The PASTAprog Web Service for Generation of Statistical Programs

John Porter, VCR/LTER (http://www.vcrlter.virginia.edu), 2015

The PASTAprog web service allows you to rapidly create and run statistical programs for the analysis of Long-Term Ecological Research (LTER) data from the LTER PASTA Data Portal. PASTAprog uses Ecological Metadata Language (EML) metadata to generate statistical programs that:

1. Download and ingest tabular data in LTER Datasets
2. Run simple statistical summaries on data columns

PASTAprog is directly integrated into the LTER Data Portal and powers its “Code Generation” features. However, as a web service, it can be used independent of any web page, or incorporated into other online pages.

PASTAprog is designed to use the information from EML metadata documents to streamline the “routine” tasks associated with ingesting LTER data. Users are expected to be knowledgeable about the statistical tools they are using so they can amend and extend the programs created by PASTAprog to support sophisticated and problem-specific analyses. Statistical packages currently supported by PASTAprog are:

1. MATLAB (stylesheet from Wade Sheldon, GCE)
2. R
3. Statistical Analysis System (SAS), and
4. Statistical Package for the Social Sciences (SPSS)

Use of PASTAprog

The PASTAprog REST-based web service is invoked using a URL consisting of at least 3 elements:

- The base URL – typically: http://www.vcrlter.virginia.edu/webservice/PASTAprog/
- An LTER dataset identifier or package ID consisting of a scope, identification number and a revision number. For example: “knb-lter-vcr.26.14”. These package IDs can be found by searching the PASTA Data Portal for datasets.
- A suffix that indicates the type of program you wish to generate:
  - .r for a R statistical program
  - .sas for a SAS statistical program
  - .spss for a SPSS statistical program
  - .m for a MATLAB program (see below for special notes regarding use with MATLAB)

Thus a complete URL for invocation would look like:

http://www.vcrlter.virginia.edu/webservice/PASTAprog/knb-lter-vcr.26.14.r

If used in a web browser, a program will be generated that can be copied and pasted into a text editor or into command interfaces for statistical packages. However, the URL can also be included directly in statistical programs using functions that retrieve code from a URL and run it without using a web browser. For example, the R "source()" function causes R to read a program from the URL specified and run it. As in:

source("http://www.vcrlter.virginia.edu/webservice/PASTAprog/knb-lter-vcr.26.14.r", echo=TRUE)

Using PASTAprog with Other Metadata Repositories

By default, PASTAprog searches the PASTA repository for the metadata document to transform. However that default can be overridden to point to any web-accessible metadata document by appending a "emlurl" attribute to the URL as in:

http://www.vcrlter.virginia.edu/webservice/PASTAprog/myprog.sas?emlurl=http://myserver.edu/myEMLdoc.xml

Where http://myserver.edu/myEMLdoc.xml is the web address of an EML metadata document and myprog.sas is the name of the SAS program you wish to generate. The easiest way to get the URL for an EML metadata document in another repository, such as the LTER Metacat, is to use a web browser and use "copy link" (or similar) to copy the URL for pasting into the web service URL.

Datasets with Multiple Data Tables

PASTAprog generates programs that read all of the data tables documented in the metadata for a specific dataset, adding sequence numbers to the names of the constructed data structures. Thus a dataset with three dataTables will generate R code for creating three data frames, dataTable1, dataTable2 and dataTable3. SAS creates working datasets named datafile1, datafile2 etc. Of special note, the program for generating SPSS creates statements that sequentially read each of the data tables into the active system file (overwriting previous data tables), therefore it is necessary to add SAVE statements to save each data table to a separate file.

PASTAprog and Data Retrieval

When run, the programs created by PASTAprog should automatically retrieve the data from within the statistical package without additional user actions. However, this feature depends on the EML metadata file correctly specifying the URLs that will retrieve the correct data files. If the data URLs in the EML document is incorrect, or point to a form, rather than directly to data files, it will be necessary to download the data manually and edit the input section of the program to reflect the location of the data file on the local PC. This should typically not be necessary for publicly-available data in PASTA, because the PASTA system itself provides a copy of the data. However, for data in the LTER Metacat or in other EML documents outside PASTA, the reliability of the links to data files are more variable, so that some automatically-generated programs will work without modification, but others will require editing and for the data to be manual downloaded.

Each time the program created by PASTAprog is run it downloads a fresh copy of the data from the PASTA server. For small datasets, this is not a problem. However for really large datasets, users may wish to store a copy of the data locally and, as discussed above, modify the program to read from local files. Alternatively, the program can be run once, unmodified, and the resulting data structure saved in the formats specific to individual statistical packages. For example, in SAS a data table can be saved as a permanent dataset, or in SPSS saved as a system file, or in R as a saved data frame or workspace. These structures can be retrieved for further analysis without requiring re-ingestion of the data.

Statistical Package-Specific Notes

MATLAB: MATLAB does not allow function file names to include any periods, other than the trailing ".m". Therefore PASTAprog allows the substitution of the underscore character "_" in place of periods when specifying the package ID. Thus, knb-lter-vcr_26_14.m is a valid replacement for knb-lter-vcr.26.14.m. Alternatively, if an emlurl attribute is specified the program name can be whatever you wish to specify, as long as it ends in ".m" to indicate that a MATLAB progam is desired. The MATLAB stylesheet (http://gce-lter.marsci.uga.edu/public/xsl/toolbox/EMLdataset2mfile.xsl) was created by Wade Sheldon (wsheldon@lternet.edu) at the Georgia Coastal Ecosystems LTER.

Note that the m-file produced by the MATLAB stylesheet contains help on function syntax and usage. For example, if you request knb-lter-vcr_26_14.m, change to the directory where the file was downloaded within MATLAB and type "help knb-lter-vcr_26_14" to view the embedded help text and "data = knb-lter-vcr_26_14" to run the program. Also, in contrast to the R, SAS and SPSS programs the MATLAB program does not perform any specific analyses. All data columns and attribute information are imported as arrays along with metadata content, which are stored in a structure variable for use with MATLAB analyses and plotting tools. For a more complete analytical solution for EML-described data in PASTA, see the GCE Data Toolbox for MATLAB (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox)

R: Data tables are ingested into R data.frames named dataTable1, dataTable2 etc. If there are irregularities in the formatting of one or more numerical data values in a column of data (e.g., a value has a letter O instead of a zero (0) in a number, or non-numeric missing value indicators other than NA), R may automatically convert a data vector from mode NUMERIC to type FACTOR. This can be tricky because the display of the vector may appear correct, but many statistical analyses operate on the index of the factor (e.g., sequential numbers from 1 to N, where N is the number of levels of the factor), instead of on the numbers themselves. Future versions of the PASTAprog-generated programs will automate testing of the vector mode in R, but until that functionality is available, users should do the testing themselves using the "mode()" and "attributes()" functions of R.

The current version of the R program does not conduct range checks or do anything with missing values codes other than NA.

SAS: In addition to statistical summaries the SAS programs created by PASTAprog include variable and value-labeling statements, range checks for numerical data and recognize missing values, as specified in the corresponding metadata document.

SPSS: As of this writing, the SPSS programs written by PASTAprog do not automatically retrieve the data from a URL. This feature will be added once I find code that will do the retrieval from inside SPSS. For now, you will need to download the data manually and modify the path to the data file in the program, as specified in the comments. Note that if there are multiple input data tables, you will need to add SAVE statements to save each of the data tables to a system file of your choice. Otherwise, only the last data table read will be available.

In addition to statistical summaries the SPSS programs created by PASTAprog include variable and value-labeling statements, range checks for numerical data and recognize missing values, as specified in the corresponding metadata document.

How PASTAprog Works

PASTAprog uses eXtensible Markup Language (XML) stylesheets to transform EML metadata documents (which are XML documents). The system is implemented as a set of PHP and Perl tools that parse the input URL, fetch the EML metadata document and apply the stylesheet transformation. Copies of stylesheets used to create the statistical programs are available in the LTER Network SVN repository (http://svn.lternet.edu). Users are encouraged to improve the stylesheets so that they create better code for subsequent incorporation into the web service.

Troubleshooting

The success of programs created using PASTAprog is ultimately dictated by the quality of the underlying EML metadata and data. If the metadata incorrectly identifies data columns, the program generated by PASTAprog will as well. Some metadata, while technically correct, incorporate bad practices that may cause specific statistical packages to fail. For example, some EML metadata may include attribute or variable names that include special characters, spaces or mathematical operations as part of the attribute names. This can be highly confusing to statistical packages such names violate the naming conventions of the statistical package, causing programs to fail when run. Some automatic fixes have been incorporated into the PASTAprog program generation, but there may still be times when this automatic correction will fail. When there are inconsistencies between the metadata and the data or problems with attribute naming conventions, it may be necessary to manually edit and correct the PASTAprog-generated program.

Dates are also a common source of problems because they are encoded so many different ways in ecological data, either as combined fields (e.g., “2012-09-07”) or as separate variables. Most programs generated by PASTAprog will correctly handle standard dates. However, metadata sometimes lists date data as strings or separate columns of numbers reflecting year, month, day etc. For this reason, you will probably need to do some additional coding within a statistical package to get dates to perform reliably.