Journal of Ecology and Environment

NOTES Open Access

Developing a semi-automatic data conversion tool for Korean ecological data standardization

Hyeonjeong Lee[1], Hoseok Jung[1], Miyoung Shin[1*] and Ohseok Kwon[2]

Abstract

Recently, great demands are rising around the globe for monitoring and studying of long-term ecological changes. To go with the stream, many researchers in South Korea have attempted to share and integrate ecological data for practical use. Although some achievements were made in the meantime, we still have to overcome a big obstacle that existing ecological data in South Korea are mostly spread all over the country in various formats of computer files. In this study, we aim to handle the situation by developing a semi-automatic data conversion tool for Korean ecological data standardization, based on some predefined protocols for ecological data collection and management. The current implementation of this tool works on only five species (*libythea celtis*, spittle bugs, mosquitoes, *pinus*, and *quercus mongolica*), helping data managers to quickly and efficiently obtain a standardized format of ecological data from raw collection data. With this tool, the procedure of data conversion is divided into four steps: data file and protocol selection step, species selection step, attribute mapping step, and data standardization step. To find the usability of this tool, we utilized it to conduct the standardization of raw five species data collected from six different observatory sites of Korean National Parks. As a result, we could obtain a common form of standardized data in a relatively short time. With the help of this tool, various ecological data could be easily integrated into the nationwide common platform, providing broad applicability towards solving many issues in ecological and environmental system.

Keywords: Ecological data, Data standardization, Data conversion, Program, Tool

Introduction

It is important to share and integrate ecological data for monitoring and studying of long-term ecological changes (Brunt et al. 2002). Currently, however, domestic data in South Korea are spread over numerous research sites, institutions, and individual researchers. Even there has been no common protocol for ecological data collection and management; the data are mainly kept in a variety of forms. For this reason, existing data are difficult to integrate, analyze, and manage for long-term ecological research, so it is very necessary to standardize domestic ecological data in a common form for data integration and further analyses (Michener et al. 2012, Bonet et al. 2014).

Until now, long-term ecological data have been globally collected in each country according to its own protocols, while being maintained in large databases in the form of Ecological Metadata Language (EML) (Fegraus et al. 2005). In particular, various long-term environmental monitoring projects, including Environmental Change Network (ECN) (Morecroft et al. 2009), the National Ecological Observatory Network (NEON) (Keller et al. 2008), and the Long-Term Ecological Research network (LTER) (San Gil et al. 2009), are providing large volume of ecological data easily accessible to the public. To follow such trends, Korea is also building a unified ecological data integration network. For this purpose, there is a need to convert already collected raw data into common form, as well as to collect new data with common protocols.

In this study, we developed a semi-automatic ecological data conversion tool that can help ecologists

* Correspondence: shinmy@knu.ac.kr
[1]Bio-Intelligence & Data Mining Laboratory, Graduate School of Electronics Engineering, Kyungpook National University, Daegu 41566, Korea
Full list of author information is available at the end of the article

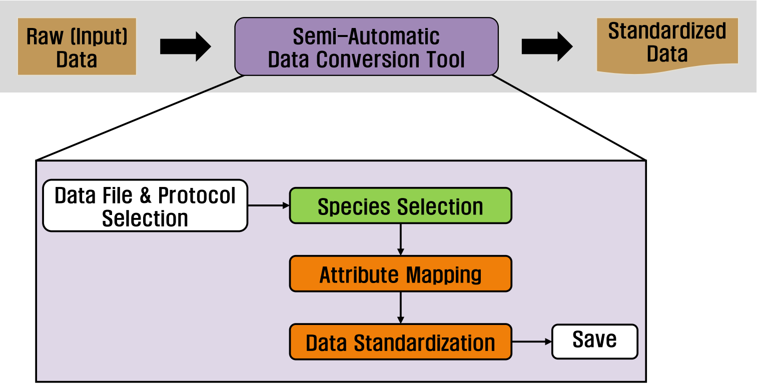


Fig. 1 The overall workflow of semi-automatic ecological data conversion tool

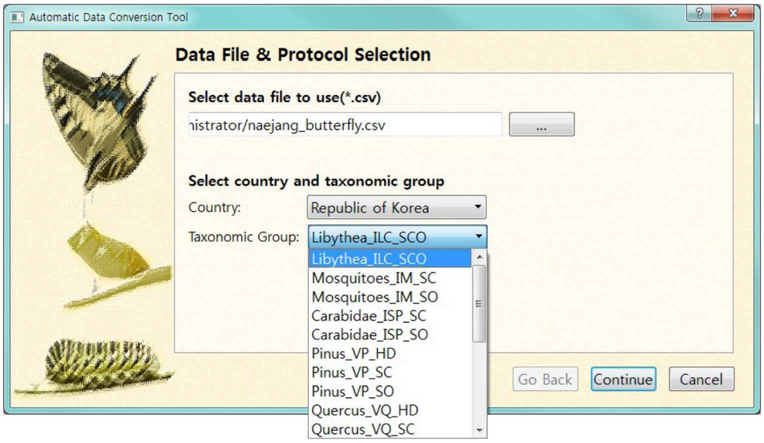
to standardize ecological data more easily and efficiently in a relatively short time, while keeping the inherent meaning of the data. The data conversion was done based on some predefined protocols for data collection and management. Figure 1 summarizes the overall workflow of conversion procedure in our program.

Materials and methods

Ecological data are mostly stored in text-based tables. Each row in the table represents a record that contains the values of many attributes (or characteristics) for target species. Each column corresponds to an attribute of the same data type and unit. For example, an attribute of "search date" includes the date when the raw data were collected, usually given in the format of YYYY-MM-DD, DD-MM-YY, and so on. With our tool, the raw data is standardized by following the four steps: (1) data file and protocol selection step, (2) species selection step, (3) attribute mapping step, and (4) data standardization step.

The first step of data file and protocol selection is to upload raw data file to be converted and select predefined protocols which define standard attributes and data types for target species (see Fig. 2). In the present version of the tool, only csv files are allowed for raw data files.

Next, the second step is to specify target species to be converted from raw data files. This is to filter out and convert only specific (target) species data matched with the chosen protocol, in case that the raw data file contains a number of species. If the raw data include only one species corresponding to the protocol, this step can be skipped. The user interface for this step to choose a list of target species that should be extracted from raw data is presented as shown in Fig. 3. Here, users can find a certain attribute containing some specific names of target species and add a particular species name to the "selected


Automatic Data Conversion Tool

Data File & Protocol Selection

Select data file to use(*.csv)

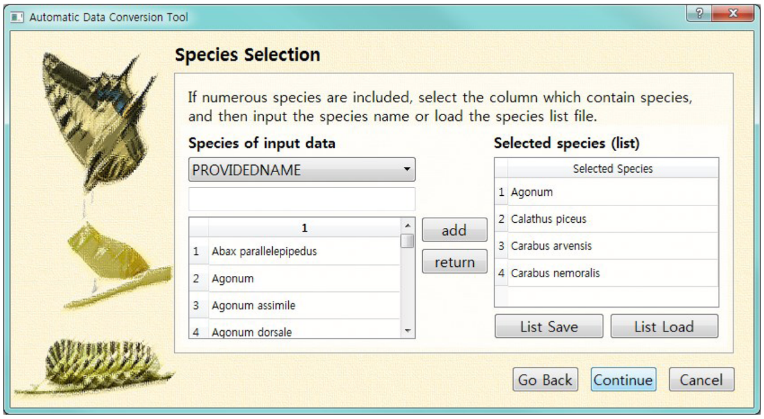
ıistrator/naejang_butterfly.csv

Select country and taxonomic group

Country: Republic of Korea

Taxonomic Group: Libythea_ILC_SCO

Fig. 2 Typical user interface used in the first step of data file and protocol selection



Species Selection

If numerous species are included, select the column which contain species, and then input the species name or load the species list file.

Species of input data: PROVIDEDNAME

	1
1	Abax parallelepipedus
2	Agonum
3	Agonum assimile
4	Agonum dorsale

Selected species (list)

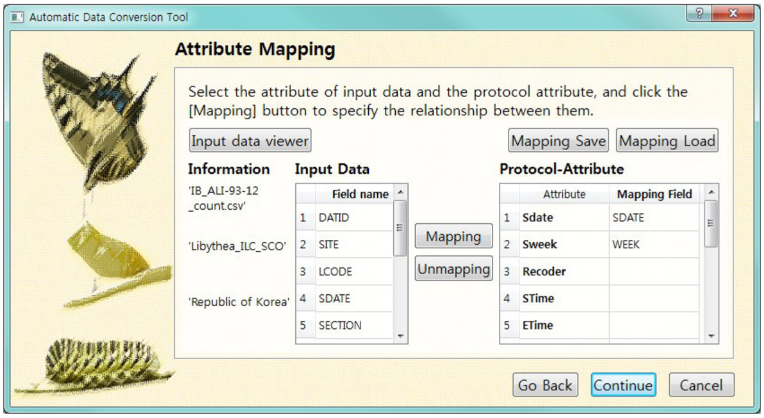
	Selected Species
1	Agonum
2	Calathus piceus
3	Carabus arvensis
4	Carabus nemoralis

Fig. 3 User interface of species selection step

species list”. Like this, users can selectively convert only a part of raw data matched with the chosen protocol. For user convenience, we provide the function of uploading a list of species names to be converted, which makes it easier and faster to select a number of species.

Then, in the step of attribute mapping, the relations between raw data attributes and standard attributes in the protocol need to be specified by users. To this end, users should specify which attributes in raw data are matched with which standard attributes defined in the protocol. Once the relation between the two attributes is specified, in Fig. 4, the “mapping” button of the screen can be pressed to realize the mapping into the data conversion procedure. Non-selected raw data attributes are excluded from the subsequent conversion process. The mapping list between the two attributes can also be allowed to use for convenience.

In the final step, data type and unit of each attribute can be properly transformed into a standardized format. For this purpose, we provide several functions like concatenation, separation, substitution, date conversion, unit conversion, and editing function (Fig. 5). Specifically, the concatenation function can be used to merge values in two or more attributes into one new value. We can insert a text or symbol as a delimiter when combining multiple values. The separation function divides a string into several chunks. For example, by the separation, the attribute of “search period” can be divided into two attributes of the “search start date” and “search end date.” The substitution function replaces certain values with different values, e.g., texts, numbers, delimiters, or symbols. The function of date conversion can be utilized to specify the desirable format of search date. For example, this function separates search date into three parts as day, month, and year,



Attribute Mapping

Select the attribute of input data and the protocol attribute, and click the [Mapping] button to specify the relationship between them.

Information

'IB_ALI-93-12 _count.csv'

'Libythea_ILC_SCO'

'Republic of Korea'

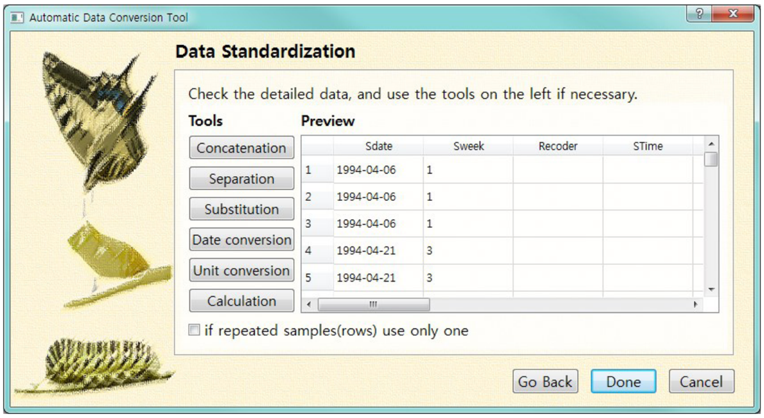
Input Data

	Field name
1	DATID
2	SITE
3	LCODE
4	SDATE
5	SECTION

Protocol-Attribute

	Attribute	Mapping Field
1	Sdate	SDATE
2	Sweek	WEEK
3	Recoder	
4	STime	
5	ETime	

Fig. 4 User interface of attribute mapping step



Automatic Data Conversion Tool

Data Standardization

Check the detailed data, and use the tools on the left if necessary.

Tools

Preview

	Sdate	Sweek	Recoder	STime
1	1994-04-06	1		
2	1994-04-06	1		
3	1994-04-06	1		
4	1994-04-21	3		
5	1994-04-21	3		

□ if repeated samples(rows) use only one

Fig. 5 User interface of data standardization step

and then rearranges them to the desired order such as YYYY-MM-DD. Unit conversion is to change the data unit, and editing function is to transform numerical data by using a formula for computation. At the end, the standardized data are saved into a new csv file in the table form.

Results and discussion

Our semi-automatic data conversion tool is a software of desktop application that works on Windows and Macintoshes. It helps ecologists to easily and efficiently create standardized data from raw collection data. To find the usability of our tool, we performed the data standardization with the six datasets from six observatory sites located in Korea National Park (for more information about the dataset, refer to Table 1). For this purpose, we need some predefined protocols about five kinds of indicator species, selected by the long-term ecological research of Kyungpook National University in Korea (refer to Table 2 for details). As results, overall, each raw data that varies widely in data types and terms was successfully standardized according to predefined protocols (refer to Table 3). For instance, search period was divided into search start date and search end date, and search date such as 01-MAY-2010 was converted to 2010-05-01, using separation and date conversion functions. The number of records that was converted according to SC protocols is equal to or smaller than that of the original raw data, because the SC protocol contain only

Table 1 Datasets of Korea National Parks used in this study

Data resource (site)	Data file to apply the protocol
Sobaeksan	Insect, plant
Jirisan	Insect, plant
Seoraksan	Insect, plant
Naejangsan	Butterfly, insect, plant
Mudeungsan	Butterfly, flying insect, insect, plant
Odaesan	Butterfly, flying insect, insect, plant

Table 2 Protocols of five species (measurements) used in this study

Species (measurement)	Protocol	Attribute
Libythea celtis	SCO	Search date, search week, recorder, start time, end time, temperature, humidity, wind speed, wind direction, weather, count, reference, description
Spittle bugs (*Carabidae*)/Mosquitoes	SC	Start date, start time, search week, recorder, end date, end time, hour, maximum temperature, minimum temperature, average temperature, humidity, wind speed, wind direction, weather, reference, description
	SO	Start date, trap ID, species, count, description
Pinus/Quercus mongolica	SC	Search date, recorder, topography, slope, fallen leaves, rock exposure, tree layer height/coverage/dominant, subtree layer height, subtree layer coverage, subtree layer dominant, shrub layer height, shrub layer coverage, shrub layer dominant, herb layer height, herb layer coverage, herb layer dominant, maximum dominant DBH, minimum dominant DBH, average dominant DBH, reference, description
	SO	Search date, layer, species, cover rate, description
	HD	Search date, plot, tree ID, species, DBH, height, vitality, description

SC survey condition, *SO* species observed, and *HD* height and DBH of vegetation

Table 3 Data conversion results from six datasets of Korea National Parks

Site	Dataset	Year(s)	Before		Species	Protocol	After	
			Attributes	Records			Attributes	Records
Sobaeksan	Insect	2007-2011	35 (19)	27	*Libythea celtis*	SCO	13 (7)	27
			35 (19)	24	Spittle bugs (Carabidae)	SC	16 (6)	15
						SO	5 (4)	24
	Plant	2007-2014	35 (30)	111	*Pinus*	SC	24 (3)	74
						SO	5 (3)	111
						HD	8 (4)	111
			35 (30)	92	*Quercus mongolica*	SC	24 (3)	62
						SO	5 (3)	92
						HD	8 (4)	92
Jirisan	Insect	2002-2012	35 (22)	18	*Libythea celtis*	SCO	13 (5)	18
			35 (22)	116	Spittle bugs (Carabidae)	SC	16 (5)	42
						SO	5 (5)	116
	Plant	2003-2014	35 (30)	159	*Pinus*	SC	24 (3)	102
						SO	5 (3)	159
						HD	8 (4)	159
			35 (30)	439	*Quercus mongolica*	SC	24 (3)	197
						SO	5 (3)	439
						HD	8 (4)	439
Seoraksan	Insect	2002-2014	35 (23)	10	*Libythea celtis*	SCO	13 (7)	10
			35 (23)	187	Spittle bugs (Carabidae)	SC	16 (6)	42
						SO	5 (5)	187
	Plant	2003-2015	35 (30)	81	*Pinus*	SC	24 (3)	43
						SO	5 (3)	81
						HD	8 (4)	81
			35 (30)	123	*Quercus mongolica*	SC	24 (3)	52
						SO	5 (3)	123
						HD	8 (4)	123
Naejangsan	Butterfly	2013	24 (18)	5	*Libythea celtis*	SCO	13 (4)	5
			24 (18)	1	Spittle bugs (Carabidae)	SC	16 (3)	1
						SO	5 (5)	1
	Insect	2013	23 (14)	26	Spittle bugs (Carabidae)	SC	16 (3)	9
						SO	5 (5)	26
	Plant	2013	28 (18)	4	*Pinus*	SC	24 (3)	4
						SO	5 (3)	4
						HD	8 (4)	4
			28 (18)	1	Quercus mongolica	SC	24 (3)	1
						SO	5 (3)	1
						HD	8 (4)	1
Mudeungsan	Butterfly	2013	23 (15)	24	*Libythea celtis*	SCO	13 (4)	24
	Flying insect	2013	26 (19)	3	*Libythea celtis*	SCO	13 (4)	3
	Insect	2013	24 (15)	27	Spittle bugs (Carabidae)	SC	16 (3)	7
						SO	5 (5)	27

Table 3 Data conversion results from six datasets of Korea National Parks *(Continued)*

	Plant	2013	28 (15)	15	*Pinus*	SC	24 (3)	11
						SO	5 (3)	15
						HD	8 (4)	15
			28 (15)	6	*Quercus mongolica*	SC	24 (3)	6
						SO	5 (3)	6
						HD	8 (4)	6
Odaesan	Butterfly	2013	23 (20)	3	*Libythea celtis*	SCO	13 (5)	3
	Flying insect	2013	30 (26)	25	Spittle bugs (Carabidae)	SC	16 (4)	20
						SO	5 (5)	25
	Insect	2013	23 (18)	217	Spittle bugs (Carabidae)	SC	16 (3)	37
						SO	5 (5)	217
	Plant	2013	28 (12)	8	*Pinus*	SC	24 (3)	5
						SO	5 (3)	8
						HD	8 (4)	8
			28 (12)	8	*Quercus mongolica*	SC	24 (3)	7
						SO	5 (3)	8
						HD	8 (4)	8

Numbers outside the parentheses indicate the total number of attributes defined in each data table, and the numbers inside the parentheses indicate how many attributes the real records contain

SC survey condition, *SO* species observed, and *HD* height and DBH of vegetation

search date and environment information, and several entities can be found in the same search date.

With the use of our tool, it is expected to possibly create standardized data of a common form in a relatively short time. Moreover, since the converted data can be stored and shared in the same format, it is possible to conduct comparative analysis with numerous ecological data more easily without regard to any organizations or project goals. Consequently, this tool can contribute to provide broad applicability to ecological and environmental data, such as towards uncovering the various effects of environmental factors on species.

Abbreviations

ECN: Environmental Change Network; EML: Ecological Metadata Language; LTER: Long-Term Ecological Research network; NEON: National Ecological Observatory Network

Acknowledgements

We would like to appreciate anonymous reviewers for their valuable comments on the manuscript.

Funding

This subject is supported by the Korea Ministry of Environment (MOE) as "Public Technology Program based on Environmental Policy (2014000210003)."

Availability of data and materials

Data are not publicly available to this article because they used under license for the current study.

Authors' contributions

HL carried out the studies, performed the analysis, and wrote/reviewed the manuscript. HJ carried out the studies. MS participated in the design of the study and wrote/reviewed the manuscript. OK participated in the design of the study. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Bio-Intelligence & Data Mining Laboratory, Graduate School of Electronics Engineering, Kyungpook National University, Daegu 41566, Korea. [2]School of Applied Bioscience, College of Agriculture and Life Sciences, Kyungpook National University, Daegu 41566, Korea.

Received: 14 February 2017 Accepted: 17 February 2017
Published online: 14 March 2017

References

Bonet, F. J., Pérez-Pérez, R., Benito, B. M., De Albuquerque, F. S., & Zamora, R. (2014). Documenting, storing, and executing models in ecology: a conceptual framework and real implementation in a global change monitoring program. *Environ Model Softw, 52*, 192–199.

Brunt, J. W., McCartney, P., Baker, K., & Stafford, S. G. (2002). *The future of ecoinformatics in long term ecological research* (Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics: SCI, pp. 14–18).

Fegraus, E. H., Andelman, S., Jones, M. B., Schildhauer, M. (2005). 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, 86*, 158–168.

Keller, M., Schimel, D. S., Hargrove, W. W., Hoffman, F. M. (2008). A continental strategy for the National Ecological Observatory Network. *The Ecological Society of America, 6*, 282–284.

Michener, W. K., et al. (2012). Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics, 11*, 5–15.

Morecroft, M. D., et al. (2009). The UK Environmental Change Network: emerging trends in the composition of plant and animal communities and the physical environment. *Biol Conserv, 142*, 2814–2832.

San Gil, I., et al. (2009). The Long-Term Ecological Research community metadata standardisation project: a progress report. *International Journal of Metadata, Semantics and Ontologies, 4*, 141–153.