSII Transactions on Internet & Information Systems, Vol. 10, No.3, 2016.
http://dx.doi.org/10.3837/tiis.2016.03.020

[6] W. Joo, and et. al. "A Trend of Data-driven Approach for Computer Simulation," ICONI2016, 2016

[7] P. Giannozzi, et al. "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials," Journal of physics: Condensed matter, Vol. 21, No. 39, 2009.
http://dx.doi.org/10.1007/3-540-35426-3-17

[8] J. Hafner, "Ab-initio simulations of materials using VASP: Density-functional theory and beyond," Journal of computational chemistry, Vol. 29, No. 13, 2008, pp. 2044-2078. http://dx.doi.org/10.1002/jcc.21057

[9] G. Blaha, and et. al. "WIEN2k, An Augmented Plane Wave+ Local Orbitals Program for Calculating Crystal Properties," 2001.
http://www.citeulike.org/user/rcollyer/article/6205108

[10] J. Anubhav, P. Kristin, C. Gerbrand, "Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases," APL Materials, Vol. 4. No. 5, 2016.
http://dx.doi.org/10.1063/1.4944683

[11] A. Jain, and et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," Apl Materials, Vol. 1, No. 1, 2013.
https://doi.org/10.1063/1.4812323

[12] T. Ogata and Y. Masayoshi, "New stage of MatNavi, materials database at NIMS," 2012,
http://mits.nims.go.jp/index_en.html

[13] NoMaD Repository, http://nomad-repository.eu.

[14] W. Kamp, and et al. "Dynameomics: a comprehensive database of protein dynamics," Structure, Vol 18. No. 4, 2010.
http://dx.doi.org/10.1016/j.str.2010.01.012

[15] T. Meyer, and et al. "MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories," Structure, Vol 18, No. 11, pp. 1399-1409, 2010.
http://dx.doi.org/10.1016/j.str.2010.07.013

[16] J. Westbrook, and et al. "The protein data bank: unifying the archive." Nucleic acids research, Vol. 30, No. 1, 2002, pp. 245-248.
https://doi.org/10.1093/nar/30.1.245

[17] P. Andrio, and et al. "BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data," Nucleic acids research, Vol 44,2016, pp. 272-278.
http://dx.doi.org/10.1093/nar/gkv1301

[18] C. Thibault, F. Julien, and C. Thomas, "IBIOMES: managing and sharing biomolecular simulation data in a distributed environment," Journal of chemical information and modeling, Vol 53. No. 3, 2014, pp. 726-736. http://dx.doi.org/10.1021/ci300524j

[19] A. Kumar, and et. al. "DCMS: A data analytics and management system for molecular simulation," Journal of big data, Vol. 2, No. 1, 2014.
https://doi.org/10.1186/s40537-014-0009-5

[20] V. Chekanov, "HepSim: a repository with predictions for high-energy physics experiments," Advances in High Energy Physics 2015, 2015.
http://dx.doi.org/10.1155/2015/136093

[21] G. Klimeck, and et. al. "nanohub. org: Advancing education and research in nanotechnology," Computing in Science & Engineering, Vol. 10, No. 5, 2008, pp. 17-23.
http://dx.doi.org/10.1109/MCSE.2008.120

[22] M. McLennan, and K. Rick, "HUBzero: a platform for dissemination and collaboration in computational science and engineering," Computing in Science & Engineering, Vol. 12, No. 2, 2010.
http://dx.doi.org/10.1109/MCSE.2010.41

[23] T. J. Hacker, and et. al. "The NEEShub cyberinfrastructure for earthquake engineering," Computing in Science & Engineering, Vol. 13, No. 4, 2011, pp. 67-78.
http://dx.doi.org/10.1109/MCSE.2011.70

[24] G. Pizzi, and et. al., "AiiDA: Automated interactive infrastructure and database for computational science," Computational Materials Science, Vol 111, 2016.
https://doi.org/10.1016/j.commatsci.2015.09.013

[25] R. Tansley, and et. al. "The DSpace institutional digital repository system: current functionality," Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. IEEE Computer Society, 2003.
http://dx.doi.org/10.1002/asi.10018

[26] D. Wilcox, and W. Evviv, "Supporting Digital Preservation and Access with Fedora," IFLA WLIC 2017, 2017.
http://dx.doi.org/10.1007/3-540-49653-x_4

[27] K. Stapelfeldt and M. Donald, "Islandora and TEI: Current and Emerging Applications/ Approaches," Journal of the Text Encoding Initiative, Vol 5, 2013.
http://dx.doi.org/10.4000/jtei.790

[28] D. Dietrich, and P. Rufus, "CKAN: apt-get for the debian of data," 26th chaos communication congress, 2009. https://ckan.org/

[29] G. King, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," Sociological Methods and Research, Vol. 36, 2007, pp. 173–199.
http://dx.doi.org/10.1177/0049124107306660

[30] C. Lagoze, and et. al. "The Open Archives Initiative Protocol for Metadata Harvesting,"
http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, 2015,
https://doi.org/10.1108/07378830310479776

[31] J. Allinson, F. Sebastien, and L. Stuart, "SWORD: Simple Web-service offering repository deposit." Ariadne, Vol. 54, 2008.
https://doi.org/10.1045/january2012-lewis

[32] W. K. Michener, and B. J. Matthew, "Ecoinformatics: supporting ecology as a data-intensive science," Trends in ecology & evolution, Vol. 27, No. 2, 2012, pp. 85-93.
http://dx.doi.org/10.1016/j.tree.2011.11.016

[33] W. Michener, and et al. "DataONE: Data Observation Network for Earth—Preserving data and enabling innovation in the biological and environmental sciences," D-Lib Magazine, Vol. 17, No. 1/2, 2011.
http://dx.doi.org/10.1045/january2011-michener

◐ 저 자 소 개 ◑

Sunghee Lee

1995 B.S in Computer science, Chonnam National Univ., Gwangju, Korea
2006 M.S in Telecom MBA, KAIST, Daejeon, Korea
1995～2014 Senior Researcher, Korea Telecom, Seoul, Korea
2015～Present Senior Researcher, NTIS Center, KISTI, Daejeon, Korea
Research Interests: Data Sharing, Data Management, Data Analysis
E-mail: sunghee.lee@kisti.re.kr

Sunil Ahn

1997 B.S in Computer science, Chonnam National Univ., Gwangju, Korea
1999 M.S in Computer Science, Seoul National Univ., Seoul, Korea
2010 Ph.D in Computer Engineering, Seoul National Univ., Seoul, Korea
2004～2006 Researcher, IITA, Daejeon, Korea
2006～Present Senior Researcher, Computational Engineering Center, KISTI, Daejeon, Korea
Research Interests: Information System, Grid Computing, Metadata Service
E-mail: siahn@kisti.re.kr

Wonkyun Joo

1997 B.S in Computer Science, Chungnam National Univ., Daejeon, Korea
1999 M.S in Computer Science, Chungnam National Univ., Daejeon, Korea
2018 Ph.D in Computer Engineering, Chungnam National Univ., Daejeon, Korea
1999～Present Principal Researcher, NTIS Center, KISTI, Daejeon, Korea
Research Interests: Semantic Web, Data Mining, Social Network Analysis
E-mail: joo@kisti.re.kr

Myungseok Yang

1999 B.S in Computer Science, Chungnam National Univ., Daejeon, Korea
2001 M.S in Computer Science, Chungnam National Univ., Daejeon, Korea
2017 Ph.D in Computer Engineering, Chungnam National Univ., Daejeon, Korea
1999～Present Senior Researcher, HPC Center, KISTI, Daejeon, Korea
Research Interests: Semantic web, Data mining, Social Network Analysis
E-mail: msyang@kisti.re.kr

Eunji Yu

2008 B.S in MIS/Fire Administration, Wonkwang University, Iksan, Korea
2012 M.S in MIS, Kookmin University, Seoul, Korea
2014~Present Researcher, NTIS Center, KISTI, Daejeon, Korea
Research Interests: Data Mining, Text Mining
E-mail: eunjiyu08@kisti.re.kr