GRIIDC Dataset Submission User Guide

Note: Before a dataset can be submitted to the GRIIDC data management system, it must first be identified through a Dataset Information Form (DIF). To complete and submit a DIF for your dataset please visit https://data.gulfresearchinitiative.org/dif. You will be able to submit your dataset file after you receive notification that your DIF has been reviewed and approved by GRIIDC staff.

Contents

Section 1: What is a Dataset Submission?

GRIIDC has developed a data management system to store datasets and related information collected and generated by researchers. Datasets can be submitted to the GRIIDC data management system directly or if housed at a National Data Archive (e.g., National Centers for Environmental Information) a link to the dataset may be provided. Dataset submission is the process that researchers use to submit data directly to the GRIIDC data management system or to notify GRIIDC that a dataset is available through a National Data Archive.

Section 2: What Do I Need Before I Can Submit a Data File?

1. User Account

In order to complete the dataset identification and submission process, you must first have an account with the GRIIDC data management system. To request an account please visit https://data.gulfresearchinitiative.org/pelagos-symfony/account. If you have a GRIIDC account, please visit the GRIIDC website (https://data.gulfresearchinitiative.org/) and click the login link in the upper right corner of the screen to open the log in form. Enter the username and password for your account.

2. Dataset Information Form (DIF)

Data providers must first have an approved Dataset Information Form (DIF) for a dataset before it can be submitted to GRIIDC. In order to complete and submit a DIF for your dataset please visit https://data.gulfresearchinitiative.org/dif.

3. Determine the Data File Submission Method

Before you begin the dataset submission process, you should determine the method that you will use to submit your dataset file to the GRIIDC data management system. The GRIIDC data management system supports a number of methods that you may use to submit your dataset file. The method you use to submit your dataset file will be based on a number of factors, including the size of your dataset file, your technical expertise, and the location of your dataset file. The following list of questions can help you determine how to submit your dataset file.

- Do you have a small dataset file (less than 20 GB) on your own computer? Use Direct Upload.
- Have you submitted a dataset file to a National Data Archive? Use Request Pull from HTTP.
- Is your dataset file posted on an institutional website that is available to the public? Use Request Pull from HTTP.
- Is your dataset file posted on an institutional FTP server that is available to the public? Use FTP.
- Do you have a large dataset file you wish to push to GRIIDC (greater than 20 GB)? Use SFTP.
- Do you have a very large dataset file you wish to push to GRIIDC (greater than 20 GB)? Use GridFTP.

FTP, SFTP, and GridFTP are methods that require servers or personal computers to be configured appropriately; therefore, you may be required to consult with your institutional IT department before using these methods. Section 5: Dataset Submission – Dataset File Transfer Details provides additional information about how to set up and utilize each method to transfer your dataset file to the GRIIDC data management system. If you have questions about what method you should use to submit your dataset to the GRIIDC data management system, please contact griidc@gomri.org.

NOTE: If your dataset is composed of multiple data files, these files must be packaged together into a single file for submission to the GRIIDC data management system (e.g., creating a zip file). For more information, please see GRIIDC Compression Guidance documents for Linux, Windows, and Mac. If your dataset is a single file (e.g., Excel file) the file can be uploaded directly and should not be compressed into a zip file. Please use consistent and descriptive names when naming files. For more information about dataset file name recommendations please see Appendix A.

Section 3: Dataset Submission – Introduction

After a data provider has created an account, received email notification that a Dataset Information Form (DIF) has been approved, and determined the method to submit the data file, they can continue to Dataset Submission: https://data.gulfresearchinitiative.org/dataset-submission.

At the top of the Dataset Submission start page in the text box below "Unique Dataset Identifier (UDI)" enter the UDI for the approved Dataset Information Form (DIF) that you will be submitting data for and select "Load Dataset". Alternatively, you can search for and select your dataset from the list of approved Dataset Information Forms (DIFs) displayed below the UDI field. This will pre-fill a number of fields including Datset Contact, Dataset Title, Dataset Abstract, Supplemental Information-Data Parameters and Units, and Data Extent with information provided when you completed your DIF. Please update these fields as appropriate. At any time you can select the Save and Continue later button if you are not ready to submit your data. For more details about what information should be provided in the fields, hover over the information icon on the webpage.

Section 4: Dataset Submission – Field Descriptions and Examples

The Dataset Submission page consists of four tabs of information. Each tab focuses on a different component of descriptive information about the dataset. Required fields are indicated with a red asterisk. A data file cannot be submitted until all the required fields are entered. A red X will be displayed on a tab if required fields have not been filled out. When all required information has been entered for a tab, a green check will appear on the tab. Once all required information has been entered for every tab, indicated by green checks on each tab, a data file can be submitted.

Dataset Contact

The Dataset Contact tab collects information about the individual who is responsible for answering questions about the dataset. Usually this is the primary investigator of the project.

Primary Point of Contact*: (Required) The person you designated as the primary point of contact in the DIF will be entered into this field. You may select another person associated with your project using the drop-down list. If a member of your research group or project is not listed in the drop-down list or if the contact information displayed is incorrect, please contact griidc@gomri.org. You may also select additional points of contact by selecting the "Add Contact" button.

Role*: (Required) This field is the person's relationship to the dataset. There are three options: Point of Contact, Principal Investigator, and Author. Point of contact is the person who can be contacted for acquiring information about the data or to acquire the data. Principal investigator is the key person

responsible for gathering information and conducting research. Author is the person who authored the dataset.

Dataset Author(s)*: (Required) This field is a list of authors who should be acknowledged if these data are cited in published materials. Please use the Modern Language Association (MLA) style, demonstrated below, when listing dataset authors.

Example Dataset Author(s)
One originator: Last Name, First Name (e.g. Smith, Joe)
Two originators: Cross, Susan and Christine Hoffman
Three originators: Lowi, Theodore, Benjamin Ginsberg, and Steve Jackson
More than three originators: Gilman, Susan *et al.*

Dataset Information

The Dataset Information tab collects information that allows others to understand your dataset.

Dataset Title* (Required)**:** Please provide a descriptive title that briefly explains the contents of your dataset and, if applicable, dates and geographic area. It should be understandable by a user unfamiliar with your methods, collection sites, or research platforms.

Example Titles

(1) Ecological
Aerial survey data for the assessment of the distribution of cownose rays (*Rhinoptera bonasus)* in the Eastern Gulf of Mexico, from May to October 2020

(2) Chemical/Molecular Engineering
Image sequences of rising bubble plumes from laboratory study to calculate bubble velocity vector under various gas flow rates and vertical density gradients

(3) Oceanographic
Conductivity, temperature, and depth data for 12 northwestern Gulf of Mexico locations, May to July 2020

(4) Model/Numerical Model
Galveston Bay Circulation Study: Stanford Unstructured Nonhydrostatic Terrain following Adaptive Navier-Stokes (SUNTANS) models simulations for 2020-2021

(5) Sociology
Cross-sectional household survey response data to assess health and wellbeing of residents of coastal Louisiana, April 2020

(6) Genetics
Mitochondrial DNA control region sequences (297 base-pairs) from 140 northern Red Snapper (*Lutjanus campechanus*) collected from the Gulf of Mexico, 2018–2020

(7) Biochemistry
Polycyclic Aromatic Hydrocarbon concentrations in liver and muscle tissue from barracuda, escolar, and common dolphin fish, northeastern Gulf of Mexico, 2020-2021

Short Title: This is an alternative short name or other language name by which the dataset might be known. This could be the title that the lab references the dataset by and could include abbreviations or cruise and platform names.

Dataset Abstract*: (Required) A narrative summary of the dataset's contents. The dataset abstract should summarize what data the dataset contains, methods used to collect or generate the dataset, and time period and location of data collection. This may be similar to a publication abstract; however, it does not need to include details about results, conclusions, or statistical analysis completed using the dataset. It should address the questions:

- What data have been collected and/or generated?
- How have the data been collected and/or generated?
- When were the data collected and/or generated?
- Where were the data collected and/or generated? If model data, the location for which the dataset was generated.

If your dataset relates to another dataset you have provided to GRIIDC, please provide the DOI of the related dataset in the abstract. Additionally, if you would like to reference a publication that uses the dataset, please include the full reference of the publication in the abstract.

Purpose* (Required): Summary of the reasons or intentions for which the dataset was created or generated.

Supplemental Information – Data Parameters and Units* (Required): Please provide descriptions of all reported data parameters/variables, list and define abbreviations for each parameter/variable, and define units of measurement for each parameter. If you provide multiple files that report different parameters/variables, indicate which files report which parameters/variables.

Supplemental Information – Methods* (Required)**:** Please provide a description of the methods used to create and/or generate the data in the dataset.

Supplemental Information – Instruments: Please provide a description of the instruments and equipment used to create and/or generate the data in the dataset.

Supplemental Information – Sampling Scales and Rates: Please provide a description of the spatial and temporal scales and rates that were used to collect and/or generate the data, if applicable.

Supplemental Information – Error Analysis: Please provide a description of any error or uncertainty analysis completed on the final data and the results of the error analysis.

Supplemental Information – Provenance and Historical References: If existing historical data were used as part of the dataset, please provide a description of the historical data used, including who created the original dataset (person and/or organization) and from where the historical data can be obtained.

Keywords

Theme Keywords* (Required): Please include commonly used words or short phrases that describe the themes or subjects related to the dataset. These may be the keywords used to describe the publication associated with the dataset. Do not include keywords that relate to the location or geography of the data.

Place Keywords* (Required): Please include commonly used words or short phrases that describe the geographic areas, locations, or places that are associated with the data. Leave blank if the dataset does not have a relevant place keyword (e.g. lab mesocosms).

Topic Category Keyword* (Required): These keywords are pre-defined by the ISO 19115-2 metadata standard that GRIIDC uses to describe datasets. Please select all the relevant topic categories from the keyword list. Hover over the keyword to see a definition.

Data Extent

All fields in this tab are required. This tab collects information about the spatial and temporal extent of your data (Figure 1).

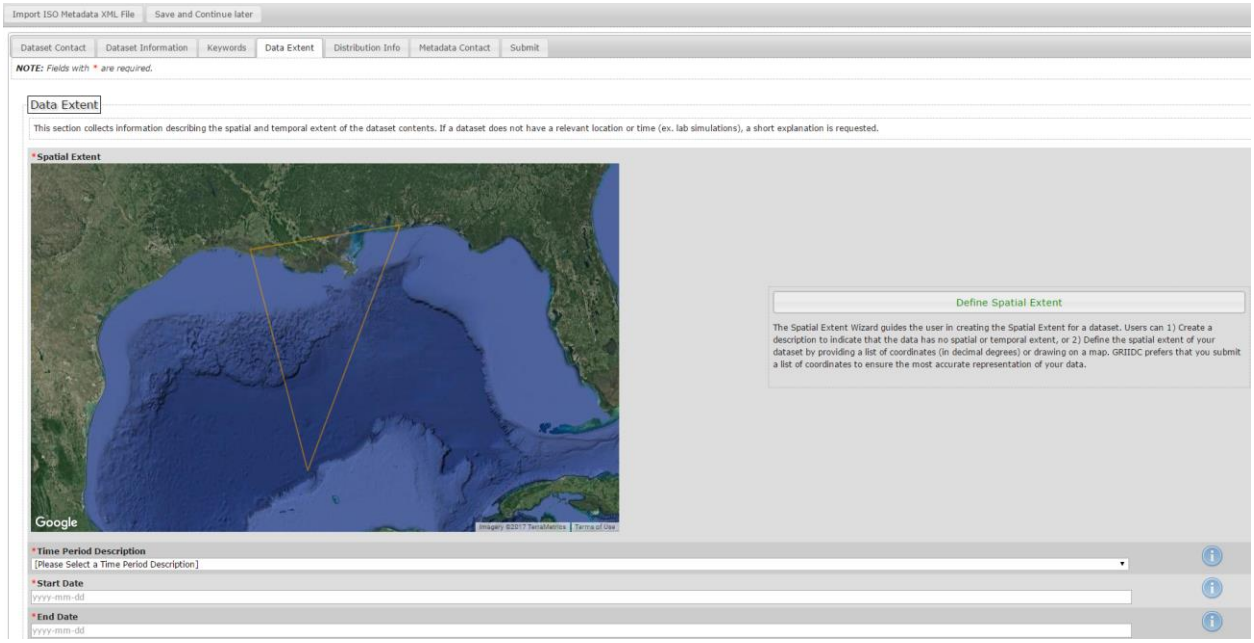

Figure 1: Data Extent tab displaying a spatial extent

Define Spatial Extent * (Required): Please use the GRIIDC Spatial Extent Wizard to provide the most accurate geographic area where the data will be collected or geographic area the data will be generated for or about, for example the area to be included in a model. If your dataset will be generated entirely in the laboratory, and therefore has no relevant geographic information, please describe this using the Spatial Extent Wizard. Please see the Spatial Extent Wizard User Guide for more information on use of the Wizard.

Time Period Description *: (Required) This is a description of what the start and end date you will provide below represents. This field has three standard values:

1. Ground condition: the data represents the actual condition of things on the ground during the time period specified. Samples or data were collected during the time period.
2. Modeled period: the data represents simulated conditions during the time period (could be future modeled period or past modeled period).
3. Ground condition and modeled period: the dataset includes data that represents the actual condition of things on the ground during the time period specified and simulated conditions during the time period.

Start Date* (Required): The beginning date when the data were collected. Alternatively, for a model the earliest date for which the data were generated.

End Date* (Required): The end date when the data were collected. Alternatively, for a model the end date for which the data were generated.

Section 5: Dataset Submission – Dataset File Transfer Details

Once all required fields on the Dataset Submission page are completed with the most accurate and up-to-date information, you can select the Submit tab to submit your dataset file to GRIIDC. You will be unable to select the Submit tab if the Dataset Submission is missing required fields; however, your work will be saved and you can continue your Dataset Submission at a later time if you do not have all of the required information or the dataset file. If your dataset is composed of multiple data files, these files must be packaged together into a single file for submission to GRIIDC (e.g., creating a zip file). Please see GRIIDC Dataset Compression Guidance documents for Linux, Windows, and Mac. Please do not include copyrighted material (e.g., published journal articles) as part of your dataset package.

1) Direct Upload

This is the simplest method for submission of small dataset files to GRIIDC. You may upload a single dataset file of any size using direct upload. Select “Select File” from the “Direct Upload” tab and select the appropriate file from your local computer. If your upload is interrupted, it will attempt to resume automatically. You may also manually pause your upload and continue later. If the window or web browser has been closed before the transfer is complete, you can return to the Dataset Submission page, select the same file, and the transfer will resume where it left off.

2) Request Pull from HTTP/FTP Server

HTTP can be used to point to a dataset that has been submitted to a National Data Archive or to pull a dataset that is available on a public website. Alternatively, your institution may have a public HTTP or FTP site where you can place your dataset file. Contact your institutional IT department to determine if an HTTP or FTP server is available for this purpose. Due to differences in institutional policies and procedures, GRIIDC is unable to assist users in establishing HTTP or FTP servers.

Once your dataset file is available through HTTP or FTP or available through a National Data Archive, visit the GRIIDC Dataset Submission page. Under the Submit tab, select the “Request Pull from HTTP/FTP

Server" tab under "Dataset File Transfer Details" (Figure 2). Please provide the web address (URL) or FTP address in the "Dataset File URL" field.

If you are providing a link to a specific data file, the data file should be available directly at the URL you provide, for example: http://www.nodc.noaa.gov/cgi-bin/OAS/prd/download/9.2.2.tar.gz. A user should not have to search the website you provide in order to find and download the dataset file. Additionally, a user should not have to complete request forms or log in to the website in order to download the dataset file.

If you are providing a link to a National Data Archive, your data should be publicly accessible through the archive. When applicable, the home page or data landing page to your dataset should be provided. By providing a link to a dataset available at a National Data Archive your data will be catalogued in the GRIIDC system.

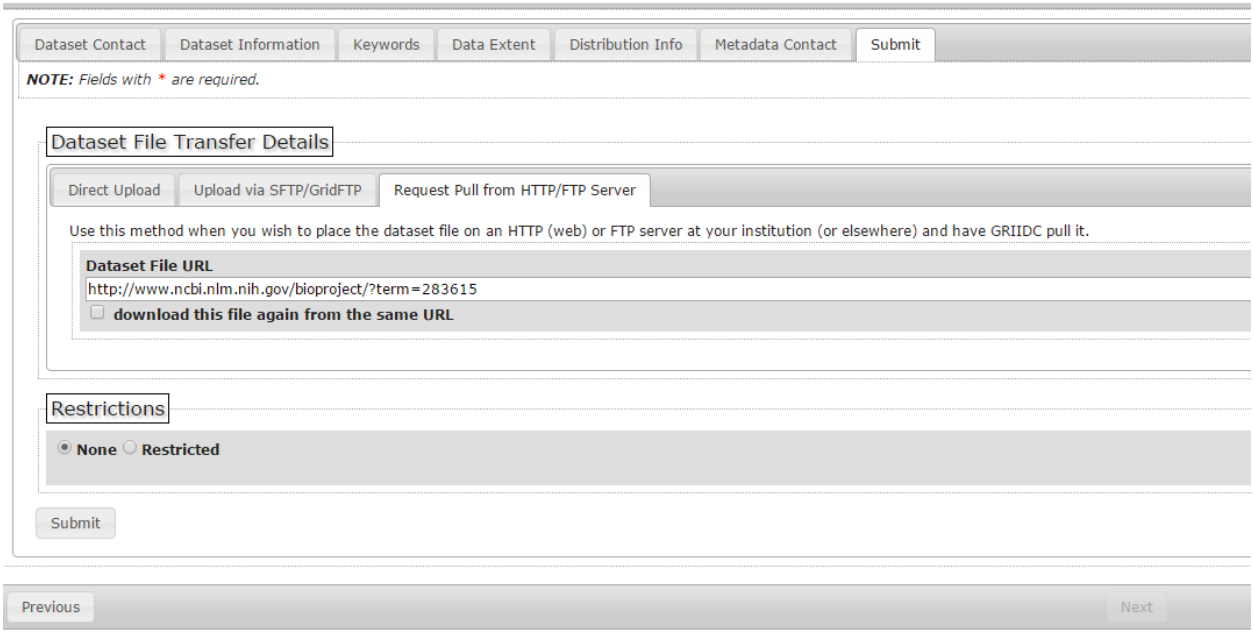

Figure 2: Submit tab: Dataset File Transfer Details, Request Pull from HTTP/FTP Server Tab

3) SFTP

SFTP is a method to submit a dataset file to GRIIDC when it is not possible or practical for you to provide your dataset file by direct upload, website, or FTP server, and if you are able to establish a SFTP client. Due to differences in institutional policies and procedures, GRIIDC is unable to set up SFTP clients for data providers. Please contact your institution's IT department to determine your institutional policies regarding SFTP and to request assistance setting up a SFTP client.

In order to transfer a dataset file via SFTP you must first configure your account to include a GRIIDC POSIX account. To do this visit the Dataset Submission page under the Submit tab in the "Dataset File Transfer Details" section select the "Upload via SFTP/GridFTP" tab and select the button "Request SFTP/GridFTP" (Figure 3). Your account will automatically be configured and a popup will display a confirmation.

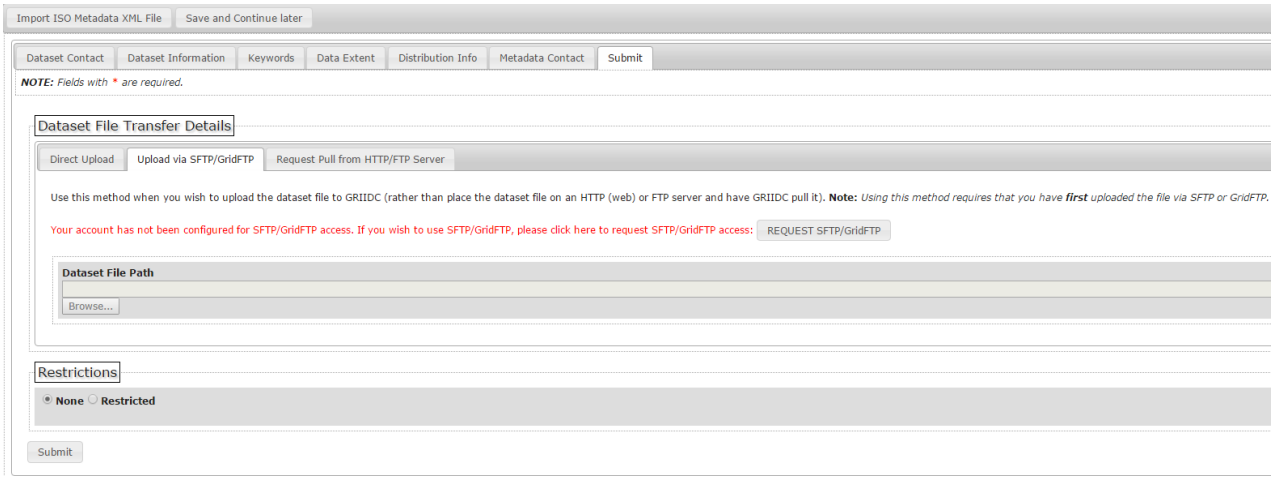

Figure 3: Request SFTP/GridFTP account configuration

Some examples of popular free to use SFTP clients include:

- WinSCP (Windows Secure Copy; http://winscp.net/eng/download.php)
- FileZilla (https://filezilla-project.org/)

To configure your SFTP client to the GRIIDC SFTP client, please use the following parameters:

- Host Name: data.gomri.org
- Port: 22

You will use your GRIIDC username and password to access the GRIIDC client. Upon connecting to GRIIDC SFTP you will be in your home directory (/). This directory is read only. In order to upload a dataset file, please change your directory (cd) to the “incoming” directory. This directory has full write access and is where you can deposit your file.

Once you have transferred your file to the GRIIDC SFTP client, you must submit your dataset. In Dataset Submission, under the Submit tab, select the “Upload via SFTP/GridFTP” tab in “Dataset File Transfer Details” section. Select “Browse” under “Dataset File Path” and choose your dataset file from the incoming directory. Once your file has been selected the window will automatically close (Figure 4).

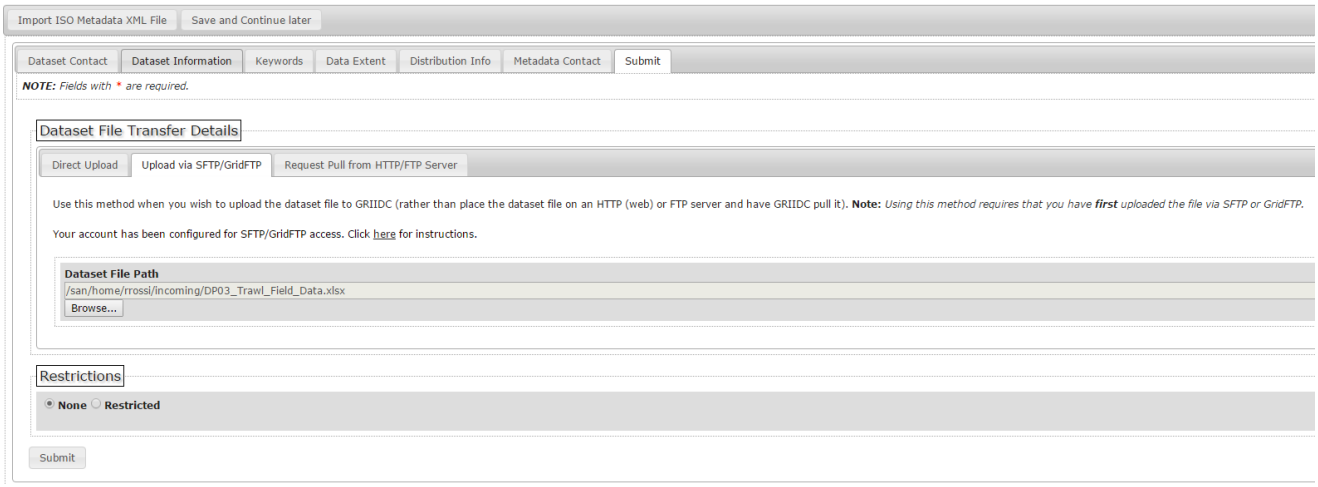

Figure 4: Dataset File Transfer Details, SFTP/GridFTP Tab

Select "Browse" under "Dataset File Path" and choose your dataset file from the incoming directory. Once your file has been selected the window will automatically close.

4) GridFTP

GridFTP is a method to submit data to GRIIDC when it is not possible or practical for you to provide data by another method, if the file size is more than 20 GB, and you are able to establish a GridFTP endpoint.

You will require a Globus account (www.globus.org) to use GridFTP. If you are setting up GridFTP on a personal computer, Globus Personal Connect software can be used to establish a GridFTP endpoint. GridFTP can also be used to conduct server-to-server data transfers, requiring Globus Connect Server Software. Due to differences in institutional policies and procedures, GRIIDC is unable to set up GridFTP endpoints for data providers. Please contact your institution's IT department to determine your institutional policies regarding GridFTP and to request assistance in setting up a GridFTP endpoint.

In order to transfer a dataset file via GridFTP you must first configure your account to include a GRIIDC POSIX account. To do this visit the Dataset Submission page. Under the Submit tab, select the "Upload via SFTP/GridFTP" tab in the "Dataset File Transfer Details" section and select "Request SFTP/GridFTP" (Figure 3).

Your account will automatically be configured and a popup will display a confirmation.

Once you have a GRIIDC POSIX account, a GridFTP endpoint established, and a Globus account you will be able to transfer your dataset file using GridFTP.

To transfer your dataset file log into the globus.org interface and activate your personal endpoint. The GRIIDC endpoint can be found by searching for "GRIIDC Primary" on the Endpoints Tab. Open both your personal endpoint and the GRIIDC Primary endpoint in the Globus File Manager. On the GRIIDC Primary endpoint select the "incoming" folder and queue the transfer of the appropriate file. Once your file is transferred you can submit your dataset. Visit Dataset Submission and under the Submit tab, select the "SFTP/GridFTP" tab of the "Dataset File Transfer Details" section. Select "Browse" below "Dataset File

Path" (Figure 4) to select your dataset file from the appropriate directory. Once you select your dataset file the window will close.

Restrictions

The default selection for restrictions is "none", indicating the dataset file is available for the general public to download after GRIIDC has performed a dataset package review. GRIIDC does allow short term restrictions to be placed on datasets. If you select "Restricted" your dataset file will not be available for download after dataset package review. It is your responsibility to make sure your data are not restricted and publicly available in compliance with the terms of your grant award.

Finish your Submission

After you have uploaded your data either by direct upload, SFTP, or GridFTP or pointed to your data by requesting a pull from an HTTP or FTP server, you can select the Submit button at the bottom of the page to finish the dataset submission process. You will be taken to a confirmation page with your dataset's information and you will also receive a confirmation email that you have submitted a dataset to GRIIDC.

What's Next?

Your submission will be reviewed by the GRIIDC team. The Dataset Package Review process includes:

- Making sure all files can be opened
- Making sure no empty files or folders are included in the dataset
- Making sure the dataset does not contain published materials or other materials that may be subject to copyright
- Making sure the descriptive information fully and accurately describes the dataset files provided

You may be contacted during the Dataset Package Review process to answer questions related to the dataset file or dataset description. Please respond to requests promptly. Datasets are not made publicly available until GRIIDC has reviewed and accepted the submission.

Appendix A: Recommendations: Dataset File Names

When a user downloads your dataset file from the GRIIDC website, the dataset will be provided with the file name with which it was submitted. GRIIDC does not change dataset file names or names of individual files within datasets. Therefore, GRIIDC recommends using the following guidelines to name your dataset file; if your dataset includes multiple files these guidelines should be followed for all files within your dataset.

- Use descriptive names that indicate what the file contains
- Use short names (less than 50 characters)
- Use simple names that are easy to understand
- Use alphanumeric characters
- Use underscores (_) or dashes (-) rather than spaces
- Avoid special characters such as: \ '= /,<>^:;()#*?%,"@!+{}~`[]

- Avoid using internal project codes or acronyms that individuals outside of your laboratory or research group would not understand
- Incorporating the temporal or spatial information when applicable is recommended

You may wish to include the UDI assigned to your dataset in the file name; however, please change the colon (:) to a dash (-). Table 1, below, provides examples of both good and bad dataset file names.

Table 1: Example Dataset File Names

	File Name	**Issues**
Bad example	#1.xls	- Undescriptive - Special Characters
Bad example	gomri data 12.zip	- Undescriptive - Spaces
Bad example	TLF5682.tar	- Undescriptive - Uses internal project code
Bad example	Sam37.pdf	Undescriptive Uses internal project name
Bad example	Dataset publication for the Journal of Examples volume 1 page 32 2012/05/06	Too long Special characters Spaces
Good example	PAHConcentrations_ClamTissue2020.csv	None
Good example	Smith_CommunitySurveyDataset_2020.txt	None
Good example	BacterialProduction_MobileBay_R6.x807.000.0002.zip	None
Good example	Johnson_BubblePlumeVelocity_Lab_Images.tar	None