International Journal of Database Theory and Application
Vol.9, No.10 (2016), pp.317-326
http://dx.doi.org/10.14257/ijdta.2016.9.10.27

Data Curation for LTER: The Case of K-ecohub

Sunil Ahn[i], Taesang Huh[1], Soonwook Hwang[1], Jihoon Jang[1] and Sunghee Lee[2]*

[1]*Faculty of Supercomputing Center,*
Korea Institute of Science and Technology Information,
245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea.
[2]*NTIS Service Center,*
Korea Institute of Science and Technology Information,
245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea.
[1]*E-mail: {siahn, tshuh, hwang, jangoqe}@ kisti.re.kr*
[2]*E-mail: sunghee.lee@ kisti.re.kr*

Abstract

In LTER, curation activities play very essential role for data discovery and retrieval, quality control, and reuse over time. The K-ecohub system is a pilot repository to preserve long-term Korean national ecological data. The repository has been designed so as to manage and share data in an efficient manner. The paper presents the workflow-based data curation process in K-ecohub, which promotes collaboration and facilitates contribution from experts in the LTER field and also provides a way to automate and customize curation activities depending on the data types.

Keywords*: Data Curation, Workflow, LTER, Quality Assurance*

1. Introduction

The LTER network aims to appropriately deal with today's environmental issues by observing climate and environmental changes and tracing the consequences [1]. To provide valuable perspectives for solving environmental problems, it is essential to collect data over a long period of time and analyse them comparatively. In addition to this, using cyber infrastructure is imperative for preserving LTER data and analysing them to engage with decision makers [2]. There have been several efforts to build a system that can collect and preserve the observed LTER data that consists of various ecological changes; the attempted systems include Metacat [3], PASTA (Provenance Aware Synthesis Tracking Architecture) [4], DEIMS (Drupal Ecological Information Management System) [5], ECN (Environmental Change Network) [6], CERN (Chinese Ecological Research Network) [7], TERN (Terrestrial Ecosystem Research Network) [8] and AEKOS (Australian Ecological Knowledge and Observation System) [9].

The KNLTER project [10] has been collecting long-term ecological data in Korea since 2004. In addition to harvesting LTER data, the goals of the project include building a system to manage collected data and help ecological researches, and planning for biodiversity conservation. The project had made a substantial contribution to the overall ecological researches. However, the project was suspended in 2013 due to several causes: poor planning and lack of consensus to derive common survey and analysis of items; lack of common measurement items shared with other sites; and absence of a shared quality assurance plan [11]. To deal with these problems, a new pilot project called K-ecohub was initiated in 2014. The aims of the new project include: developing data collection protocols for Korean LTER data, studying an efficient data model, and developing cyberinfrastructure for preserving LTER data.

ISSN: 2005-4270 IJDTA

The K-ecohub system is a repository to preserve long-term Korean national ecological data and to facilitate management and sharing of data. It has been developed to tackle issues of the KNLTER project such as: lack of consensus on data collection protocols; data fragmentation; absence of quality assurance; absence of a repository for efficient data integration; and poor linkage with the global LTER data.

Data curation is a management activity related to data integration, annotation, publication and presentation. It helps data remain available for reuse and also keeps it up-to-date and more findable. In LTER, curation activities play a very essential role in data discovery and retrieval, quality control, and reuse over time [12].

The previous LTER systems [3-9] assisted data curation in various ways such as data validation and data synthesis. Similarly the K-ecohub system curates data to enhance data usability and discoverability through various curation activities: data validation, ingestion, synthesis and expert reviews. The curation process in K-ecohub is differentiated from other LTER systems in that all the curation activities are processed by an automated workflow. The curation workflow promotes collaboration, facilitates contribution from experts in the LTER field, and also provides ways to automate and customize the curation process depending on the data types.

This paper presents the workflow-based data of the curation process in K-ecohub. In Section 2 we review how other LTER systems manage the curation process. Section 3 provides an overview of the K-ecohub, and Section 4 details its curation process. In Section 5 we conclude with discussions and directions for future research.

2. Related Work

PASTA [4] curates ecological research data harvested from LTER networks through data synthesis. Each site in the US LTER network manages collected data within its own site, and which metadata are maintained and shared with the MetaCat [3] software. This makes data in different sites fragmented and hard to be integrated. PASTA, to tackle this issue, harvests raw data from the LTER network sites and reprocesses the data to fit well in the predefined global schema such that it makes the data more connected and integrated. Data harvesting and synthesis are processed through the EML Parser and Loader, which is an automated process that helps the data get converted into a standardized format. The metadata of the synthesized data are registered into the MetaCat for sharing, reuse and analysis.

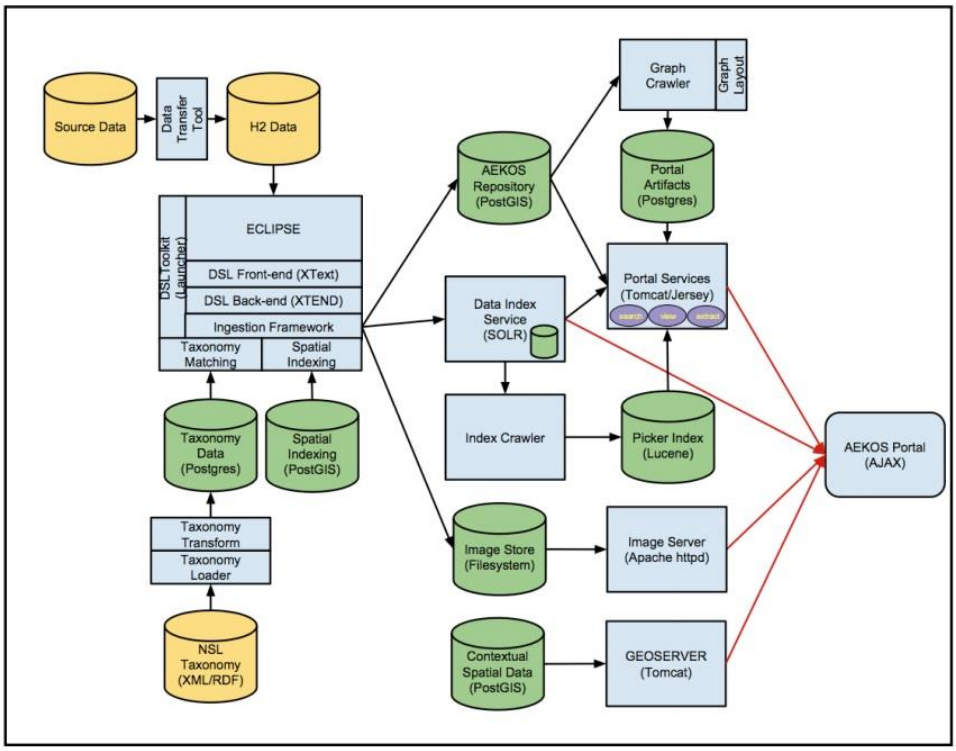

Figure 1. Data Ingestion and Processing in AEKOS [9]

AEKOS [9], a subnetwork of TERN [8], is a project that was developed to solve data fragmentation issues such as lack of data context, data diversity and lost data. AEKOS curates ecological research data collected from several sites in a manner similar to the PASTA project. The "Data Ingester" automatically harvests data from each site and the ETL (Extract, Transform and Load) script and the DSL (Domain Specific Language) convert data into the pre-defined format. The reprocessed data are then finally stored in the AEKOS infrastructure.

3. The K-ecohub System

The K-ecohub system is an LTER data repository used to collect, share and integrate data in a consistent and efficient manner. To solve the data fragmentation issues that were observed in the KNLTER project, the following various features have been requested.

The first feature is to support data validation. The K-ecohub system examines if the collected data follows the predefined protocol and schema. The data validation process includes automatic verification of data type, scope and category, followed by experts' review process.

The second feature is management of metadata to support data integration and easy searching. The K-ecohub provides a way to manage controlled vocabularies, and to convert metadata into an EML format for linkage with external data in global LTER sites. EML [13] is a standard developed for the long-term ecology, and provides a variety of searching options including multi-faceted search, map based search, keyword search and integrated search.

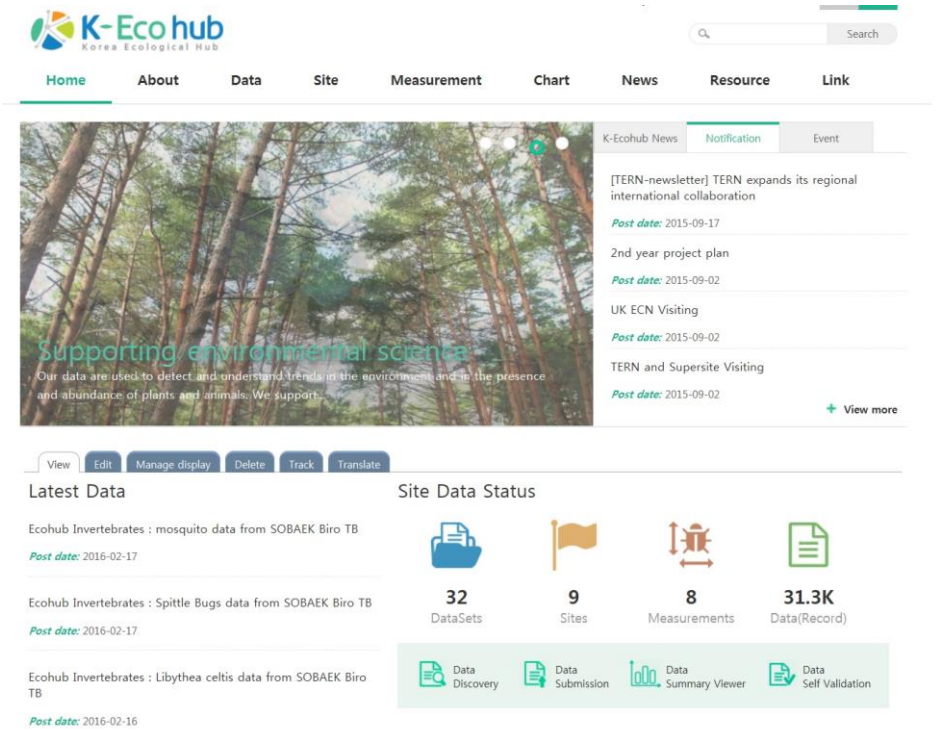

Figure 2. The K-ecohub Repository Portal

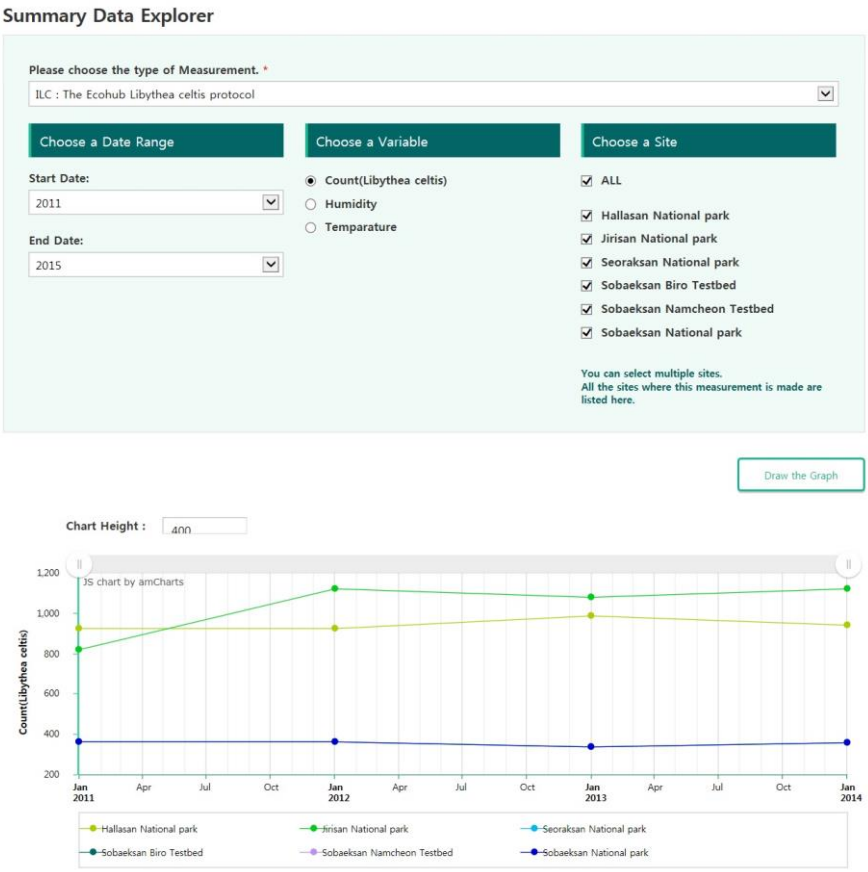

Figure 3. Data Visualization in the K-ecohub System

The Third feature is a process to transform data and enable data synthesis, data integration and data visualization. Several transformation or synthesis rules may be defined for each protocol and the data can be transformed according to the rules on new data inputs or updates. The rules are currently defined by an SQL. They will be extended to support general script languages.

The fourth feature is to support a visualization chart that presents data in a time series chart, and enables data comparison between protocols or among multiple sites. This feature may improve the efficiency of data analysis and promote effectiveness in decision making.

The fifth feature is to manage data in a scalable and flexible way. To support this, the K-ecohub system is equipped to manage data in a cloud, making it possible to handle a growing data volume and complexity.

4. Curation Process in K-ecohub

The K-ecohub data model is characterized through standardized protocols that define how to observe, collect, and store data. This standardized protocol ensures consistency in data collection by defining types, units, subjects, contents, and methods. The data collected through this kind of standard protocol are ingested into the K-ecohub system based on preset schema, and the data quality is controlled in a consistent manner.

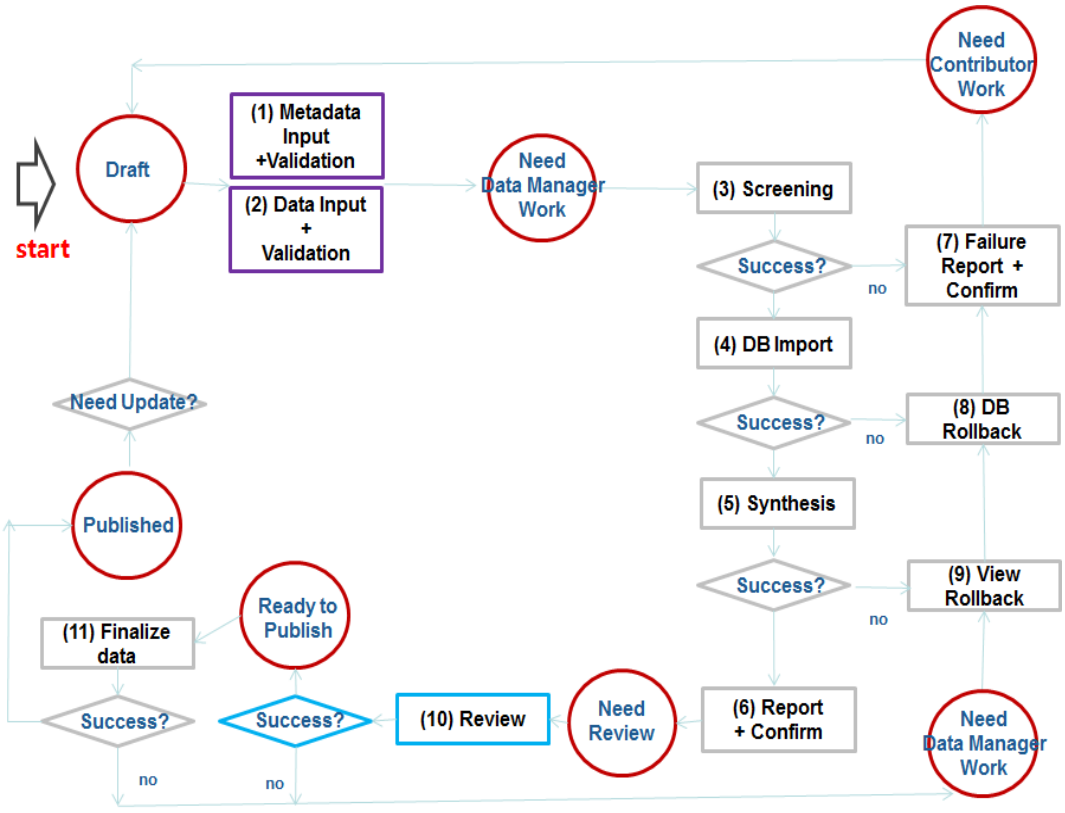

Figure 4. Data Curation Process in the K-ecohub System [11]

The curation process in K-ecohub is composed of 7 steps. Figure 4 shows the curation workflow when a dataset is submitted into the K-ecohub system. The first step is to validate metadata submitted by a data contributor. The metadata includes a protocol name, a site name, owner information, contact information, keywords, species, coordinates, access information, starting & ending date of data collection, publication date, and so on. The K-ecohub system automatically verifies whether essential fields of

metadata are provided, and if the metadata values are within an acceptable scope. In addition to this, the K-ecohub system automatically verifies the types and categories of submitted metadata by making use of predefined validation rules.

The second step is to validate the data itself. In K-ecohub, a data contributor prepares the data that conform to a predefined schema as defined by the protocol. The submitted data may include one or more text files that are recorded in a CSV file format. In this step, each column and row in the data files is validated against a predefined validation schema, which may include type, scope and category rules that the data should follow.

Data Self Validation

Please choose the rule for Data Self Validation.

ILC_SCO : The Ecohub Libythea celtis protocol

Entity Name	Field Name	Unit	Validation Rule
ILC_SCO	Sdate	date	regex("[0-9][0-9][0-9][0-9][-][0-9][0-9][-][0-9][0-9]")
	Sweek	natural	regex("[0-9]+")
	Recoder	text	
	Stime	time	regex("[0-9][0-9][:][0-9][0-9][:][0-9][0-9]")
	Etime	time	regex("[0-9][0-9][:][0-9][0-9][:][0-9][0-9]")
	Temperature	celsius	regex("[0-9.]+")
	Humidity	percent	regex("[0-9.]+")
	WindSpeed	metersPerSecond	regex("[0-9.]+|")
	WindDirection	degree	regex("[0-9.]+|")
	Count	natural	regex("[0-9]+")
	Weather	text	is("sunny") or is("cloudy") or is("rainy") or is("snowy")
	Reference	text	
	Description	text	
	Site	text	

Figure 5. Data Validation Rule Definition in the K-ecohub System

The third step is data screening. In this step, a data manager manually verifies metadata and data. If no error is found in the data and metadata, data are made ready for database import. If the data manager finds any error, the submitted data is sent back to the data contributor for modification and resubmission.

The following step is to import data into the K-ecohub database. If a new protocol is defined, then the corresponding schema and tables are created by a data manager. When an importing job is initiated by the data manager, CSV files are automatically imported into the database.

The fifth step is data synthesis. To make the data easier to find, analyze and visualize, various predefined actions are taken in this step depending on the protocol that the submitted data follows. Currently all actions are defined through sets of SQL commands that are similar to the ETL (Extract, Transform, Load) processes. Normally major actions in this stage are creating summary metadata for analysis and visualization graphs.

The next step is reviewing, where a detailed review is provided by the field experts. Reviewers may examine data manually with the help of visualized charts created in the

previous steps. If the reviewers find any error, the submitted data is sent back to the data contributor for modification and resubmission.

The final step is to publish the data. A data manager assigns a DOI to the submitted data and publishes it so that it can be viewed, searched and analyzed by users.

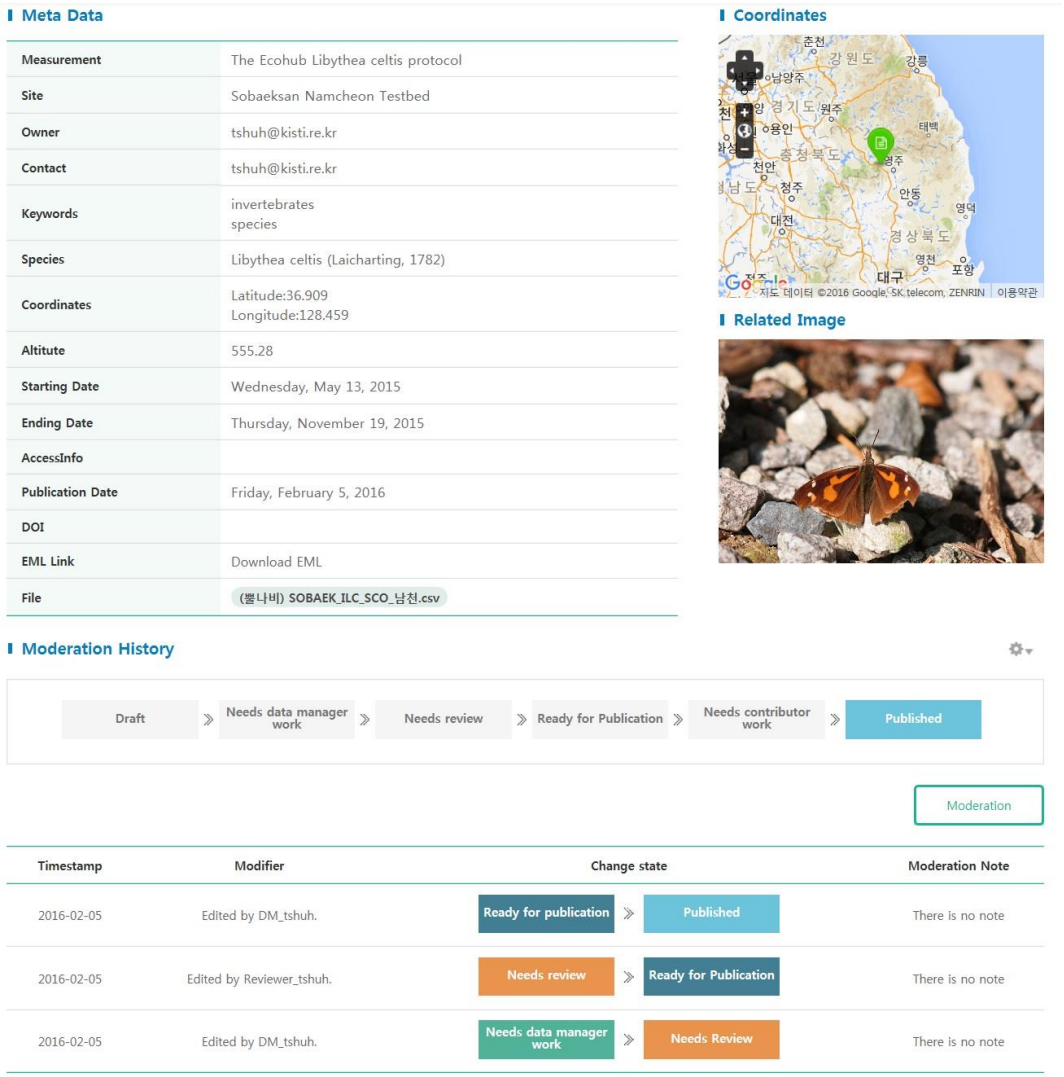

Figure 6. Provenance of Data Workflow in the K-ecohub System

Roles for data curation tasks in K-ecohub are as follows. A data contributor harvests data and screens them before submitting into the K-ecohub system. If any errors are found by the system, data manager, or reviewer, then the data contributor corrects data and resubmits. A data manager plays a central role in curating and managing the lifecycle of data. A reviewer then validates the submitted data utilizing the expertise he has in the field.

The K-ecohub system manages data in 6 stages. The “Draft” stage means that data is temporarily stored prior to submitting. “Need Data Manager Work” refers to the stage in which the data requires verification after it is submitted by a data contributor, or a reviewer has found an error and the data needs correction. The “Need Review” stage indicates that the data has to wait for a detailed review. “Need Contributor Work” means a stage of waiting for a correction when an error is detected. The “Ready for Publication” stage is where final actions such as a DOI assignment are taken by a data manager. After all these stages, the data reaches the “Published” stage. In general, the data is processed in

the order of “Draft”, “Need Data Manager Work”, “Need Review”,” Ready for Publication” and “Published”.

5. Conclusion

The K-ecohub system curates data to enhance data usability and discoverability in various curation activities: data validation, ingestion, synthesis, and expert reviews. It promotes collaboration and facilitates contribution from experts in the LTER field and also provides ways to automate and customize curation activities depending on the data types. Directions for future research include supporting data ontologies, and providing various ways to create customized curation workflows in K-ecohub.

Acknowledgments

This study was supported both by Korea Ministry of Environment as "The Eco-technopia 21 project" (Grant No.: 2014000210004) and by the EDISON (EDucation-research Integration through Simulation On the Net) Program through the NRF funded by the Ministry of Science, ICT & Future Planning (No. NRF-2011-0020576).

References

[1] LTER, “The Long Term Ecological Research Network”, https://lternet.edu
[2] S. Ahn, S. Hwang, J. Jang, S. Lee and S. Kim, “LTER Platform: Requirement, Technology and Trend”, International Journal of Software Engineering and Its Applications, vol. 9, no. 5, **(2015)**.
[3] C. Berkley, M. Jones, J. Bojilova, and D. Higgins, “Metacat: a schema-independent XML database system”, 13th Scientific and Statistical Database Management Conference, **(2001)**.
[4] M. Servilla, J. Brunt, I. S. Gil and D. Costa, “PASTA: A Network-level Architecture Design for Generating Synthetic Data Products in the LTER Network”, Ecological Informatics, **(2006)**.
[5] C. Gries, “The Drupal Environmental Information Management System Provides Standardization, Flexibility and a Platform for Collaboration”, AGU Fall Meeting Abstracts, vol. 1, **(2013)**.
[6] A. J. M. Lane, “The UK environmental change network database: An integrated information resource for long-term monitoring and research”, Journal of Environmental Management, vol. 51, no. 1, **(1997)**, pp. 87-105.
[7] B. Fu, S. Li, X. Yu, P. Yang, G. Yu, R. Feng and X. Zhuang, “Chinese ecosystem research network: Progress and perspectives”, Ecological Complexity, vol. 7, **(2010)**.
[8] TERN, “Terrestrial Ecosystem Research Network”, http://portal.tern.org.au/.
[9] A. Tokmakoff, B. Sparrow, D. Turner and A. Lowe, “AusPlots Rangelands Field Data Collection and Publication: Infrastructure for Ecological Monitoring”, IEEE 10th International Conference on e-Science, **(2014)**.
[10] J. Y. Kim, 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, vol. 34, no. 1, **(2011)**, pp. 3-10.
[11] S. Ahn, J. Jang and T. Huh, “Quality Assurance for the K-ecohub LTER Data”, Advanced Science and Technology Letters, vol. 129, **(2016)**, pp. 7-10.
[12] R. J. Miller, “Big Data Curation”, 20th International Conference on Management of Data, **(2014)**.
[13] E. H. Fegraus, S. Andelman, M. B. Jones and M. Schildhauer, “Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation,” Bulletin of the Ecological Society of America, vol. 86, no. 3, **(2005)**.

Authors

Sunil Ahn, is a research staff in the supercomputing center at Korea Institute of Science Technology and Information 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.

Taesang Huh, is a research staff in the supercomputing center at Korea Institute of Science Technology and Information in Korea. He is a Ph.D candidate in computer engineering at Chungnam National University (Korea).

Soonwook Hwang, received the B.S. degree in mathematics and the MS degree in computer science from Seoul National University (SNU), Korea, in 1990 and in 1995, respectively. He also received the Ph. D. degree in computer science from University of Southern California in 2003 under the supervision of Dr. Carl Kesselman, one of pioneers in Grid computing. He worked for Japanese National Grid Initiative (NAREGI) as a researcher, which is a Japanese National Grid project started in 2003 for five years, aiming at developing grid middleware for next-generation Cyber Science infrastructure. In 2006, he has joined Korea Institute of Science and Technology Information (KISTI) and has been a principal researcher of Supercomputing center. His research interests are in the areas of Grid computing, high throughput computing, Cloud storage, e-science and information system. Dr Hwang has been Editor-in-Chief of the Journal of Convergence Information Technology since 2009 and AMGA supervisor in European Middleware Initiative (EMI) since 2010.

Jihoon Jang, is principal researcher at the National Institute of Supercomputing and Networking (NISN), Korea Institute of Science and Technology Information (KISTI), where he does research on management of cyberinfrastructure.

Sung-Hee Lee, Lee received the M.S. in Innovation and Technology Management from KAIST, Korea, in 2006. He worked for KT Cooperation as a researcher in 1995 for 18 years. Since 2015, she has been a researcher at Korea Institute of Science and Technology Information (KISTI).

[i] *Corresponding author: Sunghee Lee.
E-mail: sunghee.lee@kisti.re.kr

