International Journal of Database Theory and Application
Vol.9, No.9 (2016), pp.241-250
http://dx.doi.org/10.14257/ijdta.2016.9.9.22

KNLTER Network: Facilitating Global Data-Sharing

Taesang Huh[1], Sunil Ahn[1], Dukyun Nam[1] and Hoe-Kyung Jung[2]*

[1]*National Institute of Supercomputing and Networking, KISTI, 245 Daehak-ro, Yuseong-gu, Daejeon 34141, KOREA*
[2]*Dept. of Computer Engineering, Pai Chai University, Doma2-Dong, Seo-gu, Daejeon 35345, KOREA*
{tshuh, siahn, dynam}@kisti.re.kr, hkjung@pcu.ac.kr

Abstract

Reliable data sharing and long-term data archiving and reuse are becoming very important in global cooperative research. Concomitantly, many types of global data-based research have been conducted on the Long-Term Ecological Research (LTER) network, with the objective of getting it to respond effectively to future changes in ecology, environment, and climate, by monitoring long-term ecological and environmental data. Korean National Long-Term Ecological Research (KNLTER), however, lacks a system for performing data collection, management, curation, and publication, and therefore global cooperative research through global data sharing is difficult in Korea. In this paper, we analyze one of the best practices link models, the TERN network, and the global data-sharing trend in the LTER area. Further, we propose a link model and necessary technologies for KNLTER and suggest a possible future direction for KNLTER.

Keywords*: KNLTER, Datasets, Data Curation, Data Publication, Data sharing, DataONE, KNB, RDA, ANDS, TERN, Network Node, Metacat, EML, Data License, DOI*

1. Introduction

The purpose of Long-Term Ecological Research (LTER) is to cope with future problems in ecological and socio-environmental issues by monitoring changes in the ecological environment for a long term. LTER began with a program created by the United in 1980 and international LTER (ILTER) was organized in 1993, in order to facilitate cooperation among scientists engaged in long-term ecological research [1-3]. Long-term data archiving and reuse and reliable data curation and publication for data sharing are becoming very important in data-based international collaborative research, Global data exchange is composed of diverse fields and networks of complex structures; therefore, further nationwide interests, such as those from data centers, are needed. Network robustness differs by nodes and the applicability of data is remarkably diverse; thus, several linking problems exist between the nodes. Past Korean National Long-Term Ecological Research (KNLTER) systems not only failed to manage the integrated data accumulation and data quality, but some terms used in the datasets were still in Korean, causing fundamental problems in global data exchange [4-5]. In addition, the system did not comply with Ecological Metadata Language (EML), a global ecological metadata standard, and lacked a software stack for data exchange—hindering data-based global collaborative research. To collect and distribute data for analytical studies in earth science, the United States constructed the Data Observation Network for Earth (DataONE). DataONE, which is based on global data sharing, provides long-term storage of multi-scale and multi-discipline data collected from various countries and allows

* Corresponding author : Hoe-Kyung Jung, hkjung@pcu.ac.kr

ISSN: 2005-4270 IJDTA

researchers, ecology managers, policymakers, students, and educators worldwide to access them. Approximately one terabyte of data are stored in DataONE and hundreds of thousands of metadata are also available [6-7].

In this paper, the global data-sharing trend for long-term ecological research is examined, and the network structure of a best practice link model, Terrestrial Ecological Research Network (TERN), is analyzed as a possible future model for KNLTER. Further, the network, data-sharing model, and necessary technologies in KNLTER among the global networks are discussed [8].

2. Related Research

2.1. DataONE – Research Data Sharing for Earth Science

As a leader in global long-term ecological research, the United States has been developing a distributed framework, called DataONE, to provide a search engine for earth environment data, including ecology and environment, and to support worldwide collaborative research based on the accumulated data. The infrastructure of DataONE, which comprises three nodes, is outlined in Table 1. Coordinating nodes are managed by the DataONE core cyberinfrastructure team and provide core DataONE services such as search and discovery for all datasets. Member nodes expose their data and metadata through a common set of interfaces and services and the investigator toolkit enables access to customized tools that are familiar to scientists and that can support them in all aspects of the data life cycle by DataONE nodes [9-10].

Table 1. DataONE Infrastructures

Infrastructure	Description	
Coordinating Node	- retain complete metadata catalog - indexing for search - network-wide services - ensure content availability (preservation) - replication services	
Member Node	- diverse institutions - serve local community - provide resources for managing their data - retain copies of data	
Investigator Toolkit	- Provide the investigator toolkits that are familiar to scientists and support all of data cycle such as data discovery, analysis, visualization and data management	
	Diverse Softwares	Kepler, Python, Morpho, DMP tool, DataONE Drive, R package, Geoportal, OPeNDAP, IDL, Matlab, etc.

DataONE thereby supports and provides a search engine and access to multi-scale, diverse, and global data, and also offers integration and processing of global-scale data. Furthermore, it facilitates sharing of data and revitalizes the community of experts and educators for profound scientific research.

As shown in Figure 1, when a user publishes a dataset in the member node, the dataset not only syncs with the coordinating nodes with a unique ID, it is also copied to other member nodes for other users to access the dataset from the search engine. In particular, the data originator (A) publishes datasets to Member Node (MN) A. MN A synchronizes with a Coordinating Node (CN) providing the metadata for the new dataset and initiating replication of the metadata across the other CNs as well as the mirroring of the datasets from MN A to MN B. The data originator (A) publishes a journal article including the ID

for the datasets which is wanted to view the data by another user (B). The user (B) access the datasets via MN B after resolving identifier and may annotate the data.

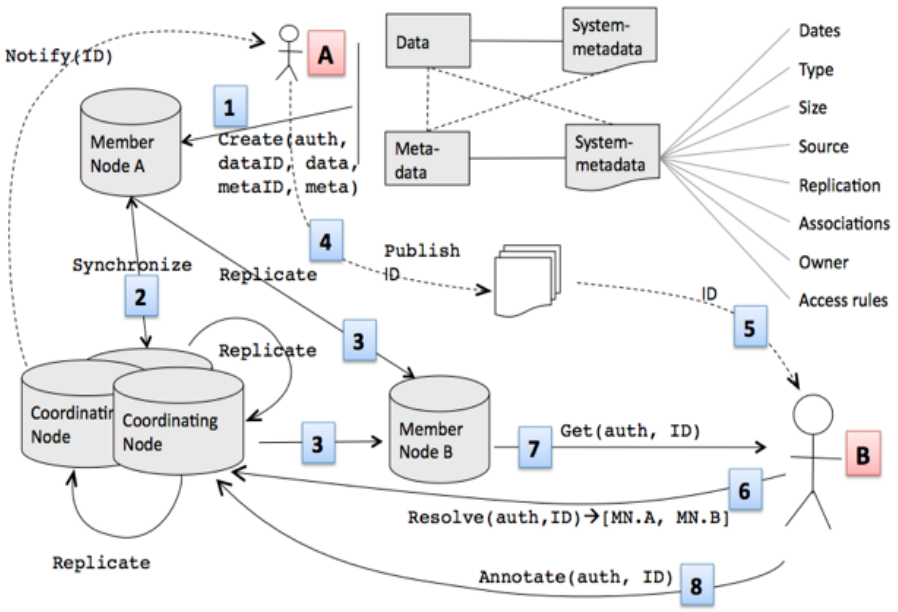

Figure 1. User Interactions with MNs as well as CNs in DataONE [8]

2.2. KNB – Representative Member Node of DataONE

Knowledge Network for Biocomplexity (KNB) is an international repository and a data center for ecological and environmental research. It is one of the main member nodes in DataONE—with the others being ORNL-DAAC and Dryad. KNB gathers projects or data that scientists generate using KNB web interfaces or its PC client Morpho. These data can then be integrated by replicating and harvesting them in the Metacat [11] server of KNB. In addition, the data in KNB can be used actively and extensively because the default data license used in KNB is CC BY. KNB is the main member node of ecological and environmental data, and users actively provide their research data. Moreover, KNB provides metadata that can be used to search for data and tools that can be used to manage and analyze those data. The software and tools provided are shown in Table 2 [12-14].

Table 2. LTER Softwares Provided by KNB

Software	Description
Morpho	Supports earth, environment, and ecology data management for scientists
rDataONE	R package capable of accessing the DataONE repository, which differs from KNB
Metacat	Metadata database
EML	XML-based ecological metadata standard

2.3. RDA – Global Research Data Sharing

RDA (Research Data Alliance), started in 2012 by the United States, Europe, and Australia, comprises Working Group, Interest Groups, Council and Secretariat Staff focusing on scientific data sharing, metadata standardization, interoperability, *etc.* Consulting groups and partnering institutions such as Casrai, ORCID, GODAN, CODATA, ISCU, and DataCite have participated in the following areas: sharing/exchange of research data, use/reuse, standardization, and search. ANDS (Australian National Data Service), a service led by the government to integrate, manage, and share public data, is constituted of lower nodes such as TERN. ANDS provides a DOI (Digital Object Identifier) mining service to the published data in each node and receives metadata to serve as a gateway for users to access the corresponding dataset. TERN is both a lower node of ANDS and a member node of DataONE [12]. It also shares metadata with KNB; TERN vigorously shares global data with the major institutions participating in RDA. To develop a data-sharing environment, the RDA community is devising a global data-sharing framework through general collaboration and considering the requirements of data scientists and domain researchers, infrastructure and data providers, budget policy decision-makers, and data policy makers with the participation of data-providing institutions, unique data ID providers, data-based journals, and institutions/communities that support computing resources and development [15-17].

3. Best Practice: TERN Network

The TERN network is composed of various facilities in a cyberinfrastructure that integrates long-term ecology monitoring organizations. The network provides an environment for ecologists to integrate ecological data through data collection, storage, and sharing. The facilities in TERN consist of five domains: ecological plot data, physical environment, biodiversity & physical environment, data cyberinfrastructure, and data analysis & synthesis. The ecological plot data domain is composed of SuperSites, Transects, AusPlots-F, and AusPlots-R; the physical environment domain Coasts, OzFlux, Soils, and AusCover; the biodiversity & physical environment domain OzFlux, Soils, and AusCover; the data cyberinfrastructure domain Eco-Informatics(AeKOS); and the data analysis & synthesis domain eMAST, and ACEAS. A network diagram of the facilities in TERN, in which each facility provides a portal service to manage datasets, is shown in

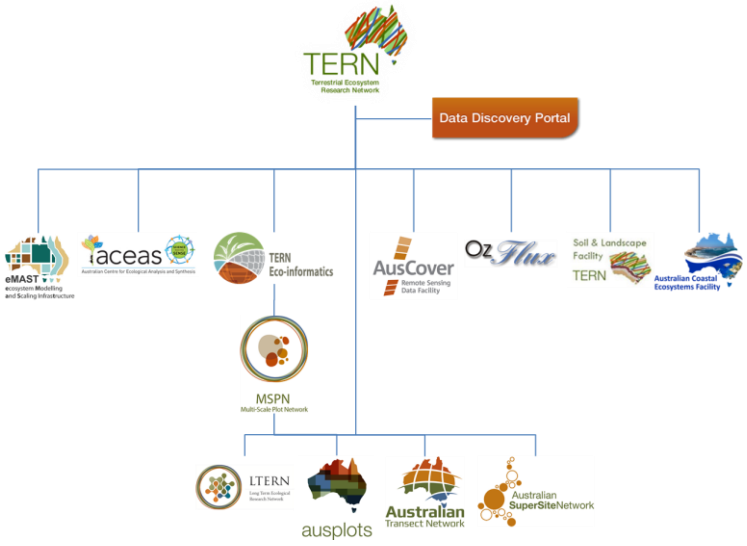

Figure 2. Facilities Network @ TERN

Figure 2 Although the higher hierarchy TERN DDP is a weak network that only shares the harvested metadata, it integrates and manages the data in the lower hierarchy network nodes, and provides search capability and links to the searched data. Each facility manages the license for the data and performs DOI minting service. AeKOS is an Eco-Informatics facility and, unlike the other facilities, does not have an ecology monitoring tool. This facility performs data aggregation and integration and integrates datasets for Fauna and Flora through MSPL. eMAST and ACEAS are computing nodes and carry out Ecosystem modeling and scaling, and ecological analysis and synthesis, respectively [16-18].

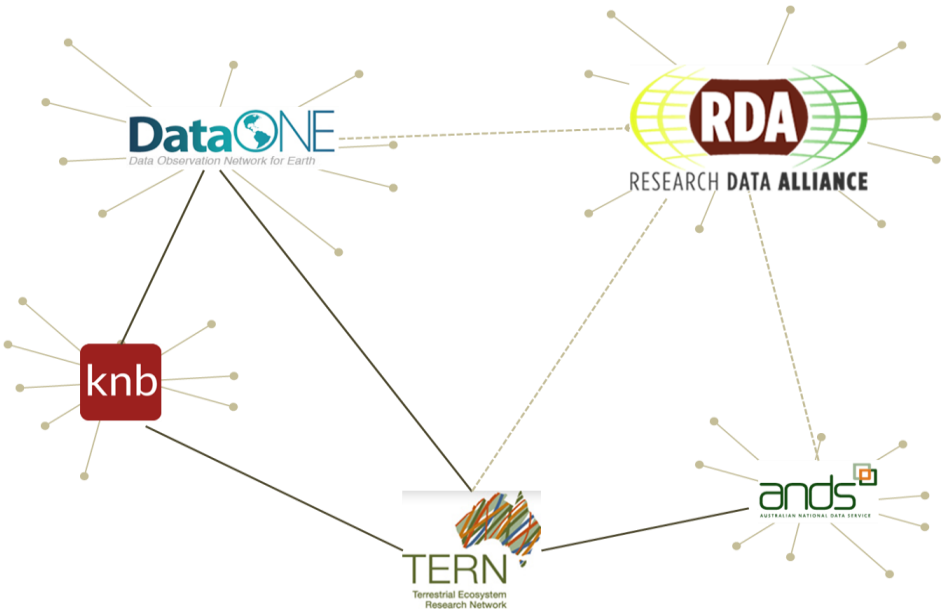

Figure 3. Internation Data Sharing Focusing on TERN

Figure 3 depicts a TERN-centered data-sharing network, which shows that TERN connects to various network gateways to exchange datasets. The Australia National Data Service (ANDS), which deals with data governance in Australia, covers science and other research areas and partners with cooperative research institutes and data-generation facilities to support cooperative research in various study areas. ANDS provides DOI minting services that assign dataset IDs to data centers and all data in the service are managed at the country level. TERN is one of the main data-generation facilities in ANDS and shares metadata with ANDS. Interestingly, even though AusPlot is one of the lower-level hierarchy nodes, it is directly connected to ANDS because it contains original data for other fields of study in addition to ecology. Furthermore, TERN shares datasets with DataONE and its lower-level member node, KNB, because it is a data center in the ecological and environmental fields, as previously mentioned. TERN also shares data with a multi-discipline data cooperative system, RDA, along with Australia's data governance and DataONE [14,16].

4. KNLTER: Global Data Linking Methods and Technologies

Long-term ecological research is leveraged when the local data monitored over a long period of time and the global data are combined. Global long-term ecological research data can be scrutinized when the data collected from each research site are saved together in a primary data repository and saved in second and third data collection repositories—expanding the parameters that can be analyzed [20-21]. The global data link can be applied in two major ways. The first method is for individuals or project-based researchers to directly submit their data to KNB through PC-based Morho or a web

interface. Subsequently, KNB links the submitted dataset with EZID [22] and grants a DOI, then copies it to DataONE through Metacat. The second approach is to construct a system exhibiting input feature, EML exchange feature, or using Metacat for a data center to directly provide the data to DataONE.

4.1. Metacat

The datasets stored in Metacat are duplicated in the member nodes in KNB for future use in the communities. As previously stated, there are two main global data-sharing methods. Both methods are illustrated in Figure 4. In the method depicted in Figure 4(A), individuals or project-based researchers directly share their data, whereas in the method depicted in Figure 4(B), data are shared through a member node in DataONE by constructing a data center. The latter method is a desirable approach for KNLTER. When data are shared through DataONE, the gathered data and metadata are, in general, provided together, but according to the sharing policy of the data center, either only metadata or the link to the data can be provided.

The license for the data shared by the member nodes should be in accordance with the Creative Commons Attribution License (CC BY), and the data center should manage the data that follow other licenses as well. In the case of Korea, CCL and Korean Open Government License (KOGL) are used interchangeably and there has been more movement towards KOGL in recent years.

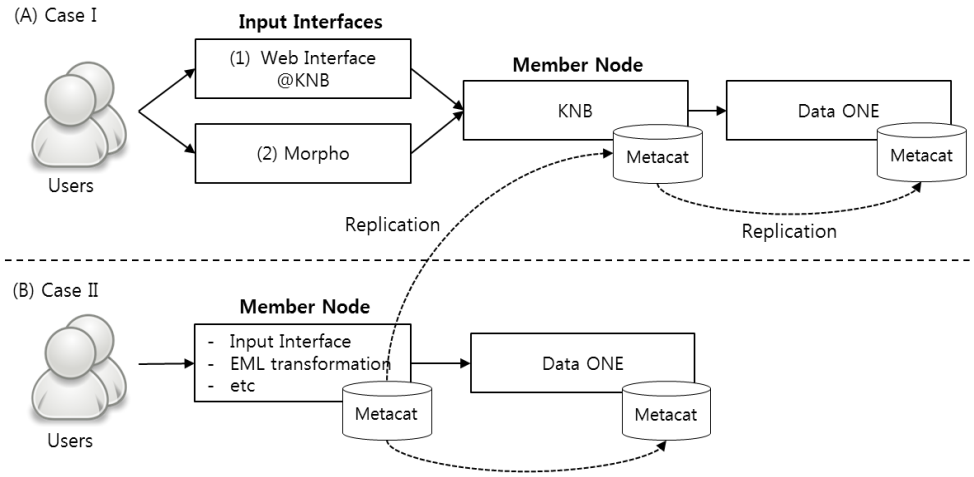

Figure 4. Two Main Methods for Global Data Sharing: (A) Case I: Direct Sharing from Individuals, (B) Case II: Data Sharing through the Data Center

Therefore, two licenses can be used as a mapping format for global linking. Furthermore, DOI needs to be assigned to the datasets for data distribution. Although the data minting service by EZID can be used to receive IDs when the datasets are submitted, the DOI needs to be issued by a separate link to manage DOIs in the data center. Other necessary features include systematic data curation. In order for a data center to register as a member node of DataONE, the published data have to be accessible via long-term storage, each dataset has to be assigned an identifier, and resource maps and reliable data for data packages have to be provided.

4.2. DOI

A DOI is a unique identifier that can be issued to all objects, including internet documents and other digital content. It provides metadata such as authors and dates, and information on the location of the object so that users can have permanent access to the object. The International DOI Foundation (IDF) manages the policies and registrations in

the DOI system. They proposed DOI as an ISO standard for an actionable identifier to ISO TC46 in 2007, after which DOI subsequently became the ISO 26324 standard. The data in DataONE are also distributed using DOI, and most data centers manage the DOI via member nodes such as the Registration Agency (RA) DataCite. In Korea, Korea Institute of Science and Technology Information (KISTI) became an RA in 2016. An identifier system is necessary for KNLTER to share and distribute data and to play a role as a global data center. Thus, constructing a DOI link system in KNLTER will improve its data accessibility.

4.3. Data License

Three important factors are associated with data sharing. First, open access to the ecological research data should be available; second, the right to use existing data should be recognized; and thirdly, cooperative effort by many organizations is necessary.

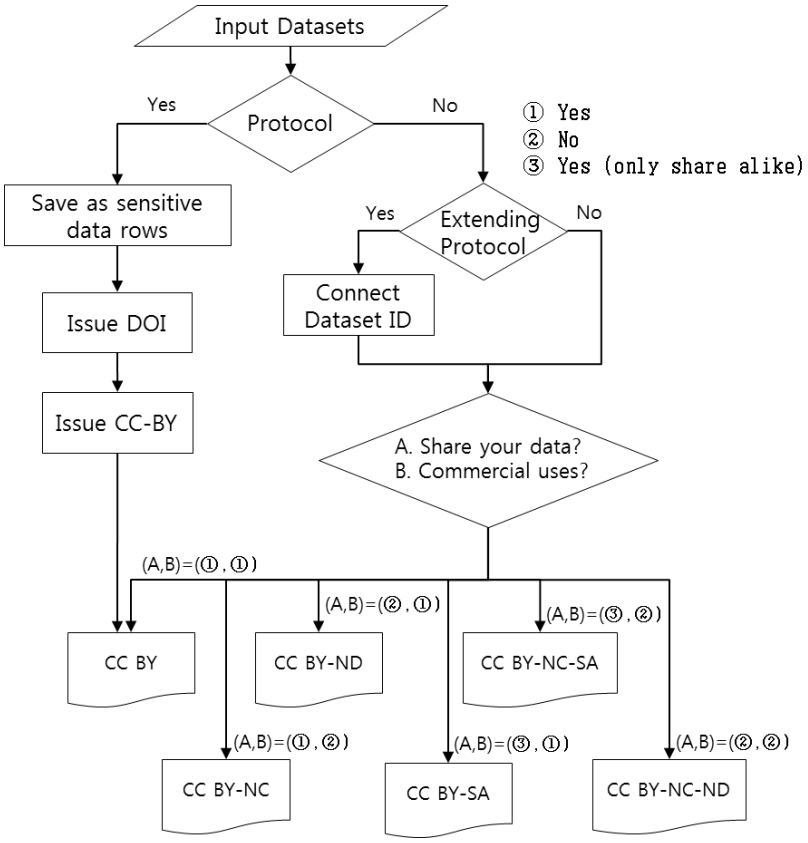

Figure 5. Data License Decision Making Flowcharttk

As shown in Figure 5, when there is an investigation method protocol for the datasets during input, any sensitive data are checked and saved separately. Subsequently, a DOI is issued to the dataset and any protocol-based datasets are issued with CC BY. For those datasets that do not have an apparent protocol, a check is made to determine if there is any extending protocol. If there is such a protocol, the protocol is issued with a foreign key for its dataset ID. Any other data that have no extending protocol are considered private survey data, and will undergo the same data license issuance procedure as the data with the extending protocol. The result is then determined from the answers to two questions: "Do you allow to share your data?", and "Do you allow commercial uses of your data?" The global data link uses CC BY as the default data license, has a research protocol, and is DOI dataset-centered.

4.3. Data Platform Model

A data platform helps users to easily manage and utilize data throughout all steps, from data generation to data sharing. In KNLTER, the data platform consists of several steps. The first step is data curation, in which datasets are submitted, the quality of the datasets managed, and the datasets corrected. The second step is data publication, in which the datasets are stored and published. The third step is data sharing, which comprises data search service, data visualization service, and global data link.

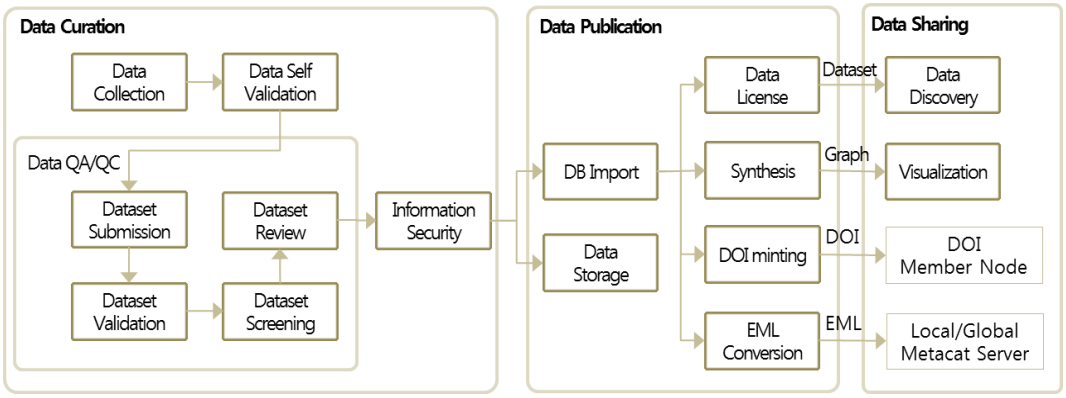

Figure 6. Data Sharing Procedure

Figure 6 shows the collected data curation and publication steps. In these steps, QA/QC and data security procedures for the submitted dataset are performed [6]. In addition, the license and DOI are issued to datasets and EML is sent to the data-sharing server, Metacat, in the data publication step.

5. Conclusions

In this paper, we analyzed global data-sharing systems that are necessary to facilitate the global data-sharing feature in the KNLTER system, and proposed a global link model for the global data link and the data management functions. In particular, we examined Australia's TERN model, which consists of networks that integrate a variety of relevant ecological facilities in Australia. The TERN model includes professional cyberinfrastructure facilities, and is linked into the global data link network through data governance within Australia at the country level. Furthermore, the model links with KNB and DataOne and is oriented to cooperate with a leading cooperative research resource, RDA, for global data sharing between various areas of study. More importantly, the proposed global link model for KNLTER needs a distribution policy between network nodes. This model can be located in various networks according to the network policy. The data link between nodes is performed by harvesting and replicating in Metacat. Further, through replication between nodes, centralized search of global data is technically possible. Data can be identified using the data DOI, and can be accessed from anywhere in the world. In addition, during data reuse, the data license, which can be used internationally to protect intellectual property and indicates the usage scope of the data, can maximize the data's usage via applicable techniques.

KNLTER should be able to construct a mutual reference system based on research data sharing, and an international collaborative research system needs to be constructed based on active data sharing by participating in the activities of RDA—a data collaborative system that analyzes various areas of study.

Acknowledgments

This subject is supported by both Korea Ministry of Environment as "Public Technology Program based on Environmental Policy (Grant No.: 2014000210004)" and the KISTI (Grant No. : K-16-L01-C03).

References

[1] S. Record, P. Ferguson, E. Benveniste, R. Graves, V. Pfeiffer, M. Romolini and B. Beardmore, "Graduate students navigating social-ecological research: insights from the Long-Term Ecological Research Network", Journal of Ecology and Society, vol. 21, no.1, **(2016)**.
[2] E. S. Kim, "Development, potentials, and challenges of the International Long-Term Ecological Research (ILTER) Network", Journal of Ecological Research, vol. 21, no. 6, **(2006)**, pp. 788-793.
[3] K. Vanderbilt, J. Cushing, J. Gao, N. Kaplan, J. Kruger, C. Leroy and L. Zeman, "Data integration challenges: an example from the International Long-Term Ecological Research Network (ILTER)", Journal of Ecological Circuits, vol. 2, **(2009)**, pp. 12-13.
[4] G. Joo, "Korea National Long-Term Ecological Research: Final Reports", **(2013)**.
[5] T. C. Rhyu and B. G. Yang, "The enterprising evaluation for the Korean National Long-Term Ecological Research (KNLTER) Project for six years (Review)", Journal of Ecology and Environment, vol. 34, no. 1, **(2011)**, pp. 11-18.
[6] T. Huh and S. Ahn, "A Study on Global Data-Sharing Model for KNLTER Network", APAIS Proceedings of Information Technology and Computer Science, vol.4, **(2016)**, pp. 62-65.
[7] DataONE, https://www.dataone.org/.
[8] T. Huh and H. K. Jung, "Data Quality Improvement for Korean National Long-Term Ecological Research", International Journal of Applied Engineering Research, vol. 11, no. 12, **(2016)**, pp. 7722-7727.
[9] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse and G. Janée, "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)**.
[10] S. Allard, "DataONE: Facilitating eScience through collaboration", Journal of eScience Librarianship, vol. 1, no. 1, **(2012)**, pp. 3.
[11] J. B. Marshall, "Metacat: A self-watching cognitive architecture for analogy-making and high-level perception", In Proceedings of the 24th Annual Conference of the Cognitive Science Society, **(1999)**.
[12] KNB, https://knb.ecoinformatics.org.
[13] C. Berkley, M. Jones, J. Bojilova and D. Higgins, "Metacat: a schema-independent XML database system", Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings of the 13th International Conference on. IEEE, **(2001)**, pp. 171-179.
[14] B. Leinfelder, J. Tao, D. Costa, M. B. Jones, M. Servilla, M. O'Brien and C. Burt, "A metadata-driven approach to loading and querying heterogeneous scientific data", Journal of Ecological Informatics, vol. 5, no. 1, **(2010)**, pp. 3-8.
[15] RDA Research Data Alliance, https://rd-alliance.org/.
[16] R. D. A. Emergence, "Editorial Building Global Infrastructure for Data Sharing and Exchange Through the Research Data Alliance", D-Lib Magazine, vol. 20, no. 1/2, **(2014)**.
[17] TERN Eco-informatics, http://www.ecoinformatics.org.au/.
[18] S. Guru, X. Shen, C. Love, A. Treloar, S. Phinn, R. Wilkinson and T. Clancy, "Sharing Australia's nationally significant terrestrial ecosystem data: a collaboration between TERN and ANDS", eScience, IEEE 9th International Conference on. IEEE, **(2013)**, pp. 53-60.
[19] AEKOS, TERN Eco-informatics, http://www.ecoinformatics.org.au/.
[20] T. Huh, J. H. Kwak, S. Kim, E. Byun, G. Park, S. Hwang and H. K. Jung, "A Conceptual Model of Ecological Observation Service Supporting Data Life Cycle", Proceedings of International Conference on Convergence Content, **(2014)**, pp. 163-164.
[21] S. Ahn, T. Huh and J. Jang, "Conceptual design of a data repository for the Korea LTER community", ASTL Proceedings of Information Technology and Computer Science, vol. 117, no. 7, **(2015)**, pp. 30-33.
[22] EZID, http://ezid.cdlib.org/.

Authors

Taesang Huh, he received the B.S. degree in electric, electronic and computer engineering from Sungkyunkwan University, Korea, in 2000 and the MS degree in information and communications engineering from Gwangju Institute of Science and Technology, Korea in 2002, respectively. He joined KISTI in 2002 and he is a

senior researcher of NISN at KISTI. His research interests include Metadata Catalog, Distributed Computing, Cloud Storage, e-science and information system.

Sunil Ahn, he is a research staff in the supercomputing center at KISTI in Korea. He obtained his Ph.D degree in parallel computing from Seoul National University (Korea). He has several published journals and conference articles largely in the grid and its application field.

Dukyun Nam, he received the BS degree in computer science and engineering from Pohang University of Science and Technology, Korea, in 1999, and the MS and PhD in engineering from the Information and Communications University, Daejeon, Korea, in 2001 and 2006, respectively. He is currently a head of supercomputer SW research lab in the NISN at KISTI (Korea Institute of Science and Technology Information). His research interests are in High Performance and Distributed Computing, low power computing.

Hoe-Kyung Jung, he received the B.S degree in 1987 and Ph. D. degree in 1993 from the Department of Computer Engineering of Kwangwoon University, Korea. From 1994 to 2005, he worked for ETRI as a researcher. Since 1997, he has worked in the Department of Computer Engineering at Paichai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, information processing, information retrieval, and databases.

