NIS data workflows best practices 0.2- 5/13/2013

Document history:

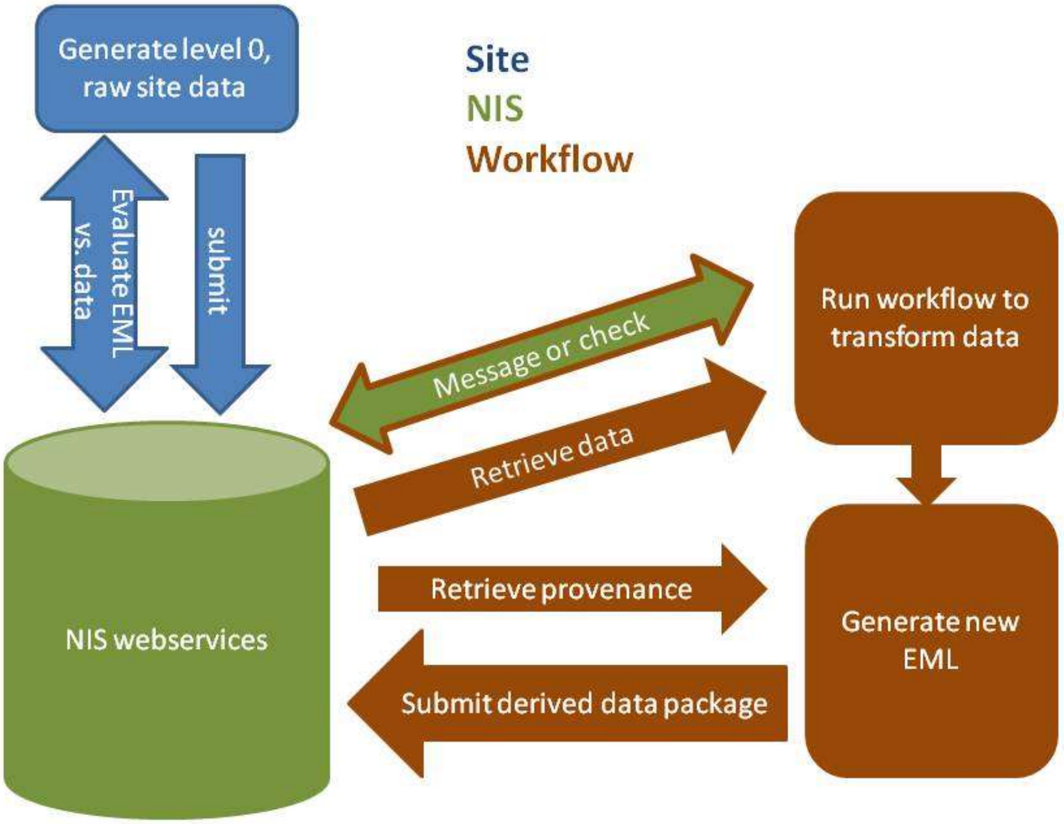
API documentation links updated 5/13/2013

NIS data workflows best practices 0.1- 12/.14/2012

Corinna Gries, John Porter, Ben Ruddell, Mark Servilla, Wade Sheldon, Jonathan Walsh

Intoduction:

General overview of methods supported by NIS webservices and employed for generating a derived dataset



Methods supported by NIS webservices considered in this document:

- Evaluate EML and datapackage: check EML/dataset congruency and general dataset quality.
- Read a PASTA data package resource map: returns URLs to the EML, entity data, and quality report
- Read a PASTA data package EML metadata document
- Read a PASTA data package data entity

- Create a PASTA data package
- Update a PASTA data package
- Get a PASTA provenance metadata fragment
- Poke the PASTA Event Manager: Request information about a certain event
- Subscribe to a PASTA event alert

API documentation

- https://pasta.lternet.edu/package/docs/api
- https://pasta.lternet.edu/eventmanager/docs/api
- https://pasta.lternet.edu/audit/docs/api

General Considerations:

Providing data (data entity + EML) to PASTA

The general sequence for data ingestion into PASTA is:

1. Generate data tables and PASTA-compatible Ecological Metadata Language (EML). Such metadata needs to include a link that will yield access to the contents of the data table (with or without authentication).
2. Use the evaluation web service to test that data will load properly into PASTA and modify the data or metadata.
3. Perform the ingestion using the web service.

Steps 1 and Step 2 pose the most challenges as data that is usable directly by researchers may still not be compatible with PASTA. For more detail please see: https://nis.lternet.edu/NIS/?q=node/54

New version vs. new data package (data entity + EML)

Long term data handling: Notifications of changes in datasets are triggered by uploading of new revisions of metadata. They may reflect either addition of new data (in existing formats) or changes to the structure/format of the data itself. Workflows reacting to these notifications will need to take appropriate steps depending on which of these changes triggered the notification. Detection of the types of changes that trigger a revision could be the responsibility of each individual workflow, or of a shared program or web service, or as part of the notification itself.

1. Data appended in the same file (same format or changed format)
    a. New version of EML

       Structural changes will break workflows

       One suggestion is to populate the maintenance element
2. Data in new entity (same format or different format)
    a. New EML file

A new EML file with new package ID will not indicate to the system that a dataset has changed. E.g. if met data are added in annual data packages, automation of triggering workflows will not work.

b. Same EML file, added entity
This will allow for automatic triggering of workflows, however, the workflow will have to detect that a new entity as been added and use that, which will require a new url to be used to access the data.

Software for accessing NIS webservices

cURL: (command line URL)

cURL is an open source command line tool for transferring data using URL syntax (http://curl.haxx.se/), and distributions are available for Unix/Linux, Windows and Macintosh systems. It functions similarly to GNU Wget, but supports many more options for controlling the HTTP/HTTPS communications.

Download an executable for your system here: http://curl.haxx.se/dlwiz/?type=bin. The cURL executable file should be copied to a directory accessible in the system path in order to call cURL from the command line (e.g. C:/Windows for Windows).

Note that some Windows systems will need an updated version of the Microsoft Visual C runtime library, or cURL will fail with a missing DLL error message. Installers can be download from here: http://answers.microsoft.com/en-us/windows/forum/windows_7-windows_programs/the-program-cant-start-becuase-msvcr100dll-is/5c9d301a-2191-4edb-916e-5e4958558090

Examples of using cURL to interact with PASTA web services are provided below. Note that the NIS security certificate is not recognized by most clients, including cURL, so you may have to substitute http for https in the PASTA URLS. Note that passwords will be transmitted in clear text with http, though, so resolving SSL certificate issues on the client is highly preferable (contact Mark Servilla for more information).

Evaluate EML vs. data entity (congruency check)

```
curl -i -u uid=userID,o=LTER,dc=ecoinformatics,dc=org:password -X POST -H
"Content-Type: application/xml" -d @eml.xml
https://pasta.lternet.edu/package/evaluate/eml
```

Replace eml.xml with the correct name. The curl command needs to be run in the folder where the eml file resides. The authentication is not necessary and this can evaluation can be done with this command as well:

```
curl -i -X POST -H "Content-type: application/xml" -d @eml.xml
https://pasta.lternet.edu/package/evaluate/eml
```

Add level 0 dataset to NIS (raw site data)

```
curl -i -u uid=userID,o=LTER,dc=ecoinformatics,dc=org:password -X POST -H
"Content-Type: application/xml" -d @eml.xml
https://pasta.lternet.edu/package/eml
```

Update data package in NIS

```
curl -i -u uid=userID,o=LTER,dc=ecoinformatics,dc=org:password -X PUT -H
"Content-Type: application/xml" -d @eml.xml
https://pasta.lternet.edu/package/eml/scope/identifier
```

Get data out of NIS

```
curl -i -X GET
https://pasta.lternet.edu/package/eml/scope/identifier/revision
```

e.g.: https://pasta.lternet.edu/package/data/eml/knb-lter-dog/2/2/5709_2009_jan_feb_v1

Get metadata out of NIS

```
curl -i -X GET
https://pasta.lternet.edu/package/metadata/eml/scope/identifier/revision
```

e.g.: https://pasta.lternet.edu/package/metadata/eml/knb-lter-dog/2/2

Subscribe to a PASTA event alert:

```
curl -i -u uid=userID,o=LTER,dc=ecoinformatics,dc=org:PASSWORD –X POST -H
"Content-type: text/xml" -d @subscription.xml
https://pasta.lternet.edu/eventmanager/eml
```

The subscription.xml file:

```
<subscription type="eml">
      <packageId>knb-lter-dog.2</packageId>
      <url>http://magma.lternet.edu/workflow-demo/eventhandler</url>
      <accessControlRule></accessControlRule>
</subscription>
```

Get provenance information out of NIS

```
curl -i -u uid=userID,o=LTER,dc=ecoinformatics,dc=org:PASSWORD –X PUT -H
"Content-type: application/xml" -d @methods.xml
https://pasta.lternet.edu/package/eml/provenance/?scope.identifier.revision
```

This webservice will add provenance information into an existing EML document between <methods></methods> tags. It will append this information to existing methods. The minimum information in the methods.xml are the tags <methods></methods>.

Read a PASTA data package resource map:

```
curl -i -X GET
https://pasta.lternet.edu/package/eml/scope/identifier/revision
```

NOTES: You may have to add the -k parameter if you receive a message regarding a problem with a certificate. Be careful of the placement. In the examples above, the -X parameter must be adjacent to the GET command. If you put the -k parameter in between it will not work.

MATLAB:

The purpose of this section is to document best practices from the perspective of a typical LTER researcher (or Information Manager) who wants to use data from PASTA within the MATLAB environment. The scope of this section is not to search for datasets or to upload or evaluate new or revised data packages, but simply to get a dataset already stored by PASTA into MATLAB in a format where it can be used. This is the minimal and most common task encountered by a typical researcher.

RELEVANT MATLAB COMMANDS

- urlread: executes a URL and retrieves the results into a string (alternative to command-line "cURL" application for HTTP addresses only)
- urlwrite: executes a URL and saves the results to a local file (alternative to command-line "cURL" application for HTTP addresses only)
- system('curl …'): calls cURL to execute a URL and save the results to a local file or string, depending on the options specified (required for HTTPS addresses)
- xmlread: reads an XML document stored as a local file or URL into a Java DOM object
- xslt: applies an XSLT stylesheet to an XML document to transform it into HTML, text, etc.
- parseXML: parses XML into a MATLAB structure variable
- fetch_eml_data (GCE Data Toolbox): executes a URL and retrieves and parses all delimited text data tables to create a MATLAB structure
- eml2gce (GCE Data Toolbox): converts EML table data downloaded with fetch_eml_data into a GCE Data Structure compatible with the GCE Data Toolbox software

Evaluate EML vs. data entity (congruency check)

This task is not currently recommended as a best practice for MATLAB because there is no support in the urlread or urlwrite function for the HTTPS protocol, sending files or setting headers that are required by the PASTA web services. The best strategy is to install cURL and call it from within MATLAB as a system command, or using a custom version of urlwrite that supports these features in the Java network libraries. This functionality may also be supported by the GCE Data Toolbox in the future.

Add level 0 dataset to NIS (raw site data)

This task is not currently recommended as a best practice for MATLAB because there is no support in the urlread or urlwrite function for the HTTPS protocol, sending files or setting headers that are required by the PASTA web services. The best strategy is to install cURL and call it from within MATLAB as a system command, or using a custom version of urlwrite that supports these features in the Java network libraries. This functionality may also be supported by the GCE Data Toolbox in the future.

Get data and metadata out of NIS and into Matlab

It is relatively simple to obtain an EML metadata file for a data package from the NIS from within MATLAB, using the PASTA web services, if the user knows the scope, the identifier, the revision number, and the specific dataset. The first three items listed above correspond to the data package, and the last is needed for a specific dataset. The example code below obtains the CSV format data file and the EML metadata file and writes them to the MATLAB working directory as "packdata.csv" and "packmeta.eml".

```
scope=('knb-lter-dog')
identifier=2
revision=2
```

```
dataset=('OceanusBorealis-B')

packdatafile=urlwrite(['http://pasta.lternet.edu/package/data/eml/
', scope, '/', int2str(identifier), '/', int2str(revision), '/'
dataset], 'packdata.csv');

packmetafile=urlwrite(['http://pasta.lternet.edu/package/metadata/
eml/', scope, '/', int2str(identifier), '/', int2str(revision)],
'packmeta.eml');
```

Once the CSV data file has been saved to disk, it can be manually imported into MATLAB using MATLAB's data import tools (e.g. uiimport, textscan, dlmread). Often there is a one-line header on the data file naming the variables and a comma or tab delimiter. Once the EML metadata file is saved on the disk, it can be viewed and edited using any XML editor (e.g. oXygen, XMLSpy) or text editor.

The "xmlread" and "parseXML" MATLAB functions may be used to read and parse EML. The xmlread function loads the EML metadata into a Java DOM object that supports standard Java methods for "walking" the document and retrieving content. The parseXML function returns the EML metadata as a deeply nested tree structure, with fields "Name", "Attributes", "Data" and "Children", for walking the document structure using MATLAB structure and array addressing syntax. However, these functions are not recommended because the user must have intimate knowledge of the EML schema and use tedious, iterative commands to retrieve critical content such as attribute information and units.

To address this problem, Wade Sheldon (GCE) wrote an XSLT stylesheet (EMLdatatable2mfile.xsl) that reads an EML document and creates a native MATLAB function m-file for download and parsing all delimited text data tables described in the document (support for fixed-width text is also planned). This stylesheet parallels the EML to R, SAS and SPSS stylesheets developed by John Porter (VCR) and will be added to the associated VCR-hosted web services. A MATLAB function ("fetch_eml_data.m") was also written to automate downloading an EML document from PASTA (or any other URL), applying the transform, and running the function to return the data, attribute metadata and high-level documentation metadata (title, abstract, creator, entity name, entity description) as a MATLAB structure. Source code for both "EMLdatatable2mfile.xsl" and "fetch_eml_data.m" will be available in the LTER SVN repository, as well as in the next GCE Data Toolbox software distribution.

In addition, the GCE Data Toolbox now includes a function ("eml2gce.m") that converts the structure returned by "fetch_eml_data.m" into a data structure compatible with the GCE Data Toolbox. A menu option is available in the data editor GUI application (Import / EML Data Table) for typing in a URL to an EML document and retrieving a data table in one step. Support for unit inter-conversions based on EML standard units is also being added to the toolbox and will be provided in the next release.

Get provenance information out of NIS

See "add level 0 data" section above. Not currently recommended within MATLAB, but support may be provided in the GCE Data Toolbox in the future.

Submit derived dataset to NIS

See “add level 0 data” section above. Not currently recommended within MATLAB, but support may be provided in the GCE Data Toolbox in the future.

Kepler:

Evaluate EML vs. data entity (congruency check)

Currently not possible with RESTService actor due to special content format requirements of NIS webservice.

Add level 0 dataset to NIS (raw site data)

Currently not possible with RESTService actor due to special content format and user identification requirements of the NIS

Get data out of NIS

The RESTService actor may be used to retrieve the following information from the NIS:

- http://pasta.lternet.edu/package/eml (returns a list of Scopes currently in the NIS, e.g. knb-lter-dog)
- http://pasta.lternet.edu/package/eml/knb-lter-vcr (returns the list of package IDs for this scope)
- http://pasta.lternet.edu/package/eml/knb-lter-vcr/25 (returns a list of available revisions)
- http://pasta.lternet.edu/package/eml/knb-lter-vcr/25/27 (returns the resource map for this dataset including the actual data and metadata URL which both may be retrieved with this actor and saved to the local system)
- https://pasta.lternet.edu/package/data/eml/knb-lter-vcr/25/27/c5b325e8182f30350782fb06be53be7c (25 MB of data)
- https://pasta.lternet.edu/package/metadata/eml/knb-lter-vcr/25/27 (metadata)

The data may also be retrieved with the fileReader and lineReader actors. However, currently the EML2dataset actor cannot access the data directly from the NIS url, but works well with metadata and data files retrieved from the NIS and saved locally. The feature of directly accessing the data may be added in the near future.

Get metadata out of NIS

See above for URL which can be used in the RESTService actor to retrieve metadata.

Get provenance information out of NIS

Currently not possible with REST actor due to special content format and user identification requirements of the NIS

Submit derived dataset to NIS

Currently not possible with REST actor due to special content format and user identification requirements of the NIS

R:

Evaluate EML vs. data entity (congruency check)

Best choice is probably system calls using cURL. RCurl support available for some platforms but has not been tested for upload functions.

Add level 0 dataset to NIS (raw site data)

Best choice is probably system calls using cURL. RCurl support available for some platforms but has not been tested for upload functions.

Get data out of NIS

read.table and read.csv statements accept a http:// URL instead of a local file. Making input routine for rectangular text files in delimited or fixed-column forms. However, this mechanism does not support https:// so access via this mechanism will be limited to "public" data.

A web service has been developed that can automatically generate an "R" program for ingesting and performing rudimentary analyses of "public" data. The "R" code can then be modified by users to perform more advanced statistical and graphical analyses.

http://www.vcrlter.virginia.edu/webservice/statprogPASTA/myprog.r?emlurl=http://pasta.lternet.edu/package/metadata/eml/SCOPE/IDENTIFIER/REVISION

as in:

http://www1.vcrlter.virginia.edu/webservice/statprogPASTA/myprog.r?emlurl=http://pasta.lternet.edu/package/metadata/eml/knb-lter-ntl/5709/2

Get metadata out of NIS

Webservice exists to read EML from NIS and generate R, SAS, SPSS programs for the data. R (SAS, SPSS) program will ingest data, run simple statistics (summary of frequencies, mean, max, min etc.). Some edits are necessary to point to the data source. (John put URL for webservice here)

Get provenance information out of NIS

Can be done, but R is probably not best tool for this.

Submit derived dataset to NIS

None other than system calls. cURL support available but probably not for upload functions.