Quality Assurance for the K-ecohub LTER Data

Sunil Ahn, Jihoon Jang, Taesang Huh

Division of Supercomputing, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea {siahn, jangoq, tshuh}@kisti.re.kr

Abstract. K-ecohub, a pilot repository of national long-term ecological data in Korea, has been developed to solve data fragmentation issues in the KNLTER project and utilize long-term ecological data in an integrated manner. One essential element in the long-term ecological research is data quality assurance because data which do not go through quality control are far less reliable. The paper presents data processing and quality assurance plans in K-ecohub.

Keywords: LTER, Research environment, Quality Assurance, Data Processing, Long Term Ecology

1 Introduction

To collect LTER data for over a long period of time and analyze them comparatively, it is essential to have a robust cyber infrastructure which monitors and manages data on ecological changes. Several countries around the world have developed and been operating these kind of systems through which they can keep collecting and managing data on ecological changes, including the Provenance Aware Synthesis Tracking Architecture (PASTA, the U.S.A) [1], Metacat [2], Drupal Ecological Information Management System (DEIMS, EU) [3], Terrestrial Ecosystem Research Network (TERN, Australia) [4], Australian Ecological Knowledge and Observation System (AEKOS) [5], Environmental Change Network (ECN) [6] and Chinese Ecological Research Network (CERN, China) [8].

In the Republic of Korea, long-term ecological data have been collected through the KNLTER project [9] since 2004. This project aimed to establish a long-term scientific ecological change management system for the Korean Peninsula and utilize it in planning for biodiversity conservation. Despite diverse contributions to the ecological researches, the KNLTER project was suspended in 2013 due to several problems; poor planning and lack of consensus to derive common survey and analysis items; absence of shared quality assurance plan.

To solve these problems, Korean government has promoted the development of ‘K-ecohub’ for the efficient management, retrieval and sharing of national long-term ecological data. K-ecohub is an integrated repository of national long-term ecological
data, developed for the purpose of solving the following problems: absence of a common protocol in the current KNLTER project; data fragmentation; absence of quality assurance; absence of a repository for efficient data integration; poor linkage with the global LTER data.

One essential element in the long-term ecological research is data quality assurance because data which do not go through quality control are far less reliable. The paper presents data processing and quality assurance plans in K-ecohub.

2 Quality Assurance in K-ecohub

In the data model of K-ecohub, it defines a standardized protocol to collect the data which are applicable to various sites just like the ECN. This protocol ensures consistency in data collection by defining the type and unit of the data to be measured as well as the subject to be monitored, contents and methods. The data collected according to this kind of standard protocol are ingested into the K-ecohub system based on preset schema, and whether or not data quality is controlled in a consistent manner is verified.

A user’s roles for data management in K-ecohub are as follows: A data manager manages a data life cycle. The data contributor for each site provides the collected data to K-ecohub. A reviewer verifies the data entered into K-ecohub. Figure 1 reveals a quality assurance workflow when datasets are submitted into K-ecohub. For quality control, six data states are managed. “Draft” is a temporarily stored state prior to the completion of the dataset input in the specified format by a data contributor. “Need Data Manager Work” indicates data that either requires verification after the input process is completed or handling of an error in a data set by a data manager. “Need Review” refers to the state of waiting for a reviewer’s detailed review. “Need Contributor Work” means a state of waiting for a revision by a data contributor after an error is detected. “Ready for Publication” is a state where a data manager takes care of finishing touches (DOI issuance, etc.) for the final publication of datasets. “Published” indicates a state where documents are finally published. In general, datasets are entered in the order of Draft, Need Data Manager Work, Need Review and Ready for Publication. Finally, they are turned into a “Published” state.

In this process, key roles in data verification are performed by a data manager, a reviewer and the automated system. Automated validation is executed at the last stage of “Draft.” Whether or not the essential values of the metadata are entered, and if the scope of the value is acceptable are automatically verified through web forms. In addition, the type, scope and category of the entered data are automatically verified.

A data manager manually verifies metadata and data entered by a data contributor during the “Need Data Manager Work” phase. If no error is found in the data, data are prepared for data synthesis and visualization. Then, the data set is converted into “Need Review” where a detailed review by a reviewer is required. If an error is found, the data set is revised into “Need Contributor Work” which necessitates revision on the data set by a data contributor.
A reviewer performs detailed verification of the data at the “Need Review” stage. A semantic verification along with visualized data created by a data manager is performed. Unless an error is detected in the data, the data set is set to “Ready for Publication.” If an error is found, the dataset is converted back into “Need Contributor Work.”

Fig. 1. Data processing and validation workflow for quality assurance in K-ecohub.

3 Conclusion

Several countries around the world have been collecting and managing data on ecological changes for the long-term ecological research and keep trying to ensure data quality and integrity. K-ecohub is an integrated repository of national long-term ecological data, and this paper presented data processing and quality assurance plans in K-ecohub.

K-ecohub supports several steps and roles in the publication workflow. It provides automated data and metadata validation in addition to semantic verification through peer-review by experts. These quality assurance procedures and standardized protocols are expected to contribute significantly to provide greater reliability in the future long-term ecological data collected.
Acknowledgements. This study was supported both by Korea Ministry of Environment as "The Eco-technopia 21 project" (Grant No.: 2014000210004) and by KISTI.

References