Instructions for Archiving Data on the EDI Data Portal

This document provides steps for

* Preparing metadata and EML for publication
* Evaluating and testing your data package in the EDI Staging Environment
* Publication of data on the EDI Data Portal

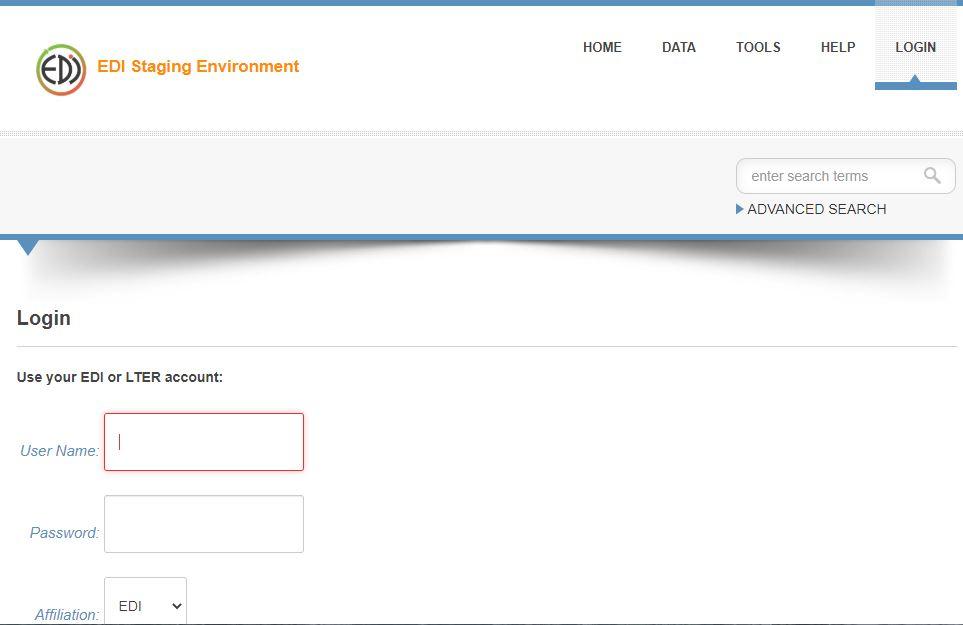
Things you will need:

* R: A language and environment for statistical computing – to create metadata templates and EML ([link](https://cran.rstudio.com/) to download R). [Link](https://rstudio.com/products/rstudio/download/) to download RStudio - a user-friendly console to run R code.
* Example EMLassemblyline R file, “run\_EMLassemblyline\_for\_DirName.R”. This R file has example information to aid in filling out the auto-created R file in step 4 on page 3.
* Link to the [EDI Staging Environment](https://portal-s.edirepository.org/nis/home.jsp) – this is a testing environment where you can evaluate your EML file and see what your data package looks like before publishing it to the EDI Data Portal.
* Link to the [EDI Data Portal](https://portal.edirepository.org/nis/home.jsp) – this is where you load and publish your data package.

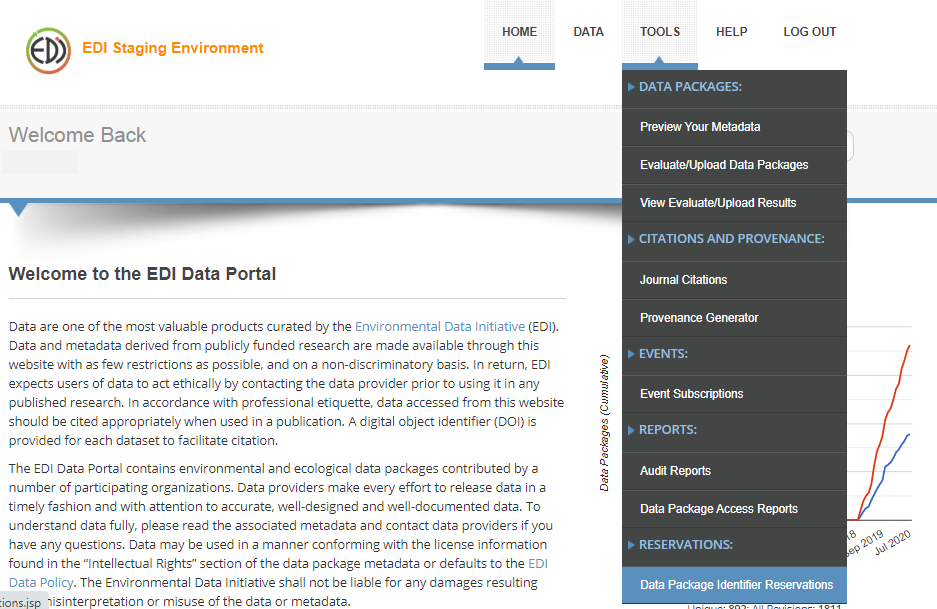
*If you have any questions or issues, email the EDI help desk at* [*support@environmentaldatainitiative.org*](mailto:support@environmentaldatainitiative.org)*, or visit EDI’s virtual office hours Wednesdays 2-3pm CT (3-4pm ET) (zoom link can be found* [*here*](https://environmentaldatainitiative.org/edi-office-hours/)*) to speak with an EDI staff member directly.*

**Prepare metadata and EML for publication**

1. Reserve a data package identifier. Navigate to the [EDI Staging Environment](https://portal-s.edirepository.org/nis/home.jsp) and login (click “LOGIN” on the top-right corner of the page)



1. On the top right corner of the page, hover over “TOOLS”. In the “TOOLS” drop down window click on “Data Package Identifier Reservations”.



1. On the “Data Package Identifier Reservations” page, click on “Reserve Next Available Identifier(s)”. You will see a table like the one below with your data package identifier information. Make note of the “Data Package Identifier” to use in step 2 below



1. Create a working directory on your computer that contains empty template folders. To do this, in R studio run each code snippet below:

remotes::install\_github("EDIorg/EMLassemblyline") #run this line if you do not have the package “EMLassemblyline” installed. Note: you just need to run this once on your machine and then can skip it during any future archiving

library(EMLassemblyline) # load the EMLassemblyline library

template\_directories((path=""), dir.name="") # fill in “ ” (see instructions below)

“path” is the file path where you want the folder to be located (i.e. path = "C:\\Users\\user\\Desktop\\EDI\_archiving\\BreedingBirdSurveyData”). **Note**: you must have the double backslashes (\\) in the path name for it to work. “Dir.name” is the name of the folder you want the template folders to be stored (i.e. dir.name = “edi.556”). **Note**: this is a brand-new folder (i.e., it is not in your directory yet). It is good practice to name the folder with the data package identifier you obtained in step 1.

After running the code above, navigate to your newly created file folder on your computer. In the folder you will now see three empty folders (“data\_objects”, “eml”, and “metadata\_templates”) and an R file called “run\_EMLassemblyline\_for\_*your dir.name*.R”, where *your dir.name* is the name you provided in dir.name = "" in the code above (e.g., “run\_EMLassemblyline\_for\_edi.556.R”).

1. Copy and paste your finalized data tables into the “data\_objects” folder. Other data entities such as zip files, geospatial data, pictures, pdf(s), and Rcode should be transferred to the “data\_objects” folder if you would also like them to be part of your data package.
2. Note: See “run\_EMLassemblyline\_for\_DirName.R” for examples and descriptions not given in the auto-created “run\_EMLassemblyline\_for\_*your dir.name*.R”.

Double click on the R file “run\_EMLassemblyline\_for\_*your dir.name*.R” to open it. Read through the text, fill in empty “ ” with appropriate information, and execute the code starting from the top of the document to the bottom. Only skip steps where it says “optional”.

Lines 16 – 69 will create empty .txt and .docx files in your “metadata\_templates” folder.

\*\* After running the code chunk, EMLassemblyline::template\_table\_attributes(), fill out the information in the text file(s) before running the chunk, EMLassemblyline::template\_categorical\_variables(). \*\*

EMLassemblyline::template\_categorical\_variables() are created when *class* is defined as “category” in the attribute text file(s)

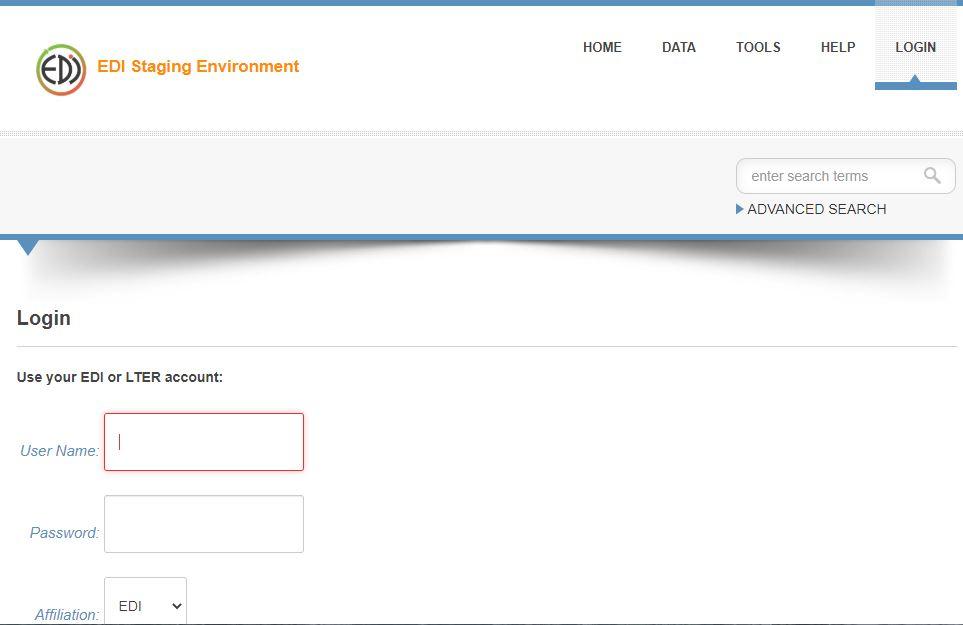
**STOP** after the code chunk “EMLassemblyline::template\_taxonomic\_coverage”

1. Now that you have all the empty files in the “metadata\_templates” folder, populate the files with metadata information (see“MetadataTemplate\_BoilerPlate.docx” for help). Further explanation and examples of how to populate the auto-generated metadata templates can be found [here](https://ediorg.github.io/EMLassemblyline/articles/edit_metadata_templates.html).
2. Once all files in the “metadata\_templates” folder are populated, go back to the “run\_EMLassemblyline\_for\_*your dir.name*.R” document. Starting at line 71 (“EMLassemblyline::make\_eml()”), read through the text, fill in empty “ ” with appropriate information, and execute the code EMLassemblyline::make\_eml(). EMLassemblyline::make\_eml will create an XML file in your “eml” folder. The XML file contains EML language that is computer readable. The EML details all the information in your data package (i.e., the information in the “metadata\_templates” folders).

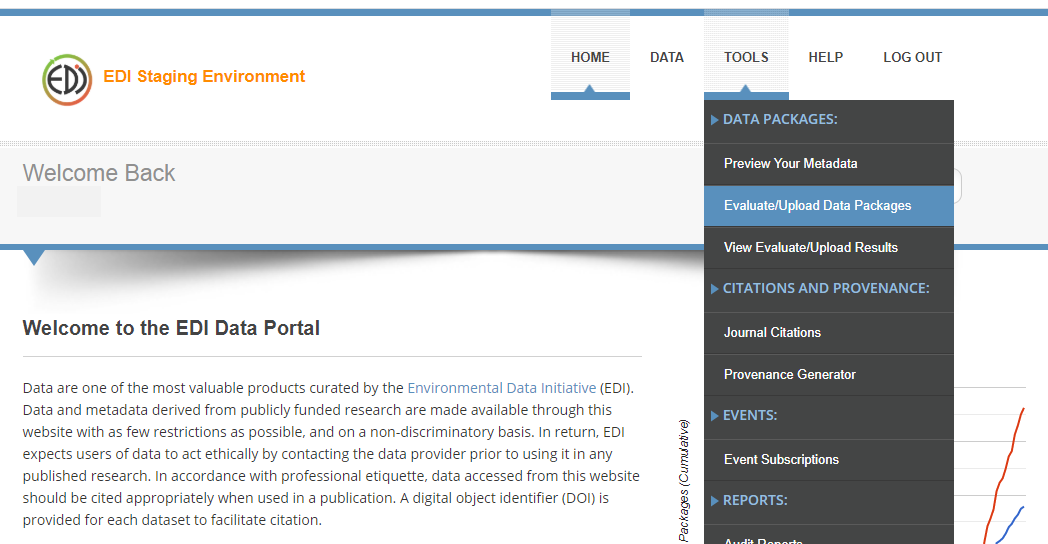
**Evaluate and test your data package in the EDI Staging Environment**

Before publishing your data package to the EDI Data Portal, first a) evaluate the EML file to generate a “Quality Report”, then b) upload the package to the staging environment to preview what it would look like when published.

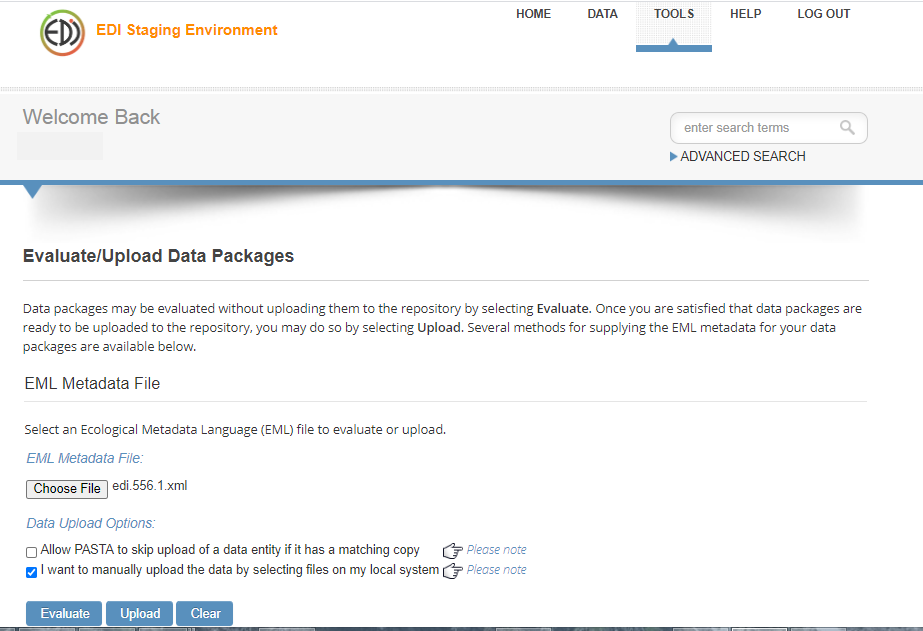
1. Navigate to the [EDI Staging Environment](https://portal-s.edirepository.org/nis/home.jsp) website and login (click “LOGIN” on the top-right corner of the page)



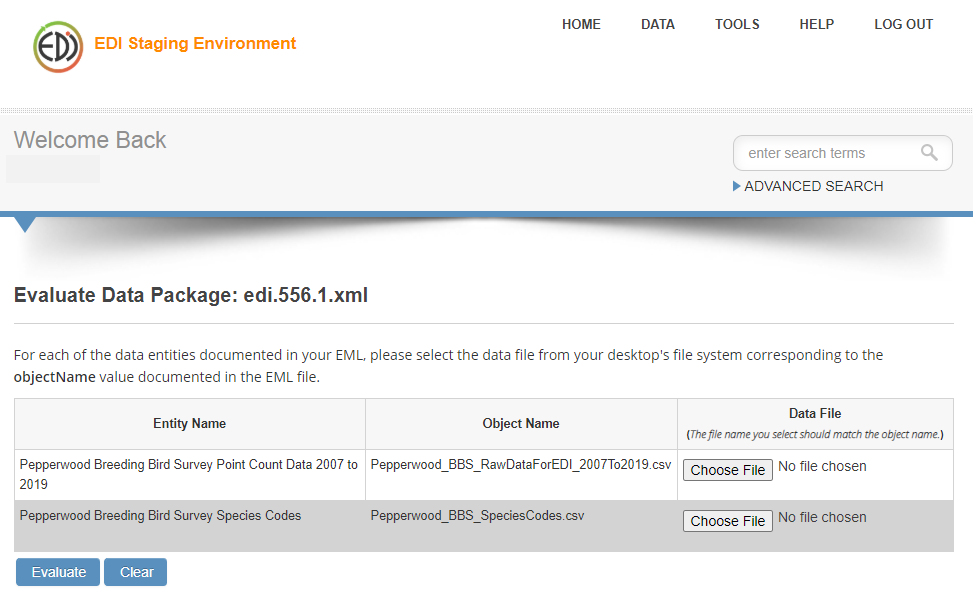
1. Hover over the “TOOLS” tab in the top-right banner then select “Evaluate/Upload Data Packages”



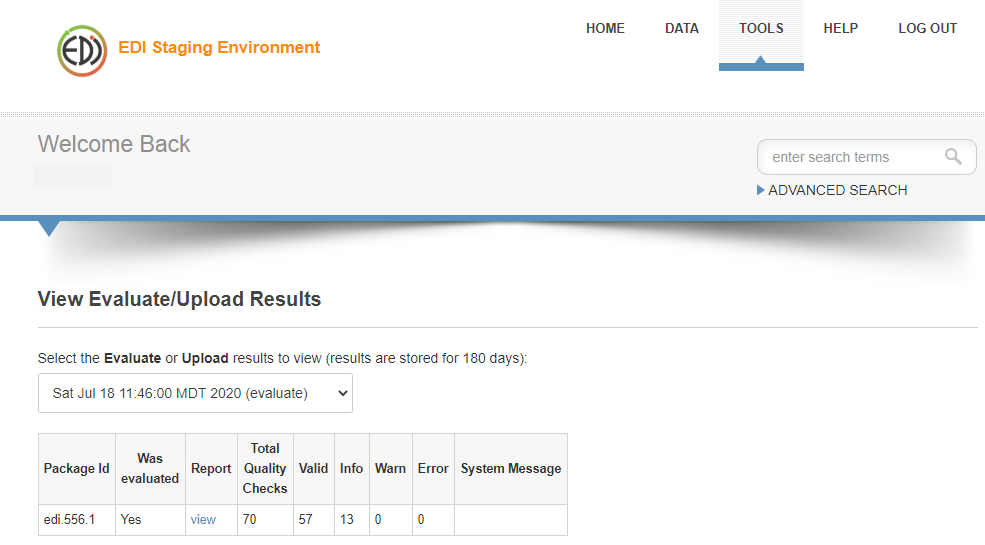
1. *Evaluate the EML file to generate a “Quality Report”*
2. Click on the “Choose File” button (underneath the blue text, “EML Metadata File:”). A file explorer window will pop up, navigate and select the XML file that contains the EML for your data package. Next, check the box to the left of “I want to manually upload the data by selecting files on my local system” (underneath the blue text, “Data Upload Options:”). Lastly, click on the “Evaluate” button.



1. Navigate to and attach the data objects associated with the data package (i.e., files in your “data\_objects” folder). The EDI Staging Environment will auto populate the “Entity Name” and “Object Name” with what was documented in the EML. Click on the “Choose File” button and attach the appropriate file that matches the “Object Name”. When all files are attached, click on the button “Evaluate”.



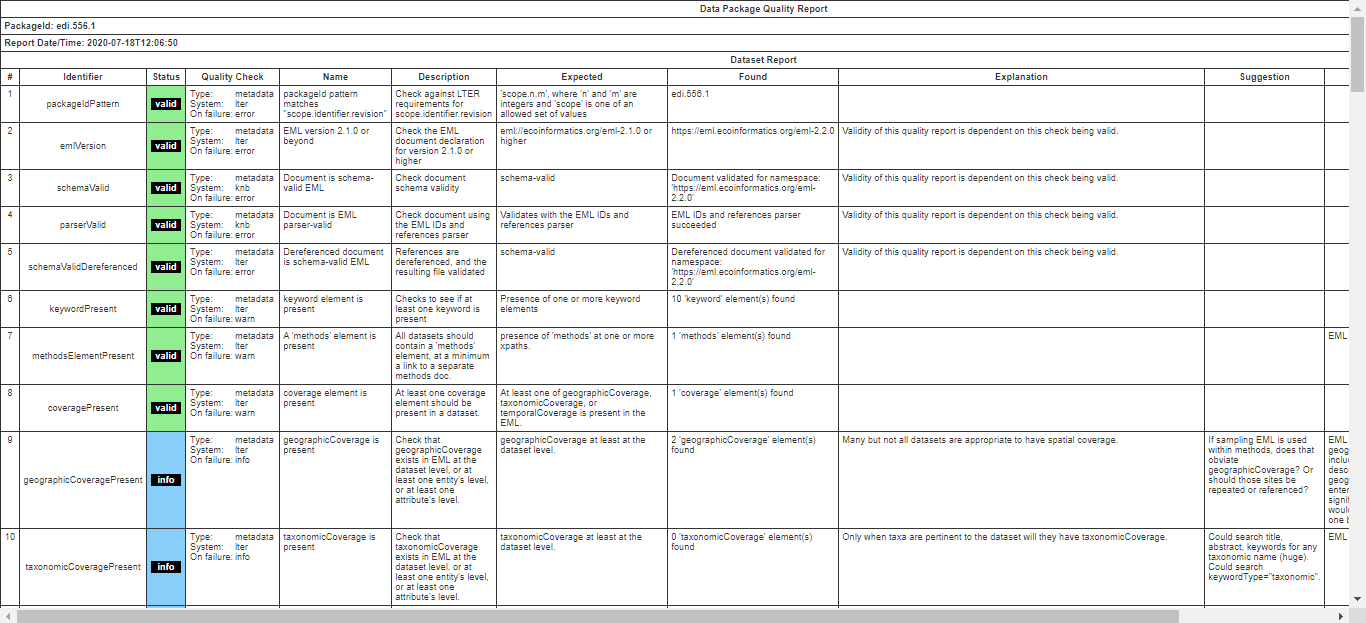
1. A table is generated with summary information from the Quality Report. Pay attention to the “Warn” and “Error” columns. If there is a number other than “0” in either of the columns, click on the “view” link to look at the quality report. Note: a data package can be published with warnings but not with errors. Contact EDI support, [support@environmentaldatainitiative.org](mailto:support@environmentaldatainitiative.org), if you have questions about resolving any warnings or errors. Additionally, this [document](https://ediorg.github.io/data-package-best-practices/EMLmetadata/index.html) has examples of what your EML structure should look like and can be helpful for fixing errors or warnings.



What is a Quality Report?:

The Quality Report is a series of quality checks for

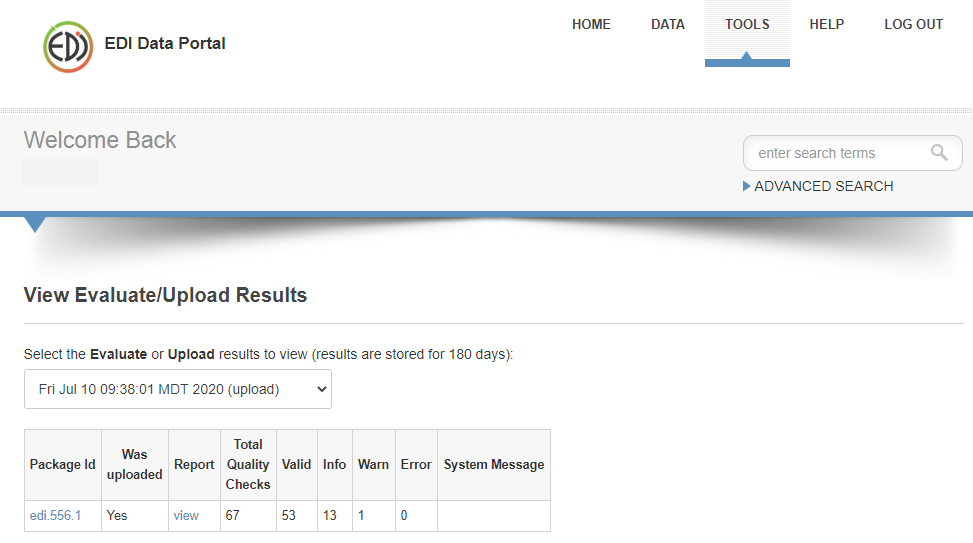
* Metadata validation
  + Ensures the EML is well-formed and schema valid
  + Content validation (does content match best practices?)
* Data validation
  + Accessible (can data be downloaded?)
* Congruence validation
  + Metadata description of data matches physical structure of data (e.g., correct number of columns, rows, datatype, delimiters)



1. *Upload the package to the staging environment to preview what it would look like when published*

After all warnings and errors in the quality report are addressed, upload the package in the staging environment

1. Navigate to “Evaluate/Upload Data Packages” (in the dropdown window under the “TOOLS” tab). Follow step a)1. above, but instead click on the “**Upload**” button.
2. Follow step a)2. above, but instead click on the “**Upload**” button.
3. A summary table will be produced like the one in step a)3. above, but now the package name will contain a link in the “Package Id” column.



1. Click on the “Package Id” link to preview what your data package will look like when it is published

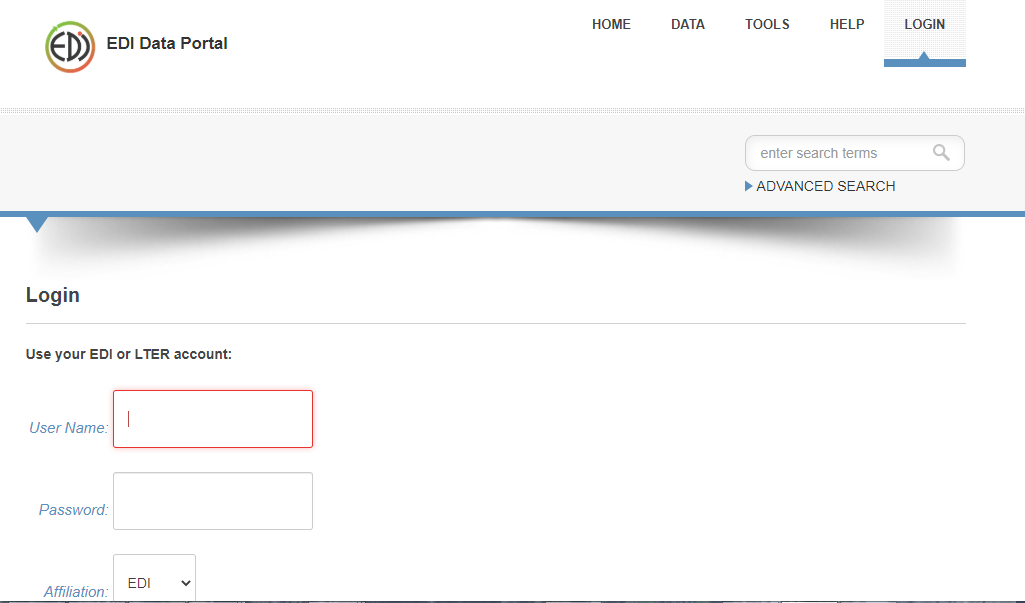


**Publish the data package using the EDI Data Portal**

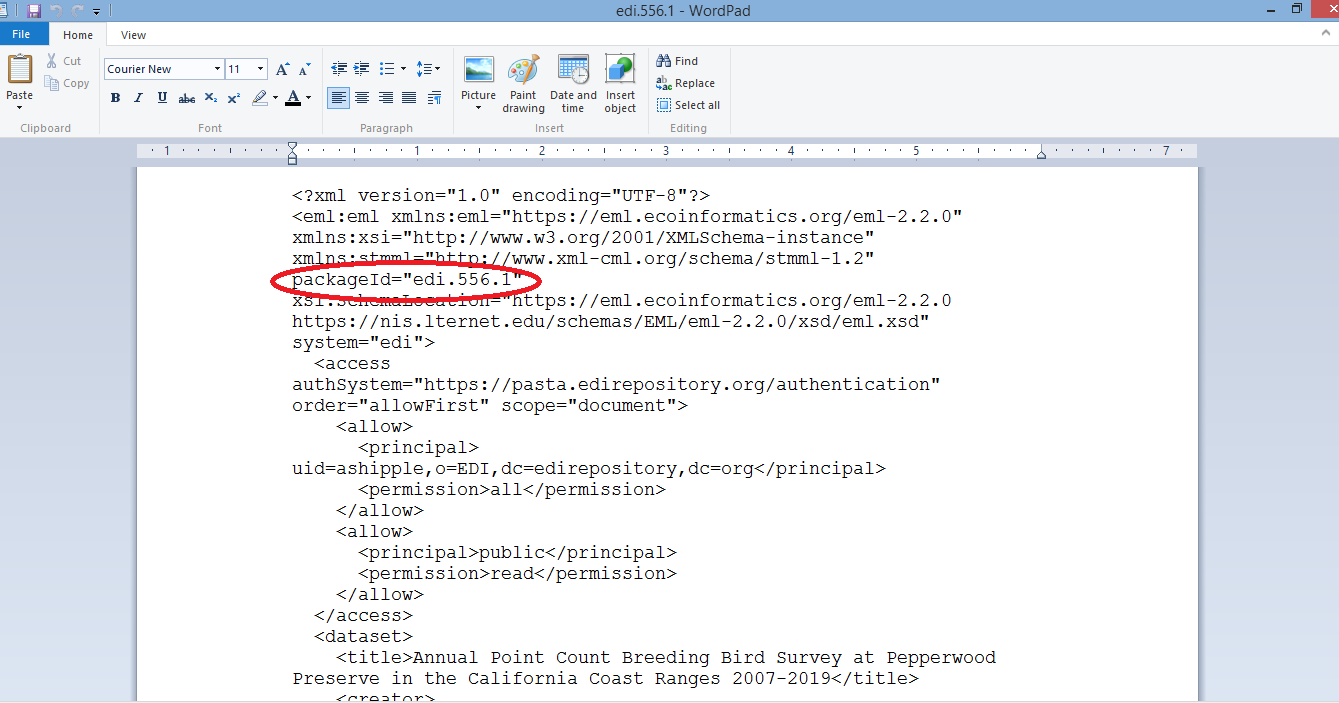
When you are happy with how your data package looks in the EDI Staging Environment, publish the package using the EDI Data Portal

1. Note: the EDI Data Portal and Staging Environment are separate entities and require a unique data package identifier for both.

Reserve a data package identifier in the EDI Data Portal. Navigate to the [EDI Data Portal](https://portal.edirepository.org/nis/home.jsp) and login (click the “LOGIN” tab in the top-right corner).

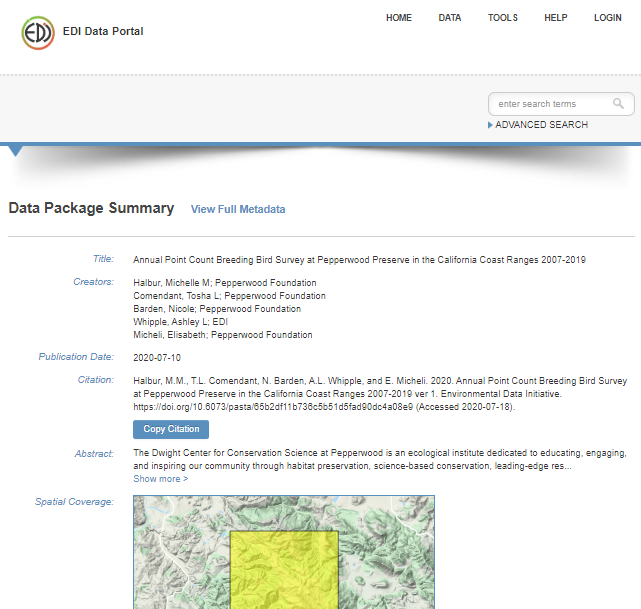


1. On the top right corner of the page, hover over “TOOLS”. In the “TOOLS” drop down window click on “Data Package Identifier Reservations”.
2. On the “Data Package Identifier Reservations” page, click on “Reserve Next Available Identifier(s)”. Make note of the “Data Package Identifier” to use in step 2 below.
3. Edit your existing EML and the name of your master folder (folder that contains the folders “data\_objects”, “eml”, and “metadata\_templates”) to match the Data Package Identifier you received in step 1.b. above.
   1. To edit the EML file, navigate to the “eml” folder and open the XML file. On the 5th line of the document, find “packageID = “edi.*xxx.x*”. Delete out the package number and replace it with your packageID received from the EDI Data Portal. Save and close the file.



* 1. Rename the XML file to also match the new Data Package Identifier. Do this by right clicking the XML file in the “eml” folder and selecting “Rename” from the drop-down window.

1. Follow the steps b)1-4 under the “Upload the package to the staging environment to preview what it would look like when published” section above.
2. Yay! Your data package is published!



* To share your data package with others use the Digital Object Identifier found in the “Citation” section of the your Data Package Summary (i.e. <https://doi.org/10.6073/pasta/65b2df11b736c5b51d5fad90dc4a08e9>), or you can create an exact url address.
  + To create an exact url address use the format https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=**\***&revision=**#**; where \* is the Data Package Identifier (i.e. 556), and # is the version number (i.e. 1). If you would like to keep the url static and will be making new versions of the package in the future, simply change the version number to read “newest” (i.e. <https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=556&revision=newest>). By replacing the version number with “newest” it ensures that the link will always go to the newest version of the data package in the EDI Data Portal.
* Your data package is now searchable on
  + [EDI Data Portal](https://portal.edirepository.org/nis/advancedSearch.jsp)
  + [DataONE](https://www.dataone.org/)
  + [Google Dataset Search](https://datasetsearch.research.google.com/)