

Bulk Upload with DERIVA Client Tools

Our underlying software system, DERIVA, has client tools for authenticating (DERIVA-Auth) and performing a bulk upload of sequencing data files (DERIVA-Upload). This is recommended if you have a large or complex sequencing data submission.

There are two versions of the client tool:

- [a graphical interface that can be run to upload files from your desktop system](#), and
- [a command-line interface that can be used to upload files from a remote server](#).

Although the process for downloading and running the above tools are different, they both use the same directory structure designed for different data types. So make sure you've read the [Organize files for bulk upload](#) section first.

Note: Currently these instructions focus on uploading sequencing data. We will be adding the ability to upload specimen imaging data soon.

Overview of Bulk Upload Process

The following are the basic steps for using the client tools to upload data files.

1. [Join the kidney-writers group](#)
2. [Organize files for bulk upload](#)
 - [2.1. Supported file types](#)
 - [2.2. Set up the directories](#)
 - [2.3. Choose naming conventions for replicate-level files](#)
 - [2.4. Choose naming conventions for study-level files](#)
3. [Download and install DERIVA Client](#)
4. [Using the GUI from a Desktop](#)
 - [4.1. Launch and Configure DERIVA-Upload](#)
 - [4.2. Upload files](#)
 - [4.3. Log out](#)
5. [Using the deriva-upload-cli command from a remote server](#)
 - [5.1. Get an authentication token from DERIVA Auth](#)
 - [5.2. Upload files with deriva-upload-cli](#)
 - [5.3. Log out](#)

Here are the details for each step.

1. Join the kidney-writers group

- Join the [kidney-writer group](#) by [clicking on this link](#).
- When you click the [kidney-writer group](#) link, if you have never used Globus before, you will be given various choices for logging in: via existing credentials (your institution, Google, or ORCID ID) or by creating a new Globus ID. We recommend using an existing credential if that is available.

- For detailed instruction on how to join different GUDMAP/RBK group, visit [Accessing GUDMAP and RBK Resources](#). If you have *any* problems, please email help@gudmap.org or help@rebuildingakidney.org.

2. Organize files for bulk upload

The upload tools will use the names of the directories and files (ie, folders) to determine what kind of files you are uploading and which metadata records to attach them to.

2.1. Supported file types

The following sequencing data file types are supported for uploading to the GUDMAP/RBK repository:

| Extension | File Type | Description (will appear in file caption) | mandatory |
|-------------|-----------|---|--------------------------|
| R1.fastq.gz | FastQ | F reads | mandatory |
| R2.fastq.gz | FastQ | R reads | mandatory for Paired-End |
| bam | bam | alignment | mandatory |
| bed | bed | positive regions | optional |
| bw | bigWig | visualization track | optional |
| rpkmt.txt | txt | expression value | optional |
| tpm.txt | txt | expression value | optional |

2.2. Set up the directories

The directory structure will use the "Internal ID" you chose for the Study and Experiments. Here's an example of the required directory structure.

```
$userid
  \- deriva
      \- Seq
          \- <Study Internal ID>
              \- <Experiment Internal ID>
```

where

- `deriva` is the name of our software
- `Seq` is a subfolder of `deriva`. This indicates that everything within is sequencing data (e.g. non single-cell and single cell)
- `<Study Internal ID>` is the Internal ID specified in the Study metadata record (e.g. `NPC_stability`).

- `<Experiment Internal ID>` is Internal ID specified in the Experiment metadata record (e.g. `mNPC_RNA`).

2.3. Choose naming conventions for replicate-level files

Within each experiment directory, you will add the data files for the replicates for that experiment using one of two naming conventions.

Note: Remember we need a replicate for each unit of tissue used in an experiment. For example, three kidney biopsies from a single person would be three replicates from a single kidney of that person.

The two file naming conventions to choose from are:

```
<Biological Replicate Number>_<Technical Replicate Number>_<Custom Text>.R1.fastq.gz
```

Or

```
<Replicate RID>_<Custom Text>.R1.fastq.gz
```

As you can see, you can either use:

- A combination of `<Biological Replicate Number>` with `<Technical Replicate Number>` (e.g. `1_1`).
- Or the `<Replicate RID>` which is the RID (Resource ID) of the replicate record (e.g. `Q-Y500`).

In either case, you can use custom text of your choice (e.g. `sorted`) after the numbering option to help distinguish the files for the user.

You can also use either file naming convention between different experiment directories in the same study.

Example 1:

If you have a study named "NPC_stability" with experiments named "mNPC_RNA" and "mNPC_ATAC". you'd create two folders:

- `deriva/Seq/NPC_stability/mNPC_RNA`
 - (on Windows the path would be `deriva\Seq\NPC_stability\mNPC_RNA`)
- `deriva/Seq/NPC_stability/mNPC_ATAC`
 - (on Windows, the path would be `deriva\Seq\NPC_stability\mNPC_ATAC`).

You would then place all your sequencing files into their respective experiment folders.

In the example below, we use the *Biological Replicate Number* and *Technical Replicate Number* convention to name the files in experiment "mNPC_RNA" and use the Replicate *RID* convention to name the files in the experiment "mNPC_ATAC".

Both file naming conventions will be accepted by the client tool.

See actual examples of metadata and files in [the NPC_stability Study](#).

```

$userid
  \- deriva
    \- Seq
      \- NPC_stability
        \- mNPC_RNA
          \- 1_1.R1.fastq.gz
          \- 1_1.R2.fastq.gz
          \- 1_1_sorted.bed
          \- 1_1_normalized_profile.bw
          \- 1_1.kpm.txt
          \- ...
        \- mNPC_ATAC
          \- Q-Y5CC.R1.fastq.gz
          \- Q-Y5CC.R2.fastq.gz
          \- Q-Y5CC_sorted.bed
          \- Q-Y5CC_normalized_profile.bw
          \- Q-Y5CC.kpm.txt
          \- ...

```

Example 2:

If you have a single-cell RNA study named "mouse_SC_RNASeq" with experiments "m1_e11_cortex" and "m2_p0_cortex", you would create two folders:

- `deriva/scRNASeq/mouse_SC_RNASeq/m1_e11_cortex`
 - (on Windows, the path would be `deriva\scRNASeq\mouse_SC_RNASeq\m1_e11_cortex`)
- `deriva/scRNASeq/m2_p0_cortex`
 - (on Windows, the path would be `deriva\scRNASeq\m2_p0_cortex`).

You would place the sequencing files into their respective experiment folders.

In the example below, we use the *Biological Replicate Number* and *Technical Replicate Number* convention to name the files in experiment "m1_e11_cortex" and use the Single Cell Metrics *RID* convention to name the files in the experiment "m2_p0_cortex".

Both file naming conventions will be accepted by the client tool.

See actual examples of metadata and files in [the mouse_SC_RNASeq Study](#).

```

$userid
  \- deriva
    \- Seq
      \- mouse_SC_RNASeq
        \- m1_e11_cortex
          \- 1_1.R1.fastq.gz
          \- 1_1.R2.fastq.gz
          \- 1_1_sorted.bed
          \- 1_1_normalized_profile.bw
          \- 1_1.kpm.txt
          \- ...
        \- m2_p0_cortex
          \- Q-Y4HM.R1.fastq.gz
          \- Q-Y4HM.R2.fastq.gz
          \- Q-Y4HM_sorted.bed
          \- Q-Y4HM_normalized_profile.bw
          \- Q-Y4HM.kpm.txt
          \- ...

```

2.4. Choose naming conventions for study-level files

If there are analysis files that are results of combining data from different experiments together, those files can be added to the **Study Analysis File** table. Follow the directory layout below to upload those files through the DERIVA client tools.

The layout of folders for analysis files to be linked at the Study level is:

```

$userid
  \- deriva
    \- Seq
      \- <Study Internal ID>
        \- <file 1>
        \- <file 2>

```

where:

- **deriva** is the name of our software
- **Seq** is a subfolder of **deriva**. This indicates that everything within is sequencing data
- **<Study Internal ID>** is what is specified in the study metadata e.g. **mouse_SC_RNASeq**
- **<file 1>**, **<file 2>** are files that you want to uploaded so they can be linked to the study.

Example 3:

If there are analysis files that are the result of combining data from different experiments together, those files can be added as follows in the Study directory:

```
$userid
  \- deriva
    \- Seq
      \- mouse_SC_RNASeq
        \- analysis.xls
          \- cluster.html
```

3. Download and install DERIVA Client

DERIVA Client is a suite of tools that include DERIVA Upload and DERIVA Auth.

Download and install the [latest version of DERIVA Client here](#). There are pre-packaged installers available for Mac or Windows desktops or you can install `deriva-client` from pip for Linux desktops or remote servers/clusters.

- DERIVA Upload - Allows you to choose a directory and upload all of the files within it.
- DERIVA Auth - Authenticates your submission if you are using DERIVA Upload on a remote server.

These are available both as GUI apps (for Windows or MacOS desktop) and command line utilities (for Linux desktops and remote servers/clusters).

You can find the GUI apps for Windows or MacOS in your Applications menu on Windows or MacOS under the "DERIVA Client Tools" folder.

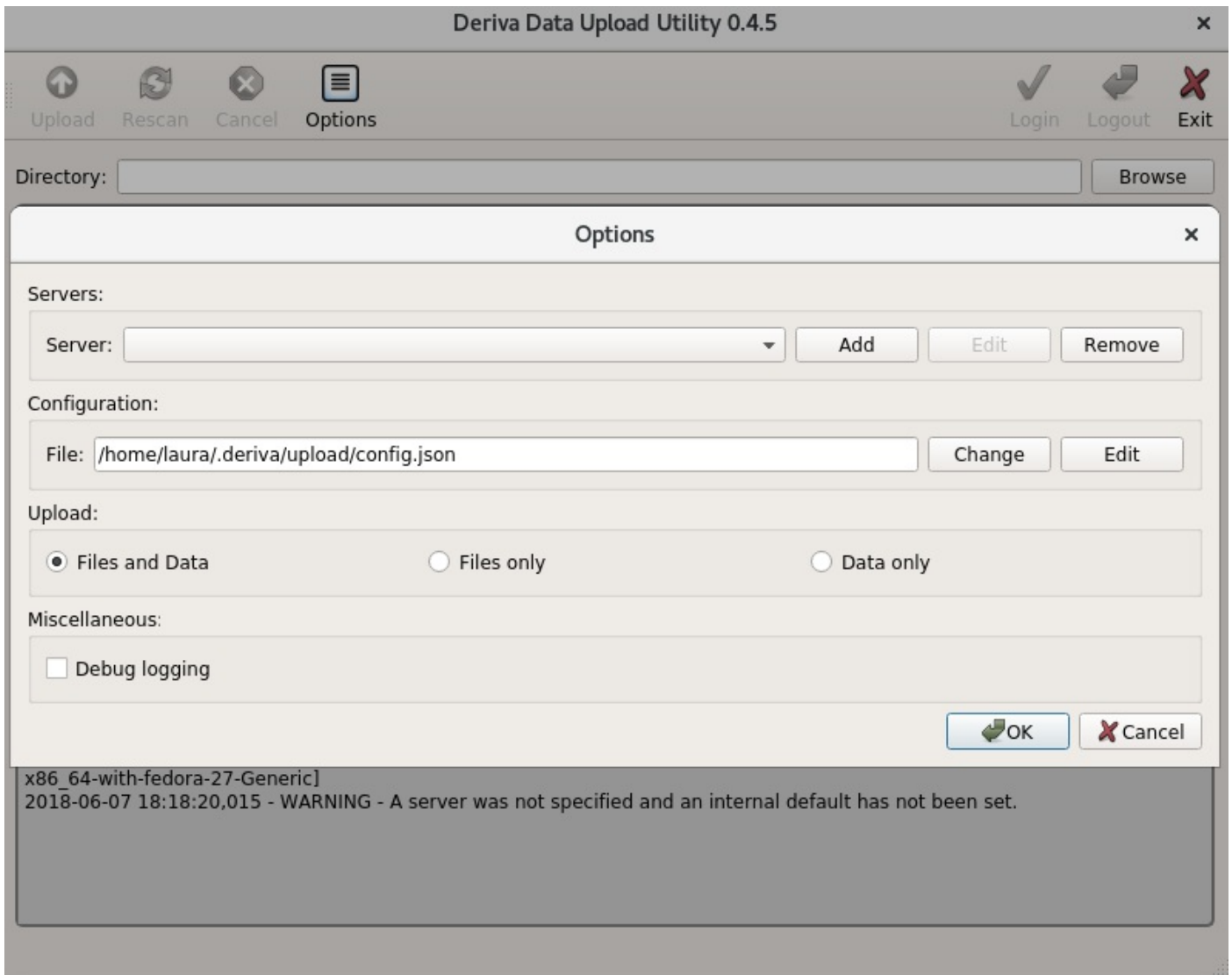
4. Using the GUI from a Desktop

4.1. Launch and Configure DERIVA-Upload

Launch the DERIVA-Upload app (through the Applications menu for Windows or MacOS desktops) or run the `deriva-upload` command (for Linux desktops).

The first time you launch it, the tool will ask you if you want to add a server configuration.

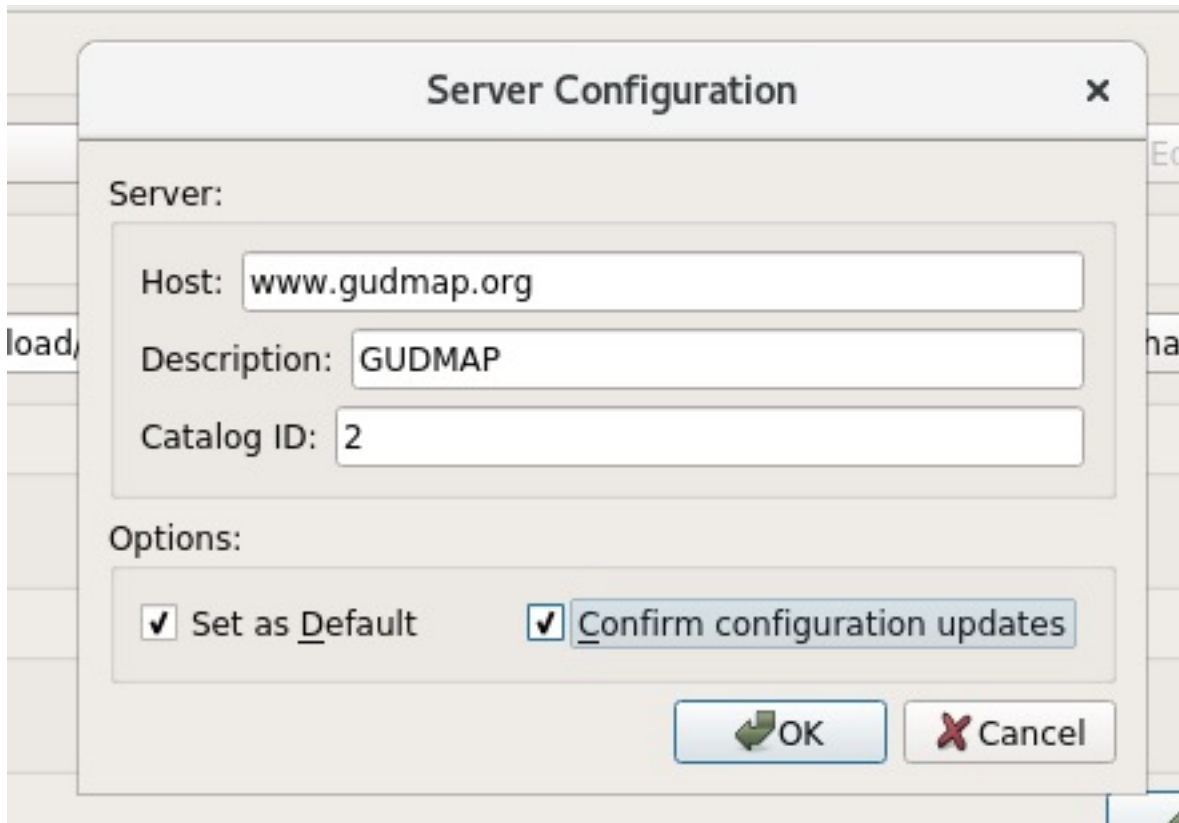
Click "Yes" to bring up the "Options" screen. You can also do this at any time by clicking the "Options" button at the top of the page.



Click **Add** to bring up the "Server Configuration" form and enter these values:

- Host: **www.gudmap.org** or **www.rebuildingakidney.org**
- Description: **GUDMAP** or **RBK**
- Catalog ID: 2

Check the "Set as Default" and "Confirm configuration updates" buttons, and click "OK".



The image shows a 'Server Configuration' dialog box with a title bar containing a close button (X). The dialog is divided into two main sections: 'Server:' and 'Options:'. The 'Server:' section contains three text input fields: 'Host:' with the value 'www.gudmap.org', 'Description:' with the value 'GUDMAP', and 'Catalog ID:' with the value '2'. The 'Options:' section contains two checked checkboxes: 'Set as Default' and 'Confirm configuration updates'. At the bottom right of the dialog are two buttons: 'OK' (with a green arrow icon) and 'Cancel' (with a red X icon).

Server Configuration

Server:

Host:

Description:

Catalog ID:

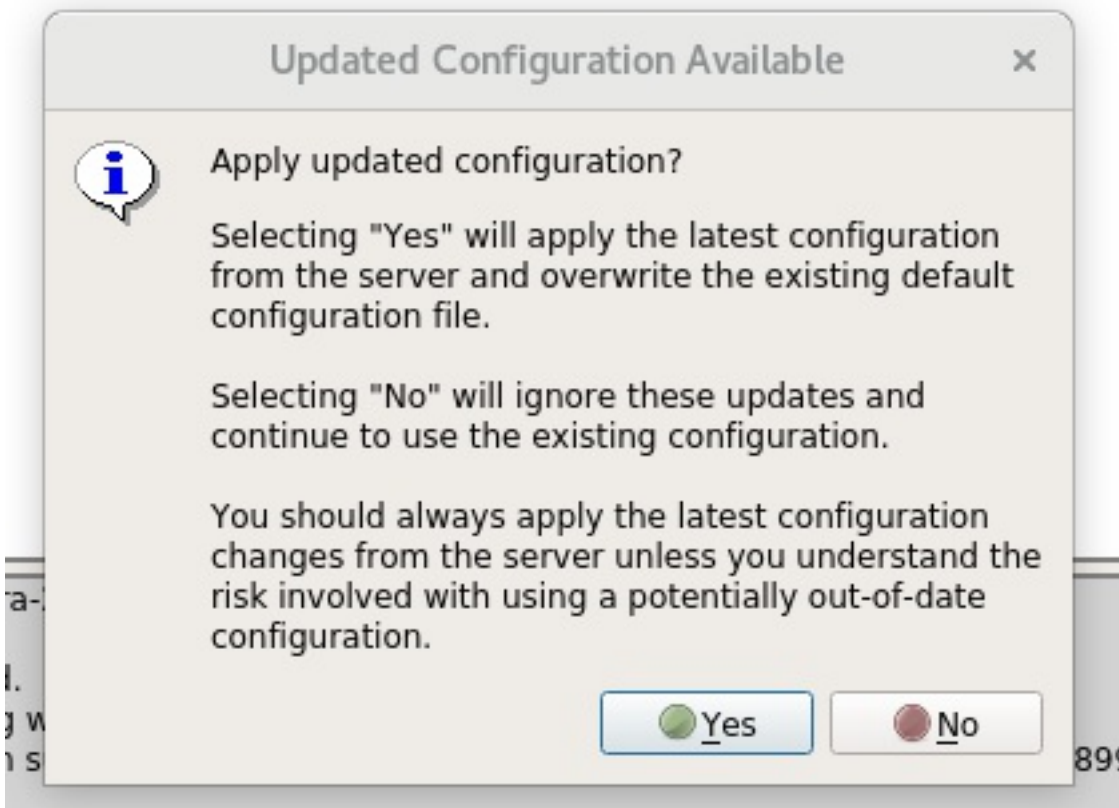
Options:

☒ Set as Default ☒ Confirm configuration updates

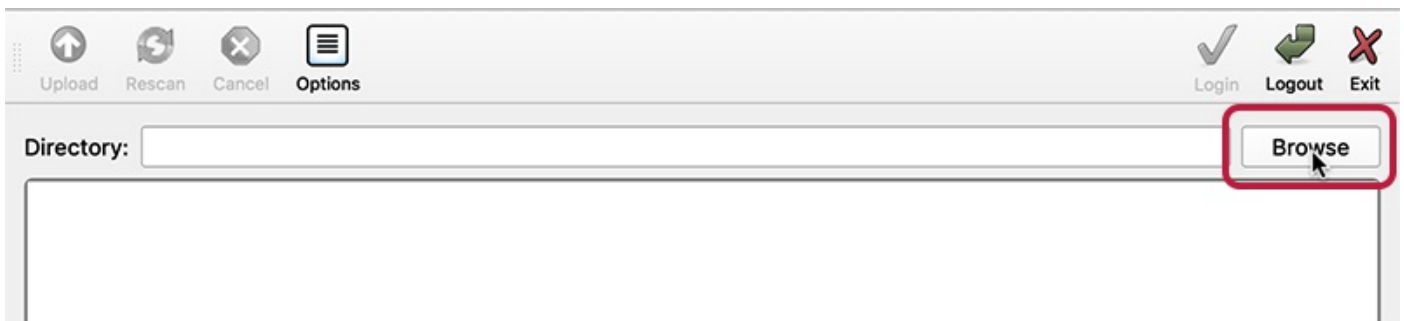
4.2. Upload files

In the main DERIVA Upload window, click the "Login" button at the top to log in. This will pop up a login dialog window.

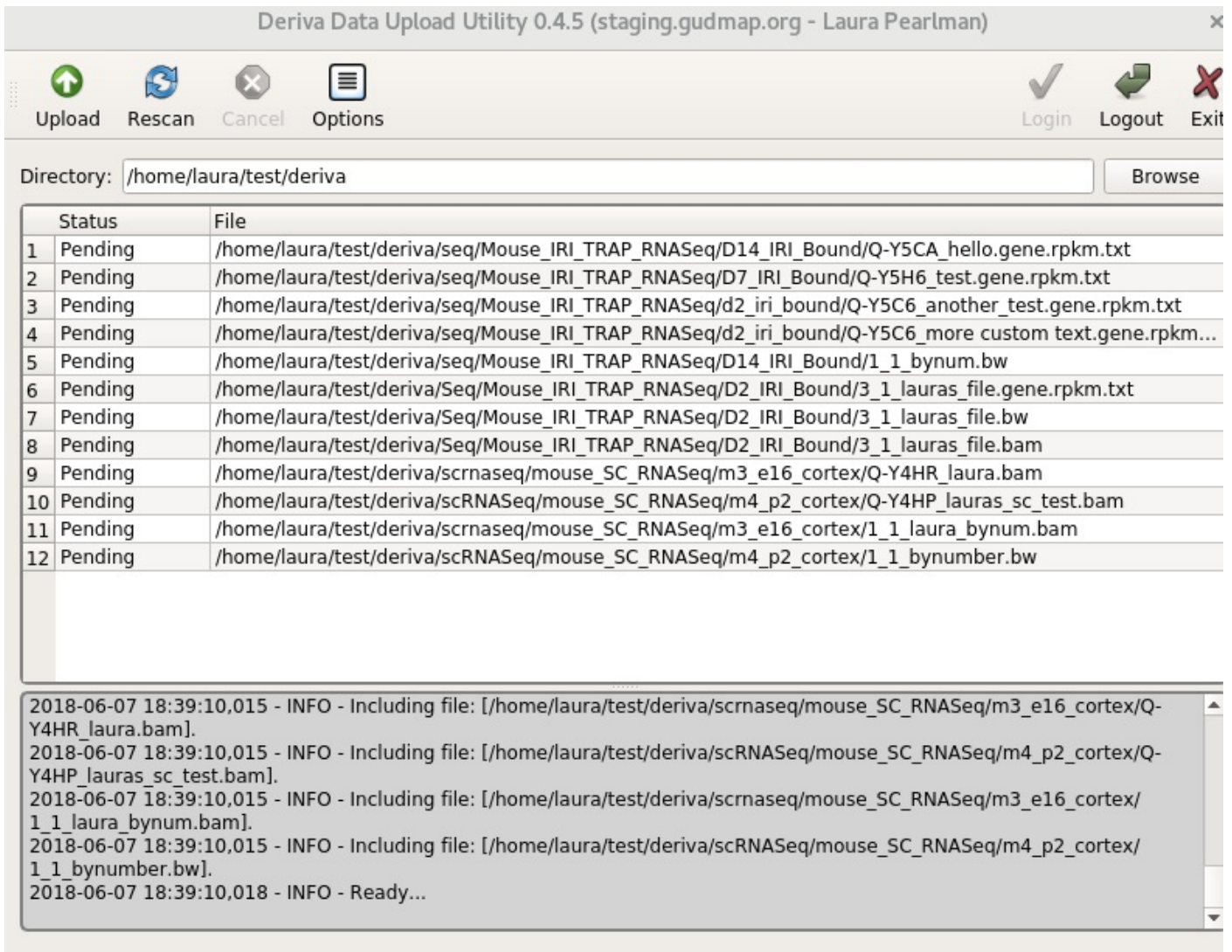
Once you've logged in, you may see a window notifying you that an updated configuration is available and asking if you'd like to apply it. Click "Yes" to update your configuration and dismiss the window.



In the main DERIVA-Upload window, click the "Browse" button.



Select the **deriva** directory you created above. You'll see all the files in your directory structure listed as "Pending".



Click the "Upload" button to start the upload process. The status of each file will change as it is uploaded. For successful uploads, the status will change from "Pending" to "Complete".

If for some reason, your upload is interrupted (ie, a network outage), DERIVA-Upload will re-try uploading a few times. If the upload is terminated, you can click the "Upload" button again and the system will automatically know which files were already uploaded successfully and skip them.

4.3. Log out

Authentication tokens expire after 30 minutes of activity but if you want to log out explicitly, click the "Logout" button at the top of the window.

5. Using the `deriva-upload-cli` command on a remote server

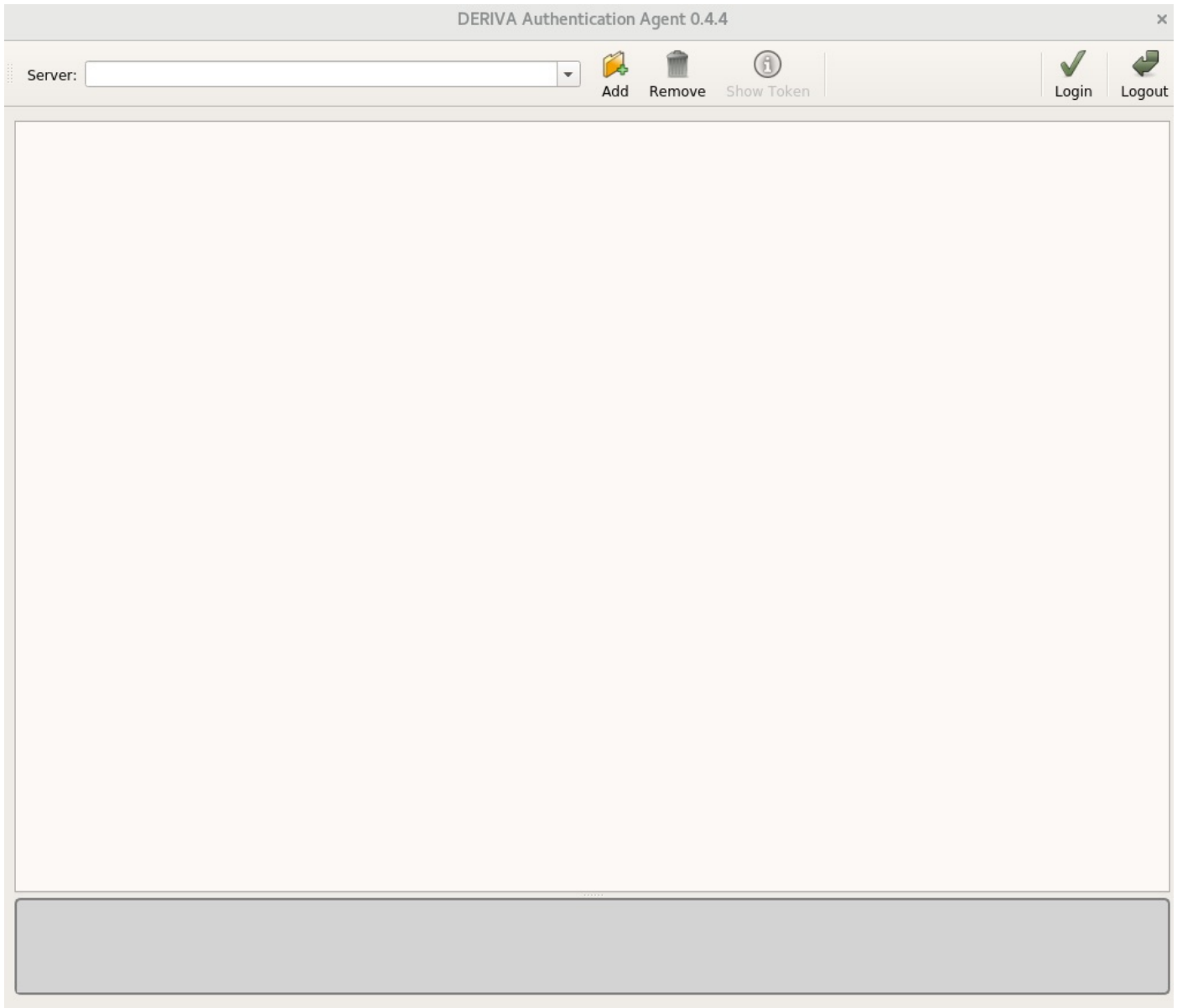
Using the command-line interface on a remote server is a bit more complicated. First, you'll need to get an authentication token by running the DERIVA-Auth tool locally on your desktop. Then you'll run the command-line tool on the remote server.

5.1. Get an authentication token from DERIVA Auth

The uploader requires an authentication token to communicate with the server.

Launch the DERIVA-Auth tool on your desktop (through the Applications menu on Windows or Mac, or with `deriva-auth` on Linux) to bring up an authentication window similar to the one used in the data browser.

The first time you log in, you'll see a mostly-empty window:



In the "Server:" area, type in the name of the target server (`www.gudmap.org` or `www.rebuildingakidney.org`) and click on `Add` . You should now see something that looks similar to the data browser login screen

The screenshot shows a window titled "DERIVA Authentication Agent 0.4.4". The interface includes a "Server:" dropdown menu with "www.gudmap.org" selected. To the right of the dropdown are three buttons: "Add" (with a folder icon), "Remove" (with a trash icon), and "Show Token" (with an information icon). Further right are "Login" (with a green checkmark icon) and "Logout" (with a green arrow icon). Below the server dropdown is a browser address bar showing "www.gudmap.org". The main content area has a blue header with the "globus" logo and a "Globus Account Log In" link. The main text reads "Log in to use RBK/GUDMAP" followed by "Use your existing organizational login" and "e.g., university, national lab, facility, project". There is a dropdown menu labeled "Look-up your organization...". Below this is a link: "Didn't find your organization? Then use [Globus ID to sign in.](#) (What's this?)". A "Continue" button is present. Below the "Continue" button is an "Or" separator. At the bottom are two buttons: "Sign in with Google" (with the Google logo) and "Sign in with ORCID iD" (with the ORCID logo). At the very bottom, a log window shows two messages: "2018-06-07 19:03:27,772 - INFO - Initializing authorization provider: AuthWidget v0.4.4 [Python 3.6.5, Linux-4.16.6-202.fc27.x86_64-x86_64-with-fedora-27-Generic]" and "2018-06-07 19:03:27,932 - INFO - Authenticating with host: https://www.gudmap.org".

Server:

Add Remove Show Token Login Logout

www.gudmap.org

globus Globus Account Log In

Log in to use RBK/GUDMAP

Use your existing organizational login

e.g., university, national lab, facility, project

Didn't find your organization? Then use [Globus ID to sign in.](#) (What's this?)

Continue

Or

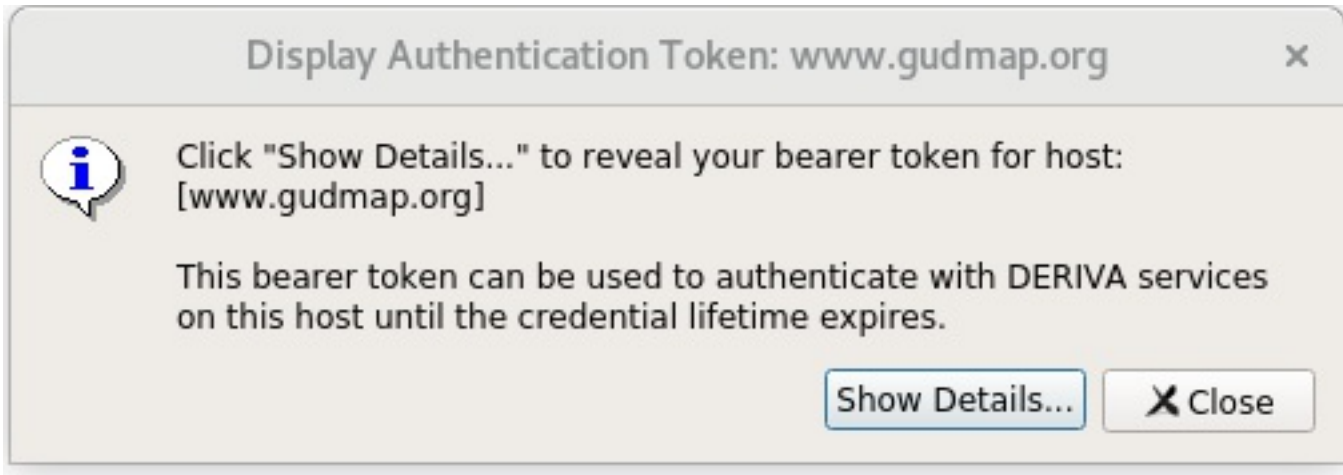
Sign in with Google Sign in with ORCID iD

2018-06-07 19:03:27,772 - INFO - Initializing authorization provider: AuthWidget v0.4.4 [Python 3.6.5, Linux-4.16.6-202.fc27.x86_64-x86_64-with-fedora-27-Generic]
2018-06-07 19:03:27,932 - INFO - Authenticating with host: https://www.gudmap.org

Note: In subsequent runs, DERIVA-Auth might take you directly to this window (skipping the blank screen at the beginning). It's always a good idea to look at the server URL before you log in.

After logging in, you'll see an "Authentication Successful" message. Click the "Show Token" button.

This will bring up another dialog box to verify that you really want to view the token. Click on "Show Details" to display the token. Copy and store for use in the upload command.



5.2. Upload files with `deriva-upload-cli`

On the server, run the command:

```
deriva-upload-cli --catalog 2 --token token --catalog 2 host /path/to/deriva
```

where:

- *token* is the token copy-and-pasted from your DERIVA-Auth session
- *host* is `www.gudmap.org` or `www.rebuildingakidney.org`, and
- */path/to/deriva* is the path to the `deriva` directory you created above.

For example:

```
deriva-upload-cli --catalog 2 --token xxxxxxxXxxxxXxxxxX www.gudmap.org
$HOME/deriva
```

5.3 Log out

Authentication tokens expire after 30 minutes of activity but if you want to log out from DERIVA Auth explicitly, click the "Logout" button at the top of the window.