Publishing Dataset to ORNL Gateway

fwang2@ornl.gov

Contents

1 Overview **1**

2 Preparation On the Gateway **2**

3 Preparation on Data Node **3**
3.1 Stage Data 3
3.2 Obtain Credential 3
3.3 Customize esg.ini 4

4 Project Customization **4**

5 Publishing Steps **4**

6 Unpublish Your Dataset **5**

1 Overview

This is a how to document describing general steps involved to publish a dataset (with AmeriFlux as an example) from ORNL data node to ORNL gateway, which is running ESG system (Earth Grid System). This note is to complement any official ESG data publication document, not a replacement.

In its current form, a successful publication means: you have a dataset stored somewhere at ORNL data node (`esg-tb.ornl.teragrid.org`), which contain one or more data files you want to publish. After going through the following steps, you should be able to browse and search your dataset on the web portal at `http://esg-gw.ornl.teragrid.org`.

2 Preparation On the Gateway

First step, you should register an account at ORNL gateway web portal: `http://esg-gw.ornl.teragrid.org`. The password you pick now will be used when you acquire proxy certificate.

After you receive email notification of successful registration and login, you need to know the project you want to publish under **and** user group that is associated with that project. If none of these are established yet, then you need to ask portal administrator to set it up for you, with the following steps.

These steps requires Gateway Portal root privilege:

- Create top level project "AmeriFlux" with a description, the description is important as it will be used as hyperlink. There was a bug (fixed now since Beta-1) in ESG that if you don't provide a description, you won't be able to get back to it as the hyperlink would be missing.

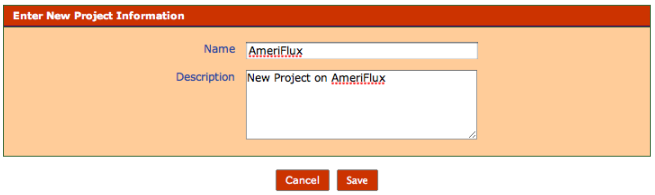

Figure 1: Creating new project

- Create an access group, "AmeriFlux User Group".

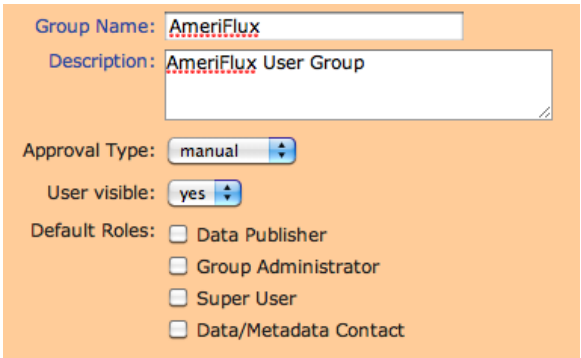

Figure 2: Creating user group

- Under the project just created, create a top level collection, with identifier `ornl.ameriflux`, the identifier needs to be unique on gateway system

- Now you need to edit "metadata properties" for this collection and associate access control group "AmeriFlux User Group" with it.

If the admin has done all above for you, what's left for you to do as a data publisher is to make sure you have the right **role** for the project and user group you want to publish. You need to subscribe for a role. ESG doesn't allow you to specify what role you want to subscribe, yet.

Group Name ▲	Description
AmeriFlux	AmeriFlux User Group
ARM	ARM Users Group
CDIAC	CDIAC Users Group
C-LAMP	C-LAMP Users Group

Subscribe

3 Preparation on Data Node

3.1 Stage Data

If you want to store your data to `/lustre/esgfs/projects`. You need write permission and you must run the following command to switch your group ID:

```
newgrp esg-publishers
```

3.2 Obtain Credential

Before anything else, make sure you have set proper environment variables, see Appendix for a template, specifically for current ORNL configuration.

Next, you need to obtain a fresh proxy certificate. As you push the thredds catalog to the gateway at the final step, the proxy certificate is checked for authentication purpose.

```
% myproxy-logon -o ~/mycert/proxycert
```

You can manually check if this certificate *looks* right by:

```
% openssl x509 -text -in ~/mycert/proxycert
```

You should expect to see some roles defined, such as:

```
X509v3 extensions:
     1.2.3.4.4.3.2.1.7.8:
    ...esg.vo.group.roles=group_ARM_role_default;group_ARM_role_publisher;\
    group_FeiyiTest_role_default;group_FeiyiTest_role_publisher;\
   group_User_role_default:esg.vo.openid=https://esg-gw.ornl.teragrid.org/myopenid/fwang2
```

3.3 Customize esg.ini

If you have never published using ORNL data node before, you should run the following command to get a few template files in your `$HOME/.esgcet`, which should include: `esg.ini`,

cf-standard-name-table.xml and esgcet_models_table.txt.

```
% esginitialize -i
```

However, as each site may have different configuration setup and this has to reflected in each esg.ini - it is better to start with a customized template for ORNL site, which is given in the appendix, along with an example for AmeriFlux project.

Next, I'd assume you have saved your data to somewhere on the ORNL data node, and that location needs to be defined in esg.ini.

```
thredds_dataset_roots =
    ameriflux | /data2/esgdata/cdiac/ameriflux_test1
```

When starting a new project, decide a project name, and define it in esg.ini. If there are other projects defined in the template you are using, it is advisable to remove or comment them out first - to limit the scope if things go wrong.

```
project_options:
    ameriflux | AmeriFlux Data  | 1
```

Here, we define project as ameriflux, the number at the last field defines the **search order** this project is handled when multiple projects are being published at once.

4 Project Customization

For each project, a project specific section needs to be defined: See appendix for a complete sample.

```
[project:ameriflux]
directory_format = /data2/esgdata/cdiac/ameriflux_test1/%(experiment)s
dataset_id = %(root_id)s.%(project)s.%(experiment)s
experiment_options =
    ameriflux | ARM_SGP_Burn | notes
    ameriflux | Fort_Dix | notes
```

This is one of the murky areas as far as whole publication process goes. And I need to expand this section. Overall, it can be figured out by reading Bob Drach's page at http://www2-pcmdi.llnl.gov/Members/bdrach/.personal/esg-publisher-configuration/

5 Publishing Steps

Step. 1 Add to database

```
% esginitialize -c
```

By running this, whatever you put in the experiment_options will be added to the databases.

Although it may not report on the console, it seems the esginitialize also add project and its model information to ”project” and ”model” table, respectively.

Step. 2 Scan data

```
% esgscan_directory --project ameriflux \
    --dataset ornl.ameriflux.armburn \
    /data2/esgdata/cdiac/ameriflux_test1/ARM_SGP_Burn > armburn.txt
```

Step. 3 Parse datafile and write metadata to DB

```
% esgpublish --map armburn.txt --experiment ARM_SGP_Burn
```

Step. 4 Publish to local THREDDS

```
% esgpublish --use-existing ornl.ameriflux.armburn --noscan --thredds
```

Step. 5 Publish to remote THREDDS to ORNL Gateway

```
% esgpublish --use-existing ornl.ameriflux.armburn --noscan --publish
```

6 Unpublish Your Dataset

For whatever reason, you might want to retract a publication to republish it. If you have followed above steps, the following command should do the trick:

```
% esgunpublish --map armburn.txt
```

Again, you should expect to see “SUCCESSFUL” message on the console.

In the case that you don’t have a map file anymore, here is an easy way out: Identify all dataset ids and put them in a plain file, for example:

```
ornl.arm.data.mytest
ornl.arm.data.hpsstest
```

Then run command:

```
% esgpublish --use-list dataset.txt
```