PNIE PROCEEDINGS OF THE NATIONAL INSTITUTE OF ECOLOGY OF THE REPUBLIC OF KOREA

Original Article

PNIE 2021;2(2):139-14
https://doi.org/10.22920/PNIE.2021.2.2.139
pISSN 2765-2203, eISSN 2765-2211

Analysis of the Current Status of Data Repositories in the Field of Ecological Research

Suntae Kim*

Department of Library and Information Science, Jeonbuk National University, Jeonju, Korea

ABSTRACT

In this study, data repository information registered in re3data (re3data.org), a research data registry, was collected. Based on collected data, the current status was analyzed for 354 repositories (approximately 14% of total repositories) in the field using keywords in the ecological field suggested by two experts. Major metadata formats used to describe data in ecological research data repositories include Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC/CSDGM), Dublin Core, ISO 19115, Ecological Metadata Language (EML), Directory Interchange Format (DIF), Darwin Core, Data Documentation Initiative (DDI), and DataCite Metadata Schema. The number of ecological repositories according to country is 102 in the US, 34 in Germany, 31 in Canada, and one in Korea. A total of 771 non-profit organizations and 12 for-profit organizations are involved in the construction of the ecological field research data repository. Data version control ratio of the ecological field research data repositories registered in re3data was analyzed to be somewhat higher (86.6%) than the total ratio (83.9%). Results of this study can be used to establish policies to build and operate a research data repository in the ecological field.

Keywords: Data Repository, Ecology, Ecological Data Repository, Research Data Repository

Introduction

Regulations on Management, etc. of National R&D Projects were recently revised to take effect from September 1, 2019. Contents of research data included in Regulations on Management, etc. of National R&D Projects were included in the National R&D Innovation Act which was in effect from January 2021. According to this law, in the case of research and development projects that the head of a central administrative agency deems necessary, when selecting a research and development project, the faithfulness of research data production, preservation, and management according to the data management plan and the possibility of joint use should be reviewed.

Therefore, researchers submitting project plans must manage research data produced in the research process in a data repository and establish a plan to disclose it to the outside. In addition, as data journals grow rapidly, raw data described in data papers must be managed in data repositories. For this reason, research data repositories are being built and operated by various organizations. A registry service that registers such research data repositories so that they can be easily found is operated. For the above reason, the registry arose from two separate projects, re3data.org and DataBib. It is now managed by DataCite (Klump & Huber, 2017). This move is the same in the field of ecological research. The objective of this study was to analyze the current status of research data repositories in the ecological field. The current level of operation was examined in terms of metadata, the status of repositories by country, and version management of research data to derive implications.

Received March 16, 2021; **Revised** April 1, 2021;
Accepted April 1, 2021

***Corresponding author:** Suntae Kim
e-mail kim.suntae@jbnu.ac.kr
https://orcid.org/0000-0002-8726-6367

Materials and Methods

Theoretical background

Research data repository and re3data

Data repositories play increasingly larger role in academic research. Reliable storage and fair re-use of the re-

search data are of paramount importance in terms academic ethics, and thus become an imperative for any research institution (Kim & Choi, 2017). Pampel *et al*. (2013) have classified and presented types of research data repository into institutional research data repositories, disciplinary research data repositories, multidisciplinary research data repositories, and project specific research data repositories. Tropical Ecology Assessment and Monitoring Network (TEAM), Australian Drosophila Ecology and Evolution Resource (ADEER), and Neotoma Paleoecology Database are representative data repositories in the ecological field. TEAM repository is identified as r3d100010606 in re3data. It is devoted to monitoring long-term trends in biodiversity, land cover change, climate, and ecosystem services in tropical forests. ADEER from the Hoffmann lab and other contributors is identified as r3d100011630 in re3data. It is a nationally significant life science collection. The Drosophila Clinal Data Collection contains data on populations along the eastern coast of Australia. It remains an excellent resource for understanding past and future evolutionary responses to climate change. Neotoma is identified as r3d100011761 in re3data. It is a multiproxy paleoecological database that covers the Pliocene-Quaternary, including modern microfossil samples.This database is an international collaborative effort among individuals from 19 institutions, representing multiple constituent databases. There are over 20 data-types within the Neotoma Paleoecological Database, including pollen microfossils, plant macrofossils, vertebrate fauna, diatoms, charcoal, biomarkers, ostracodes, physical sedimentology and water chemistry (Scientific Data, 2021).

Meanwhile, re3data was a research project funded by the German Research Foundation (DFG) from 2012 until 2015 to create a Registry of Research Data Repositories called re3data (Kindling *et al*., 2017). The main goal of re3data is to offer researchers orientation in the heterogeneous Landscape of RDR. Researchers are both data producers and data users. Other target groups are research funders and infrastructure facilities such as data centers and academic libraries (Pampel *et al*., 2013). As of December 21, 2020, 2,607 data repositories were registered in re3data.org.

Publisher and data repository

As open access publishing models are diversifying around the world, data journal publications are increasing by various actors (Jung *et al*., 2020). When a researcher submits a manuscript to a data journal, sometimes they are guided to deposit raw data in a separate data repository. With this background, the importance of data repositories is increasing day by day.

Nature Publisher publishes Scientific Data journals. As a journal that publishes research data, Nature recommends publishing data papers and submitting research data to reliable data repositories. In other words, Scientific Data mandates the release of datasets accompanying our Da ta Descriptors. However, we do not host data ourselves. Instead, we ask authors to submit datasets to an appropriate public data repository. Data should be submitted to discipline-specific, community-recognized repositories where possible, or to generalist repositories if no suitable community resource is available.

Table 1 shows ecological field data repositories recommended by Scientific Data Journal and the metadata information registered in re3data for each repository. Nature Publisher recommends raw da ta to be submitted to Global Biodiversity Information Facility (GBIF), The Knowledge Network for Biocomplexity (KNB), Environmental Data Initiative, and Australian Ecological Knowledge and Observation System (AEKOS) for data papers submitted to Scientific Data Journal in the field of ecology. It was confirmed that all repositories were registered in re3data, a global data registry service.

Table 1. Ecological data repositories recommended by Scientific Data Journal and metadata information registered at re3data.org for each repository

Data Repository recommended by Scientific Data Journal	Repository Name	Repository URL	Size	Start Date	Entry Date	Nation Codes
Global Biodiversity Information Facility (GBIF)	Global Biodiversity Information Facility	https://www.gbif.org/	964.313.520 occurrence records; 37.614 datasets;	2001	2013-01-31	DNK
The Knowledge Network for Biocomplexity (KNB)	KNB Data Repository	https://knb.ecoinformatics.org/	26.886 public datasets	1999	2012-10-02	USA,USA,USA
Environmental Data Initiative (formerly LTER Network Information System Data Portal)	Environmental Data Initiative Repository	https://portal.edirepository.org/nis/home.jsp		1980	2013-05-13	USA,USA,USA,USA, USA,USA
AEKOS - TERN Ecoinformatics	AEKOS Data Portal	http://www.aekos.org.au/index.html#/home	3.432.272 records	2011	2015-01-13	AUS,AUS,AUS

Data collection and analysis

To collect re3data's data, Crawler program developed in 2017 was used. Collected data (totally 2,607 records) were stored in a relational database and evaluated against the proposed re3data schema. Since 2017, the problem caused by the diversity of the length and type of data for each item provided by re3data has been resolved. The crawler operating environment is as follows.

- OS: Windows 10 Pro
- Database Server and Client: MySQL Server 8 / MySQL Workbench version 8
- IDE: Eclipse Java EE IDE / Luna Service Release 1 (4.4.1) / build 20140925-1800
- Programming Language and VM: Java 1.8.0_144
- Analysis SQL Client: SQLyog Community – MySQL GUI v13.0.1 (64bit)

Data collected from re3data were saved in the MySQL database. After that, analysis was performed using SQLyog, an SQL client program. Current status of the repository in the ecological field and the format of metadata were investigated and analyzed. Research and analysis were conducted for the current state of ecological repositories and version control of research data by country.

Results

Repository distribution in the ecological sector

As of December 21, 2020, 2,607 data repositories were registered in re3data.org. Among these repositories registered in re3data.org, 9 repositories registered in Korea were identified, including the one operated by Seoul National University College of Veterinary Medicine (https://vet.snu.ac.kr/en). Among all data repositories registered in re3data.org, the number of search results in the repository name for the ecology keyword was 3, the number of search results in the repository description part was 18, and the number of search results in the keyword registered by the repository was 78. In this study, an expanded keyword list (Ecology, species, restoration, biodiversity, ecosystem, wildlife, ecological, eco-tourism, ecoinformatics, climate, change, ecological database) was used to identify ecological repositories with the help of two experts. To identify ecological repositories, search was performed using one or more keywords from the list of expanded keywords. The number of search results was 26 when the search was performed against the repository name, 207 when the search was performed against the repository description, and 241 when the search was performed againstthe keyword registered in the repository (Table 2). Excluding duplicates, the total number of ecological reports was 354, accounting for about 14% of the total numberof repositories registered in re3data. In this study, repositories to be analyzed were finally determined through the above steps.above steps.

Metadata format of the ecological field repository

Major metadata formats used in ecological repositories included Federal Geographic Data Committee Content (EML), Directory Interchange Format (DIF), Darwin Core, Data Documentation Initiative (DDI), and DataCite Metadata Schema. These types of metadata format for the entire ecological field were analyzed (a total of 19 cases). Five cases were surveyed as 'other' metadata formats and four of them were judged with ABCD-access criteria for biological collection data as a result of analyzing their actual URL (http://www.dcc.ac.uk/resources/metadata-standards/abcd-access-biological-collection-data).

Table 3 below shows metadata format used in the ecological field research data repository registered in re3data. The number of registered metadata format registrations was 155 (43.8%) out of a total of 353 repositories analyzed.

Number of ecological repositories by country

As a result of analyzing the ranking by the number of countries operating ecological repositories, the United States, which operates 102 repositories, ranks the first. Germany, which operates 34 repositories, ranks the second. Canada, which operates 31 repositories, ranks the third. Japan, which operates 7 repositories, ranks the 7th. Korea is operating one ecological repository. Fig. 1 shows the above information schematically.

Meanwhile, the number of repositories depending on whether the institution was profitable or not was surveyed. A total of 771 non-profit organizations and 12 for-profit organizations are participating in the operation of the ecological research data repository. In the case of Korea, two non-profit organizations ('Korea Science & Engineering Foundation' and 'Seoul National University, College of Veterinary Medicine') were surveyed to build an ecological research data repository.

Research data version control status

Research data version management can provide confidence in the data to other researchers who want to use the research data. In addition, version management of research data guarantees a systematic preservation process. It is judged as a function that must be provided by an institution operating a research data repository. Table 4 below shows the current status of ecological research data repositories registered in re3data managing the version of research data. As of March 2021, it was confirmed that 83.9% of the total repositories (n = 2,62,607) registered in re3data and 86.6% of the ecological repositories (n = 354) were managing the research data version.

Table 2. The number of data repositories registered according to the location where the analysis keyword appears

division	The number of appearances in the repository name	The number of occurrences of keyword in Description	The number of occurrences in the registered keyword
Before keyword expansion	3	18	78
After keyword expansion	26	207	241

Table 3. Metadata format used in the ecological field research data repository registered in re3data

Metadata Format	Count
ISO 19115	25
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata	21
EML-Ecological Metadata Language	19
Repository-Developed Metadata Schemas	19
Dublin Core	15
Darwin Core	14
DataCite Metadata Schema	13
ABCD - Access to Biological Collection Data	7
DIF - Directory Interchange Format	5
DDI - Data Documentation Initiative	5
CF (Climate and Forecast) Metadata Conventions	5
RDF Data Cube Vocabulary	1
MIBBI - Minimum Information for Biological and Biomedical Investigations	1
Genome Metadata	1
CIM - Common Information Model	1
DCAT - Data Catalog Vocabulary	1
ISA-Tab	1

Table 4. Data version control ratio among ecological research data repositories registered in re3data

Number of cases Null / except Null / Yes / No / Total	Ratio (Inc. Null) Null / Yes+No / Yes / No	Ratio (except Null) Yes / No
1366 / 1241 / 1041 / 200 / 2607	52.4 / 47.6 / 39.9 / 7.7	83.9 / 16.1
167 / 187 / 162 / 25 / 354	47.2 / 52.8 / 45.8 / 7.1	86.6 / 13.4

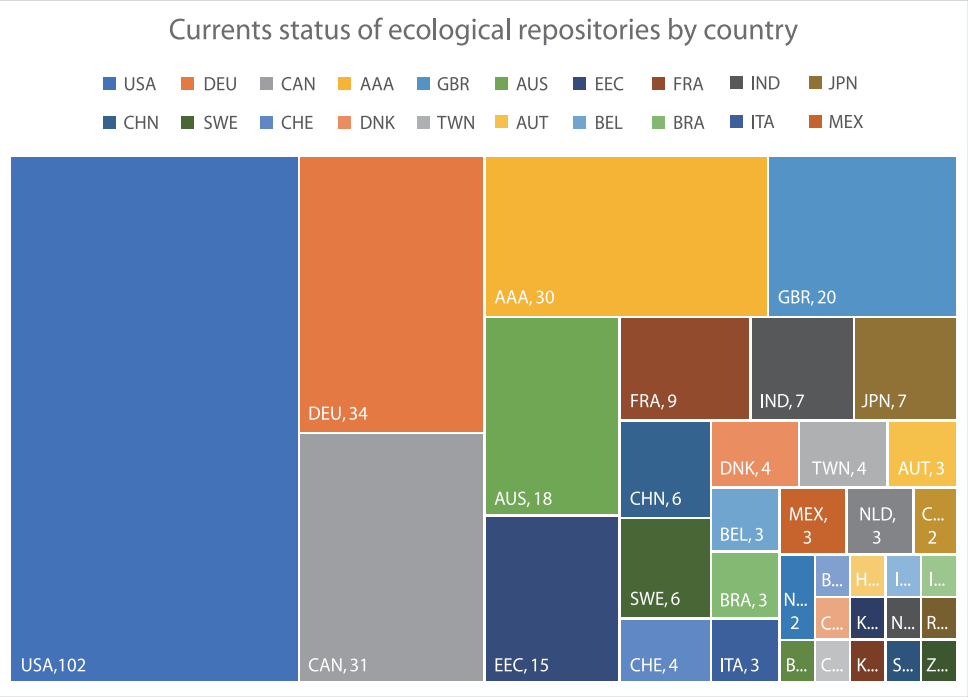

Fig. 1. Current status of ecological repositories by country (as of 2020. 12. 21).

Discussion

In this study, data repository information registered in re3data, a research data registry, was collected. Based on the collected data, the current status was analyzed for 354 repositories (approximately 14%) in the field by using keywords suggested by two experts in the ecological field. Main metadata formats used to describe data in ecological research data repositories have emerged as ISO 19115, FGDC, EML, Dublin Core, Darwin Core, and so on. As for the number of ecological repositories by country, the US, Germany, and Canada have 102, 34, and 31 repositories, respectively. A total of 771 non-profit organizations and 12 for-profit organizations are involved in the construction of the ecological field research data repository. The data version control ratio of the ecological field research data repositories registered in re3data was analyzed to be somewhat higher (86.6%) than the total ratio (83.9%). Results of this study can be used to establish policies to build and operate a research data repository in the ecological field. This is a time when the open science movement for the reuse of research data is actively unfolding in the era of data-intensive science. In this flow of research culture, the role of research data repositories is becoming very important. Korea's ecological research data repositories should be built in line with the international level. This study examined the current status and level of international research data repositories in the ecological field. Results of this stud could be used as benchmarking data by organizations that build and plan research data repositories in Korea.

Conflict of interest

The author has declared that no competing interests exist.

References

Jung, Y., Kwon, O., Kim, K., Kim, S., Seo, T., and Kim, S. (2020). A study on the strategies for publishing data journals in the field of ecology: Focused on K institution. *Journal of Korean Library and Information Science Society*, 51, 83-100. doi:10.16981/kliss.51.4.202012.83

Kim, S., and Choi, M. (2017). Registry metadata quality assessment by the example of re3data.org schema. *International Journal of Knowledge Content Development & Technology*, 7, 41-51. doi:10.5865/IJKCT.2017.7.2.041

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., *et al*. (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, 23. doi:10.1045/march2017-kindling.

Klump, J., and Huber, R. (2017). 20 years of persistent identifiers – Which systems are here to stay? *Data Science Journal*, 16, 9. doi:10.5334/dsj-2017-009

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump J., *et al*. (2013). Making research data repositories visible: The re3data.org Registry. *PLoS ONE*, 8, e78080. doi:10.1371/journal.pone.0078080

Scientific Data. (2021). *Recommended Data Repositories*. Retrieved December 13, 2020 from https://www.nature.com/sdata/policies/repositories