Check for updates

© 2022 The Authors. The *Bulletin* published by Wiley Periodicals LLC on behalf of The Ecological Society of America.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Contributions

Publishing Ecological Data in a Repository: An Easy Workflow for Everyone

Kristin Vanderbilt[1], Jon Ide[2], Corinna Gries[2], Susanne Grossman-Clarke[2], Paul Hanson[2], Margaret O'Brien[3], Mark Servilla[1], Colin Smith[2], Robert Waide[1], and Kyle Zollo-Venecek[2]
[1]Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA
[2]Center for Limnology, University of Wisconsin, Madison, Wisconsin 53706, USA
[3]UCSB Marine Science Institute, University of California, Santa Barbara, California 93106, USA

Abstract

For many ecologists, publishing data in a data repository is a new, unfamiliar task. To reduce the learning curve, the Environmental Data Initiative has developed user-friendly software to make capturing and submitting data and metadata a simple process. In this article, we introduce ezEML and discuss use cases for researchers who publish data infrequently or information managers who regularly update multiple datasets.

Introduction

Many funders and publishers have recently adopted open access policies to facilitate the broadest reuse of research data and to support open science (ESA 2022, Wiley 2022). Ecologists are, therefore, faced with the question of where and how to publish their data to comply with these new requirements. While there are several repositories that accept environmental data, they differ with respect to the tools and support they provide for data contributors (Waide et al. 2017). One repository, operated by the Environmental Data Initiative (EDI), has developed a workflow that makes data publishing accessible for all ecologists. At the heart of this workflow is a new metadata editor called ezEML that offers a simple approach to capturing the rich metadata needed to support reproducible research and data reuse.

Metadata, the documentation that describes a dataset, are key to data reuse. Metadata provide details about the data that are essential for a data user to understand, such as the creators and contacts of the

Vanderbilt, K., J. Ide, C. Gries, S. Grossman-Clarke, P. Hanson, M. O'Brien, M. Servilla, C. Smith, R. Waide, and K. Zollo-Venecek. 2022. Publishing Ecological Data in a Repository: An Easy Workflow for Everyone. Bull Ecol Soc Am 103(4):e02018. https://doi.org/10.1002/bes2.2018

dataset, how, when, and where the data were collected, taxa studied, accessibility of the dataset, and any licenses that apply to data reuse. Metadata elements, like keywords, abstract, and title, aid in data discovery while grant numbers allow funding agencies to query a repository to find products of the research they sponsored. A digital object identifier (DOI) added to the metadata by the hosting data repository enables data users to cite the dataset so that its creators get credit for their efforts to preserve and share their data.

For submission to a data repository, metadata must often be declared in a structured, machine-readable format called a metadata standard. The Ecological Metadata Language (EML) is one such standard and is used widely in ecological repositories (Jones et al. 2006), including the EDI repository. The EML standard is specified as an XML (extensible markup language) schema, which poses significant barriers to data creators unfamiliar with XML.

Researchers wishing to publish their data are challenged to create the detailed metadata necessary to support data reuse and the corresponding EML needed by data repositories. While there are R programming packages such as EML (Boettiger et al. 2022) and EMLassemblyline (Smith 2022) to assist with this process, these tools are not helpful for non-R users. To facilitate metadata capture and EML generation by all ecologists, EDI has released ezEML, an openly accessible, online, form-based metadata editor that creates EML. In this article, we discuss the features of this ezEML editor, how it facilitates swift data publication in the EDI repository, and how it supports several data management use cases.

ezEML: typical workflow

The ezEML editor simplifies the workflow whereby ecologists create EML metadata and submit data and metadata to the EDI repository. A user must log in to ezEML with Google, ORCID, or GitHub credentials before any work can begin. Once authenticated, the user is presented with a menu of items to populate. The basic workflow is as simple as following the ezEML wizard through the forms, or the user may enter metadata in any order they wish via the menu. Fig. 1 shows important features of the ezEML interface.

While ezEML can be used to document most kinds of data, including spatial raster and vector, model code, and documents, most datasets are tables of data. ezEML makes detailed documentation of tables efficient. When a table is uploaded to ezEML, ezEML infers several characteristics of the table that are required metadata, including the filename, record delimiter, MD5 checksum, and the number of records. ezEML also infers the table's column names and column types (e.g., text, date time, numerical, categorical) and category codes. Information on each column can then be edited to define the column, categories, missing values, and units. Non-tabular data can be documented using the "Other Entity" selection in the menu.

The final product of a metadata editing session is an "ezEML data package," a .zip file that includes the EML file, all data objects associated with the package, and a manifest that lists the contents of the package. One ezEML data package may contain several types of files. This permits a collection of data files and supporting files (e.g., code to process and analyze the data, a .pdf map of study sites) to be published together. The contents of the ezEML data package are available to the user for closer inspection, if desired, by downloading the package. Researchers can collaborate on metadata development

23276096, 2022, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1002/bes2.2018, Wiley Online Library on [27/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

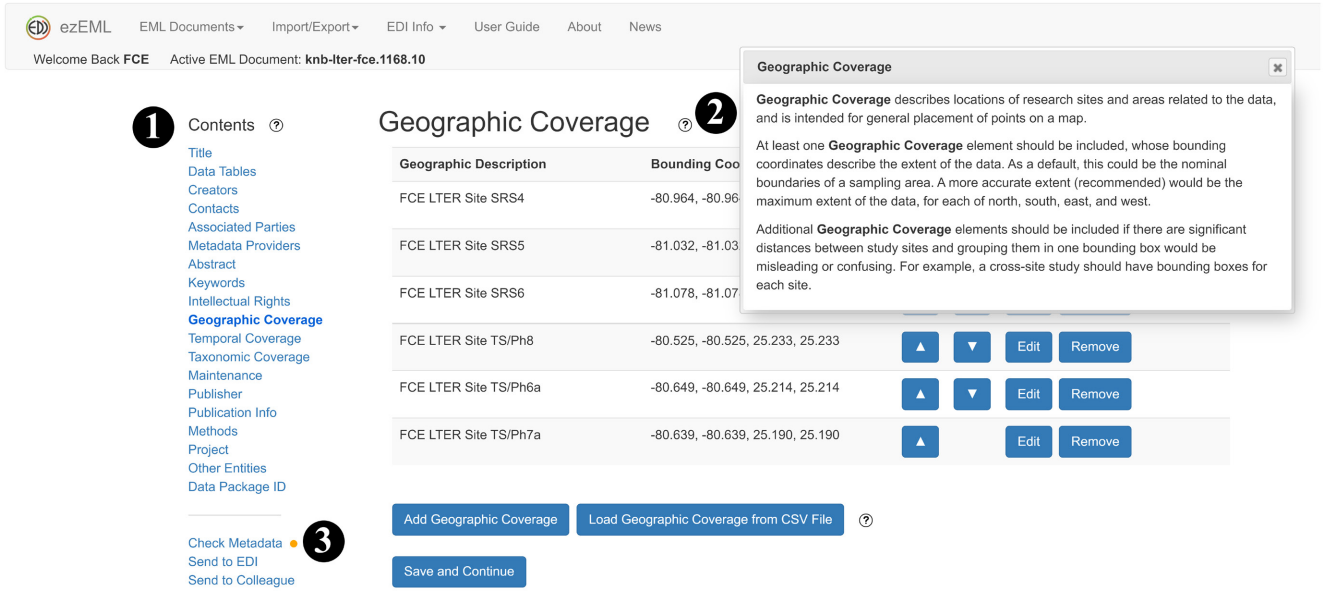

Fig. 1. This form, for capturing geographic metadata, shows features common to most windows in ezEML. 1. Users can enter metadata in any order by accessing forms through the Contents panel. 2. Clickable circled question marks yield pop-up windows to help the user understand what to enter. 3. The circle next to Check Metadata will be red if the user still needs to fill in some required metadata elements.

by emailing an ezEML data package to another user who can then import it into ezEML, modify the metadata, and email it back to the original user.

After completing the necessary metadata, a typical user would choose the "Send to EDI" link in the menu, and ezEML sends the ezEML data package to the attention of the EDI repository's Data Curation Team. A curator typically responds to the request to publish data within the same day. The metadata creator may be offered suggestions for improving their metadata, after which the curator publishes a proof of the dataset in a staging environment for the submitter to review and approve. Most datasets are published to the EDI production repository within 24 hours of approval. The use of ezEML is provided at no cost to end users.

Other use cases

The ezEML editor is ideal for the "one-off" type of data submission, such as a dataset used in a publication, but it is also a great tool for users who have a number of datasets to publish. Content from previously created ezEML data packages in the user's account can be imported into new packages, simplifying the publication of related datasets for a large project. Elements that can be imported include creators, contacts, research site locations, taxonomic information, project details, and funding awards. If a user needs to publish a new data table that has a similar structure as a table in an existing ezEML data package, there is also a function to "Clone Column Properties from Another Data Table." Thus, ezEML makes creating new datasets more efficient by reusing existing metadata, which also makes the metadata more consistent across data packages.

ezEML is being adopted by information managers (IMs) across the U.S. who manage datasets for multiple users, such as for long-term ecological research sites of the US LTER Network. ezEML facilitates periodic updates to long-term datasets by allowing an IM to fetch metadata and data directly from the EDI

23276096, 2022, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1002/bes2.2018, Wiley Online Library on [27/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

repository. The IM can then re-upload a new data file to replace the previous version without having to re-enter all the column properties metadata. ezEML also supports the use of templates, which provide a customizable pattern of metadata for reuse by members of a research site or team. An ezEML template can be pre-populated with the LTER's core research sites, project and funding information, contact information, and site-specific keywords. Users may import the template from the ezEML menu and delete core research sites they did not use, for example, which is a much easier process than re-entering site coordinates. Template usage improves metadata consistency across data packages related to a project.

While infrequent data submitters usually request that an EDI data curator complete the ezEML data package upload to the repository, EDI also supports the use case where an IM prefers to curate and upload packages themselves. An IM may ask researchers to use the "Send to Colleague" selection from the ezEML menu to email the package to the IM rather than directly to EDI. The IM can then edit the metadata and email the ezEML data package back to the researcher for approval if needed. EDI will provide the IM with an EDI account with which to manage ezEML data package uploads.

Summary

Publishing data is an activity most ecologists must now undertake to satisfy open science policies of funders and publishers. Recognizing that existing tools for capturing and structuring rich metadata are not suitable for all ecologists, EDI developed a user-friendly web-based editor that anyone can use. ezEML is gaining popularity in the ecological community among individual scientists as well as information managers who support groups of ecologists. To view a demonstration of ezEML, please access the EDI YouTube Library.

Acknowledgments

Development of ezEML by EDI was supported by National Science Foundation grants #1931143 and #1931174.

Literature Cited

Boettiger, C., M. B. Jones, M. Maier, B. Mecum, M. Salmon, and J. Clark. 2022. EML: read and write ecological metadata language files. https://CRAN.R-project.org/package=EML

Ecological Society of America. 2022. Open research policy. https://www.esa.org/publications/data-policy/

Jones, M. B., M. P. Schildhauer, O. J. Reichman, and S. Bowers. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution & Systematics 37:519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Smith, C. 2022. EMLassemblyline (v3.5.4). Zenodo, Software. https://doi.org/10.5281/zenodo.6625344

Waide, R. B., J. W. Brunt, and M. S. Servilla. 2017. Demystifying the landscape of ecological data repositories in the United States. Bioscience 67:1044–1051. https://doi.org/10.1093/biosci/bix117

Wiley. 2022. Wiley's data sharing policies. https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html

23276096, 2022, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1002/bes2.2018, Wiley Online Library on [27/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License