International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 3 (2017) pp. 336-342


# Design Criteria of Korean LTER Data Platform Model for Full Life-cycle Data Management

**Taesang Huh[1#], Geunchul Park[1], Sunil Ahn[1], Soonwook Hwang[1] and Hoekyung Jung[2*]**

*[1]National Institute of Supercomputing and Networking, KISTI,*
*245 Daehak-ro, Yuseong-gu, Daejeon, South Korea.*

*[2]Department of Computer Engineering, Pai Chai University,*
*Doma 2-Dong, Seo-gu, Daejeon, South Korea.*

**[1#] ORCID 0000-0001-5158-7524**

**Abstract**
Long-term ecological research (LTER) is conducted in the form of collaborative research that comprehensively accumulates data for ecology, environment, and biodiversity over a long-term based on accurate data management. Previously, LTER in Korea had many data-management issues, including the absence of integrated repository, making it difficult to conduct data-based research and limiting integrated data use. In addition, international collaborative research was impossible because of the lack of global data-sharing. Against this backdrop, this study was conducted to develop the design criteria for a Korean LTER data platform model that enables full life-cycle LTER data management through various stakeholders and international data-linking.

**Keywords:** Long-Term Ecological Research, Data Platform, Data Quality Control, Metadata, Data Publication, Data Curation, Data-Sharing

## INTRODUCTION

The objective of long-term ecological research (LTER) is to enhance the preservation of nature, biodiversity, ecology, and environment and achieve sustainable and reasonable problem-solving by comprehensively investigating and observing ecological research including the social aspect of human beings in the long-term [1, 2]. It is difficult to understand the correlations among ecology, environment, and climate change with short- and long-term data collection; thus, comparison and analysis must be conducted. This type of LTER requires a data platform that can support the consistent management and analysis of long-term monitoring data on changes in the ecosystem. Countries around the world have built data platforms that can consistently collect, manage, and utilize the monitoring data on changes in ecosystem for their LTER. Representative examples include PASTA and Metacat of the US, DEIMS of Europe, TERN and AEKOS of Australia, and CERN of China [3-5].

South Korea collected LTER-related data through the Korean National LTER (KNLTER) project for ten years since 2004. The objective of the KNLTER project is to establish a scientific and long-term framework to manage changes in the ecosystem of South Korea brought about by climate changes, and use the outcome to develop measures for the preservation of biodiversity. To this end, a number of activities were conducted including research and monitoring of long-term ecological changes in Korea driven by climate change, identification of correlation between changes in climate and environment and those in ecosystem, research on the changes of biodiversity, protocol selection, research on ecological and behavioral changes in animals resulting from climate change and environmental pollution, and research on policy measures for the preservation of biodiversity and biological resources regarding ecological changes[6, 7].

Despite its significant contribution to ecological research, the KNLTER project was provisionally suspended in 2013 due to data-management issues. Owing to the lack of pre-agreed protocol in the KNLTER project, the data collection method and content varied by researcher and site, making it almost impossible to integrate and analyze the collected data. This implied the poor management of planning meetings to draw common survey and analysis items, lack of joint measurement and analysis items with other sites, and failure to use the same equipment and methods for common measurement items [8].

Therefore, as a preliminary step for this study, the data-management issues of the past LTER were analyzed, the trends in global data platforms were identified, and global data-sharing networks were investigated. Through this process, this study attempted to define the main features and element technologies of LTER and present a method

[*] Corresponding author: Hoekyung Jung, Email: hkjung@pcu.ac.kr

for building a data platform model capable of a full life-cycle data management and a global data-sharing network, so as to solve issues regarding data management and integrated use and lay out design criteria for international data-sharing.

## RELATED STUDIES

### Analysis of the problems of the past LTER

The LTER of the past 10 years has revealed several problems. First, the selection of research points (sites) was rather inappropriate. Second, the development of a climate-change-driven ecosystem change model was insufficient. Third, the lack of data-based international collaborative research was identified in the research assessment and consulting results. In addition, the lack of joint and collaborative research among research sites made it difficult to draw items for common survey, analysis, and measurement. The failure to use the same equipment and investigation methods made it impossible to accumulate data for common use. The building and managing of a research information database was identified as the weakest part. There was no conversion to the ecological metadata language (EML) [9], a metadata standard, which would allow for international joint uses, and the metadata was not managed in English. In addition, there was no discussion on data-sharing prior to research and no integrated management system covered data collection to use according to self-assessment [6].

A closer look from the perspective of data management reveals that there was no prior definition on the protocol for common investigation and analysis items and investigation method for sites and measurements. In addition, there were no integrated metadata and data schema for datasets. Thus, the research was conducted around the site according to the propensity of researchers. This led to data fragmentation during the national LTER project because data managed in the individual PCs of researchers were difficult to integrate, inevitably resulting in the accumulation of low-quality data. As a result, owing to the lack of an integrated data repository, the intangible assets of the nation could not be used in an integrated manner, and data exchange with the international LTER (ILTER) network was impossible [8, 10].

### Research on the Trends of the Data Platform Model

The models for building an integrated repository for LTER data can be largely classified into three models by the level of integration. The first is the most loosely-coupled integration model centering on metadata. There is no predefined protocol for data but standard ways to express metadata are defined in this case. Metadata includes not only titles, authors, keywords, and summaries of data but also the schema information of data. This model typically provides controlled vocabulary for effective search of metadata and provides a function to publish metadata based on the EML standard to globally connect metadata. This model is usually used in cases involving many sites with different characteristics from multiple countries. Examples include the DEIMS of European LTER and the TERN of Australia. The second type is a moderately-coupled integration model, which supports data integration through data conversion and synthesis. This type of model not only uses standard ways to express metadata like the most loosely-coupled integration but also alters and processes some of data into standard types for integrated management. In addition, this model is used when there are extensive existing data, such as the PASTA of the USA and AEKOS of Australia. The third type is the most tightly-coupled integration model, which collects and manages data according to the predefined standard protocol. In this model, not only the standard protocol but also the data expression method are predefined according to the protocol, making it ideal for data integration and analysis. This model can be used in a country with no large land and a good implementation system. A typical example of this data integration model is the ECN [11] of the UK. It is also suitable to apply from the initial stage of data management system such as in South Korea.

### Analysis of global data-sharing network

For connecting with international data, many different disciplines and a network of complicated structures are required. Thus, it requires consideration at the level of a data center or a nation. In addition, the maturity of networks varies from one node to another, resulting in great differences in terms of data usability in nodes; the node connectivity also has many problems. An inquiry into the current related networks is necessary for global data-sharing, and the related networks are described in the following sections.

***Knowledge network for biocomplexity :***

The knowledge network for biocomplexity (KNB) is an international repository for ecological and environmental research, and is a representative member node of Data Observation Network for Earth (DataONE). The data created by projects or scientists can be collected through the web interface provided by the KNB or Morpho, which is a PC client, and can be integrated using the Metacat of the KNB. CC-BY based-data license that enables the active use of data is recommended. The KNB, one of the major member nodes of ecological and environmental data receives data from users and provides a data-search function by using metadata and toolkits used for managing and analyzing data, for example, Morpho, DataONE, Metacat, and EML [12].

***DataONE :***

The United States, as a leader in ILTER, has built DataONE as a distributed framework to provide a search engine for earth's environmental data, including ecology and environment, and to support worldwide collaborative research based on the accumulated data. The DataONE infrastructure is mainly composed of three parts. The first part includes the coordinating of nodes, maintaining of catalogues for all data, and provision of core services, such as search based on stored metadata, to DataONE. In the

second part, member nodes are composed of a distributed network based on organizations such as a data center that manages data. Data is shared through the member node interface, and not only scientific data but also computing resources and data replication are provided. The third part, which includes the Investigator Toolkit, refers to software toolkits provided by DataONE familiar to and easily accessible by scientists; they support utilization from all aspects of the data life-cycle. When a user publishes datasets in a member node, they are synchronized with the coordinating node through a unique ID and replicated to another member node so that other users can obtain the data through search. This enables DataONE to support search for and access to multiscale, multidiscipline, and multinational data and provide data integration and synthesis at a global scale. Moreover, it promotes communities of education and expertise, thus strengthening scientific study and enabling data-sharing[13].

***Research Data Alliance :***

The Research Data Alliance (RDA), which started in 2012 by the United States, Europe, and Australia, focuses on the standardization and interoperability of scientific data-sharing and metadata and consists of a working group, interest groups, a council, and a secretariat staff. It is an alliance of consulting groups in the following areas: sharing/exchange, use/reuse, standardization, and search of research data. Their partnering institutions include Casrai, ORCID, GODAN, CODATA, ISCU, and DataCite. To create a data-sharing ecosystem, the RDA community is building a global data-sharing system by collaborating with data providers, budget policy decision-makers, and data policymakers to meet the requirements of data and domain scientists, through the participation of infrastructure and data providers, unique data ID distributors, data-based academic journals, and organizations and communities supporting computing resources and development [14].

## ELEMENTARY TECHNOLOGY FOR DATA PLATFORM

Ecology is a field of study of the interactions between living creatures and the environment. Here, the environment refers to the surroundings of living creatures, including both biological and nonbiological factors. Thus, ecology is intertwined with other disciplines much more than any other study. As interdisciplinary research is emphasized in ecology, it draws the attention of researchers from a wide range of disciplines and requires long-term research. Thus, organizing, classifying, and storing research results achieved by researchers from different disciplines over a long-term according to structured and common standards is needed in ecology research more than in any other research field. The lack of sustained and common standards can make it impossible to share ecological observation data that are widely exchanged across the globe, leading to the misunderstanding of data because of the subjective judgment of researchers when such data are stored over decades. Therefore, standard-based accumulation of data is required.

- **Data Characteristics:** Data characteristics include data heterogeneity, as data has various types of datasets and different volumes. Collected raw data also generates linked data because the raw data is processed for research and analysis, resulting in data complexity. With the development of sensors driven by the automation of research equipment, data forms have become diverse; thus, the data processing methods change; this is referred to as data variability.
- **Data Accumulation:** There are two types of data: the first comprises automated records obtained from using measurement equipment based on a protocol for long-term spatial data research and the other comprises the data directly recorded by researchers. Data includes variable specifications, units, accurate records, terrain size, resolution, reference systems, and quality assurance procedures. Data is stored in the integrated repository and metadata is built in the RDBMS for easy access to all data.
- **Data Integration:** Based on biological and environmental data, the environmental, soil quality, chemical, meteorological, and spatial information can be utilized as external data. Regional meteorological information services by the meteorological authority, map coordinates of the geographical statistical system, old and new address services, and Google map data can become the targets.
- **Private Data Management; created by private users:** Although it is crucial to collect and accumulate data pursuant to the standard protocol, accumulating private survey data separately can present an important starting point for full life-cycle data management. Therefore, it is necessary to integrate and manage user-defined datasets according to the standard protocol.
- **Data Conversion and Synthesis:** A data model that can be jointly used through conversion and synthesis in a single format is needed to improve the utilization of ecological data. Measurement values must be separately managed in the form of defined time series to enable statistical and correlation analysis. This requires a global schema for integrated management is needed separately from local schemas corresponding to a variety of protocols. The conversion of units, language integration using controlled vocabulary, and conversion and synthesis into time series according to sampling rules [15].
- **Data Exchanging:** International data exchange involves various areas and a network of complicated structures, thus requiring consideration at the level of a data center or nation. Moreover, standardized metadata must be used for ecological research to connect the metadata to global data.
- **EML:** This was developed by ecologists and is a set of XML schema documents that allow for the structural

expression of metadata. EML is scalable by using the modules standardized by units. Each module is designed to describe metadata by categories and can be regarded as a major technology for international data exchange.

- **Data Quality:** Low-quality data cannot be trusted and hinders utilization. Quality management is critical in the entire process, that is, from data collection to storage and use. Quality control must include protocols, maintenance schedules, calibration specifications, and clear instructions on the manipulation of measured data, and provide the management of data quality deviations according to the collection cycle. In addition, data validation is required to maintain the quality of datasets by following the validation process before data is loaded to the database.
- **Data Publication:** The collected data goes through a data curation process for quality improvement. Here, a wide range of stakeholders validate the data through a data platform [16]. The validity of the submitted data is verified when data owners submit the data. Furthermore, the data manager and reviewer separately conduct data screening and semantic review before publishing the data for long-term data preservation. At this moment, a data identifier, such as the digital object identifier (DOI), is assigned to facilitate data reference and quotation [14].
- **Data Access:** The user classification for data access in the system must give priority to ecological researchers. This also has an educational purpose, including the nurturing of the younger generation. In addition, summarized information must be provided for the media and policy makers.
- **Data Security:** Although ecological data in principle must be made public, it is also important to maintain the security of sensitive data that requires restricted access and nondisclosure.
- **Data License:** Most of ILTER data platforms use the international standard Creative Commons License [17], whereas Korea has established the Korea Government Open License (KOGL) [18], which requires permission for free use of copyrighted public works. Thus, both data license rules must be applied through mapping to deliver both national and international services.
- **Controlled Vocabulary:** The provision of a standard list of vocabularies by using tags on data and information makes it easy to communicate and exchange knowledge among stakeholders. In datasets with heterogeneous schema, controlled vocabulary makes it possible to search and use similar datasets.

## DESIGN CRITERIA OF FULL LIFE-CYCLE DATA PLATFORM FOR KNLTER

### Main Features Derivation

The main features that meet the requirements of a data platform for the full life-cycle management of LTER information are as follows [19]. The first feature involves the entering and validating of datasets. For LTER data, quality is the most important characteristic. Therefore, data is collected according to the defined protocol, and the collected data is validated to determine whether it follows the defined schema. As for data-quality validation, the data type, scope, and category are automatically validated first through the system, followed by semantic validation by experts using visualization information on the data. The second feature involves metadata management. The main metadata includes datasets, measurement sites, users, organizations, and validation rules. Each metadata can provide low-level semantics by providing mutual references. In addition, metadata provides quality control for input. The third feature involves search functions for datasets and data. Search includes faceted, map, keyword, and unified searches. The fourth feature involves the management of controlled vocabulary that improves the utilization of heterogeneous data. The fifth feature is the EML conversion of metadata. EML is a standard for long-term ecological metadata technology, and metadata must be stored and converted to the EML format to connect to international ecological research sites. The sixth feature involves data conversion and synthesis. To support the integration of data stored in different formats, the conversion of data in a unified format must be supported. To this end, conversion and synthesis rules are defined in each measurement and a transformation process is conducted each time the data is changed. The seventh feature involves data visualization. This includes the expression of data in a time-series chart based on measurements and sites, and the comparison between two measurement data to ease data analysis. The eighth feature involves the management of scalable data. We can respond to the KOGL by internally using NoSql, and increase data volumes by managing large external data in a cloud. The ninth feature is a user-convenience function that can enhance accessibility to data platforms. This service is delivered through mobile devices.

### Data Platform Model for Full Life-cycle Data Management

Low-quality data in the LTER cannot be trusted and hinders data use. Quality management is crucial in the entire process, that is, from data collection to storage to use. Data must be created in compliance with the measurements defined in the protocol. Data handling must comply with the standard operational procedure by applying quality criteria, such as accuracy, and the target specification for record resolution must be included in the standard operational procedure. Data distribution must be performed by validating data entered by user, continuously monitoring data by the data manager for better quality, and conducting semantic verification of the input data in the community.

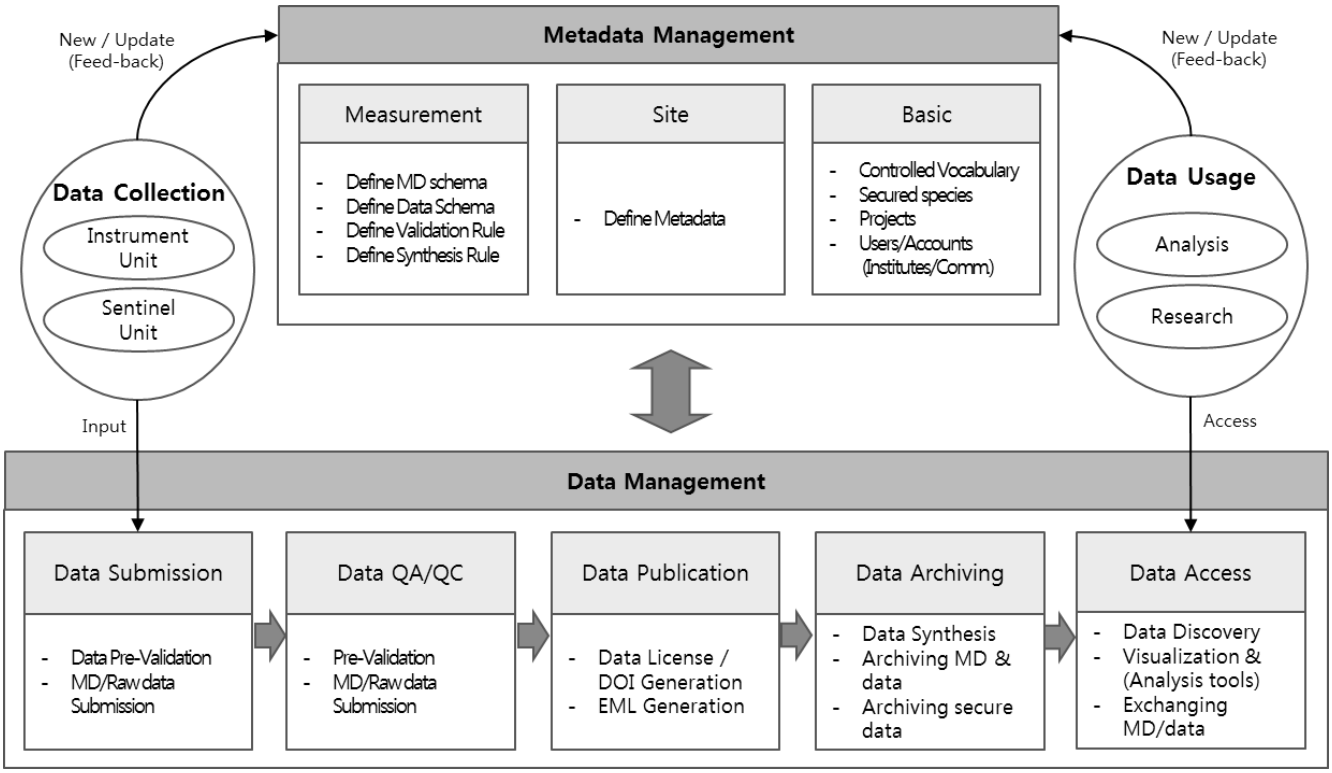


**Figure 1.** Workflow of LTER Life-cycle Management

Figure 1. illustrates the workflow of the LTER data management that reflects the solutions to the problems discussed earlier. The absence of a measuring protocol is defined in advance in the community and managed by metadata, while the problems of data fragmentation and low-quality data are addressed through systematic data management and data quality management. In addition, an integrated repository is provided to store structured and unstructured data by applying the same system structure to both database and storage. Further, data processing, map-based faceted search, and controlled vocabulary management improve data utilization. Data license is assigned for data distribution and the DOI, which is a data identifier, is generated for connection. Moreover, the EML, which is an international standard metadata, allows compatibility through Metacat, enabling global data-sharing [7, 20-22].

**Network Model for Global Data-Sharing**

Accurate analysis becomes easy in the LTER when global data is gathered from local data monitored over the long-term. Sophisticated interpretation of ILTER data becomes feasible when the data collected from each research site is integrated and stored in the primary data repository, and then stored in the secondary and tertiary data collection repositories; this increases the number of parameters used in data analysis and enables more elaborate analysis. Data can be directly provided to DataONE from the data center by building a system with an input function, an EML transformation function, and Metacat. The datasets stored in Metacat are replicated in the member nodes of the KNB so that they can be used in the community.

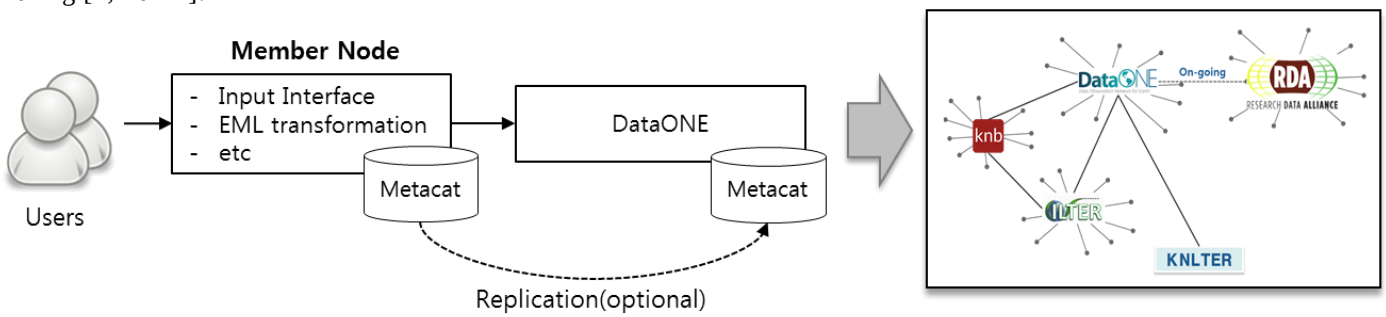


**Figure 2.** Global data sharing through the data center

Figure 2 shows the sharing of data through the member nodes of DataONE by building a data center for international data-sharing. The desirable model for the KNLTER system also involves the sharing of datasets collected at DataONE and KNB through a data center. In principle, not only the actually collected data but also metadata must be provided when sharing data with DataONE. However, depending on the sharing policy of the data center, it is also possible to provide only metadata with link information for the data [1]. This allows DataONE to exchange data with KNB, a member node and the ILTER network, and the possibility of data-sharing with the RDA can be expected in the future.

## CONCLUSIONS

High-quality data is essential for data-based research. To ensure this, data quality must be managed throughout the entire cycle of data curation from data collection to publication. For the LTER, the integrated use of datasets according to their purpose must be considered through correlations among the datasets, and the design of the LTER data platform must reflect characteristics such as data heterogeneity, complexity, and variability. Flexible system scalability is required because long-term large data are necessary for the research of ecology, environment, and climate change. In some cases, global data exchange is required to use international data with similar geographical characteristics for association analysis.

In this study, systems and data were analyzed and the trends of global data platforms were investigated to identify data management problems of the KNLTER. In addition, the global data-sharing networks were analyzed as a preliminary study. As a result, the main features and major element technologies were derived, and a data platform model capable of full life-cycle data management and a network construction method for global data-sharing were proposed. This resulted in the development of design criteria of a data platform model for data management, integrated data use, and international data-sharing. When compared with global data platforms, the proposed data platform model stands out in terms of the composition of essential features for a single platform. In addition, it is expected to support the data management of the researchers of the KNLTER community and serve as a data center node for the ILTER network.

## ACKNOWLEDGEMENTS

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-17-L01-C01).

## REFERENCES

[1] Huh, T., S. Ahn, D. Nam, and J. Hoe-Kyung, KNLTER Network: Facilitating Global Data-Sharing. International Journal of Software Engineering and Its Applications, 2016.

[2] LTER. Long Term Ecological Research Network, Available from: https://lternet.edu/.

[3] Jang, S.A.J. and T. Huh, Quality Assurance for the K-ecohub LTER Data. 2016.

[4] Servilla, M., J. Brunt, I. San Gil, and D. Costa, PASTA: a network-level architecture design for generating synthetic data products in the LTER network. LTER Databits, 2006.

[5] Fu, B., S. Li, X. Yu, P. Yang, G. Yu, R. Feng, and X. Zhuang, Chinese ecosystem research network: progress and perspectives. Ecological Complexity, 2010. 7(2): p. 225-233.

[6] G. Joo, et al., "Korea National Long-Term Ecological Research: Final Reports", **(2013)**.

[7] Rhyu, T.C. and B.G. Yang, The enterprising evaluation for the Korean National Long-Term Ecological Research (KNLTER) Project for six years. Journal of Ecology and Environment, 2011. 34(1): p. 11-18.

[8] Huh, T. and H.-K. Jung, Data Quality Improvement for Korean National Long-Term Ecological Research. International Journal of Applied Engineering Research, 2016. 11(12): p. 7722-7727.

[9] EML. Ecological Metadata Language, Available from: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html.

[10] Vanderbilt, K., J. Cushing, J. Gao, N. Kaplan, J. Kruger, C. Leroy, J. Mallett, K. Ramsey, and L. Zeman, Data integration challenges: an example from the International Long-Term Ecological Research Network (ILTER). Ecological Circuits, 2009. 2: p. 12-13.

[11] Lane, A., The UK environmental change network database: An integrated information resource for long-term monitoring and research. Journal of Environmental Management, 1997. 51(1): p. 87-105.

[12] KNB. Knowledge Network for Biocomplexity, Available from: https://knb.ecoinformatics.org/.

[13] DataONE. Data Observation Network for Earth, Available from: https://www.dataone.org/.

[14] Emergence, R., Guest Editorial Building Global Infrastructure for Data Sharing and Exchange Through the Research Data Alliance. D-Lib Magazine, 2014. 20(1/2).

[15] Huh, T., J.-H. Kwak, S. Kim, E. Byun, G. Park, S. Hwang, and H.-K. Jung, Data Conversion and Synthesis According to Ecological Observation Datasets. International Conference on Convergence Content (ICCC) 2015, 2015. 13(2): p. 205-206.

[16] Klump, J., R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Höck, M. Lautenschlager, U. Schindler, I. Sens, and J. Wächter, Data publication in the open access initiative. Data Science Journal, 2006. 5: p. 79-83.

[17] Corporation, C.C. Creative Commons, Available from: https://creativecommons.org/.

[18] KOGL. Korean Government's Open License, Available from: http://www.kogl.or.kr/.

[19] Ahn, S., J. Jang, and T. Huh, Conceptual design of a data repository for the Korea LTER community. Advanced Science and Technology Letters, 2015. 117.

[20] Berkley, C., M. Jones, J. Bojilova, and D. Higgins. Metacat: a schema-independent XML database system. in Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on. 2001. IEEE.

[21] Kim, J.Y., G.-J. Joo, G.-Y. Kim, B. Yang, M. Kim, and C.S. Lee, Korea National Long-Term Ecological Research: provision against climate change and environmental pollution (Review). Journal of Ecology and Environment, 2011. 34(1): p. 3-10.

[22] Record, S., P. Ferguson, E. Benveniste, R. Graves, V. Pfeiffer, M. Romolini, C. Yorke, and B. Beardmore, Graduate students navigating social-ecological research: insights from the Long-Term Ecological Research Network. Ecology and Society, 2016. 21(1).