Python for Data Management

General Information

Core Science Analytics, Synthesis, and Library is planning a technical webinar series to improve data managers' and scientists' skills with using Python to perform basic data management tasks.

Who: These training events are intended for a wide array of users, ranging from those with little or no experience with Python to others who may be familiar with the language but are interested in learning techniques for automating file manipulation, batch generation of metadata, and other data management related tasks.

Where: WebEx (See WebEx Connection Information below).

When: June 11th, June 18th, and June 25th from 3:00 - 4:30pm ET (1:00 - 2:30pm MT). See Schedule below for more information.

Requirements: This series will be taught using Jupyter notebook and the Python bundle that ships with the new USGS Metadata Wizard 2.x tool (download | more information). See Setup below for more information.

Contacts: Drew Ignizio (dignizio@usgs.gov) and Madison Langseth (mlangseth@usgs.gov)

Post-Training Survey: Please take a few minutes to let us know how the training series went by completing this short survey.

Schedule

June 11, 2018: Part 1 - Working with Local Files

- Introduction to the 'os', 'shutil', 'sys' modules
- Creating folders and copying files
- Listing files
- String manipulation and parsing
- List iteration
- Python print statements
- 'Try / catch' code construction
- Strategic organization of data resources

June 18, 2018: Part 2 - Batch Metadata Handling

- Documentation considerations and strategies
- Batch metadata creation
- Batch metadata validation
- Batch metadata updates

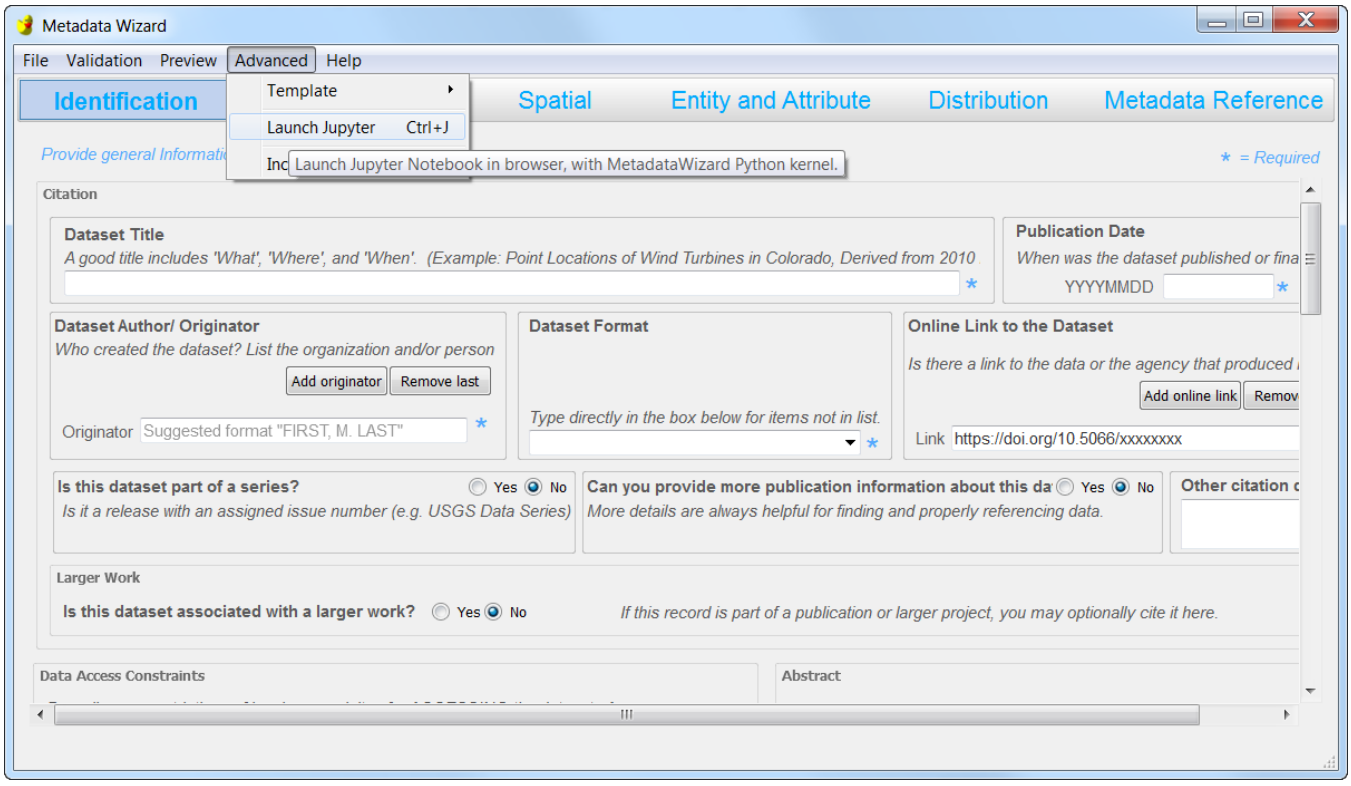
June 25, 2018: Part 3 - Using the USGS ScienceBase Platform with PySB

- ~~Reading files for data quality~~ **
- Understanding the ScienceBase item model
- Bulk upload and item creation in ScienceBase
- Reading from and writing to ScienceBase

*** Due to the interest we've had in learning about using Python with ScienceBase and the time constraints for the training, we've decided to use the time we have for the interactive tutorial to focus solely on using Python with the ScienceBase API. We will consider adding another training to focus more specifically on data review and QA/QC for datasets at a later point point in time, if there is adequate interest. If you are interested in this type of training, please respond to our post-training survey to let us know!*

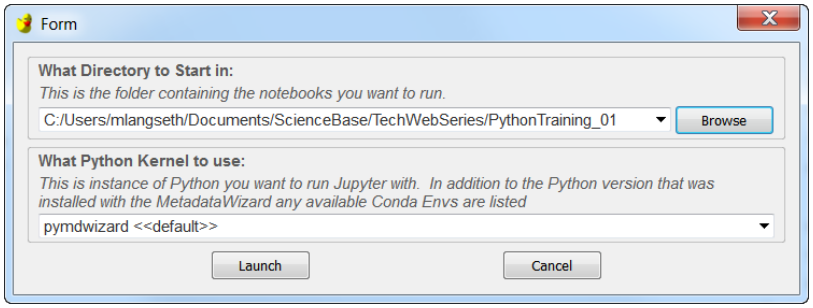
Setup

Prior to the course, you should download and install the Metadata Wizard (version 2.0.4) (download | more information) and ensure that you are able to run the tool properly. If you are on a Mac, you will need to download the MetadataWizard_osx_2.0.3.dmg version (version 2.0.4 is not available yet for Macs). O nce the tool has been downloaded and installed, open the Metadata Wizard and select Advanced Launch Jupyter.

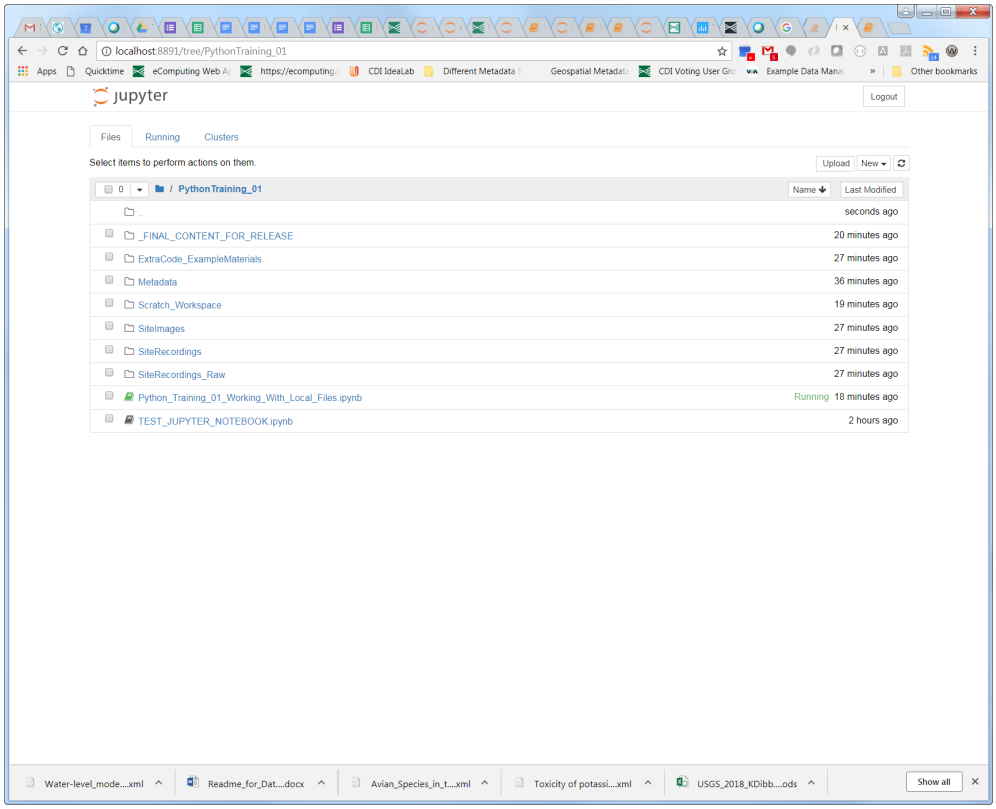


Under **"What Directory to Start in:"** navigate to the location where you saved the Resources Bundle (unzipped) (see Course Resources below). Example: "...\PythonTraining_01"

Under **"What Python Kernel to use:"** select the <<default>> kernel. Then, select **"Launch"** and click **"OK."**

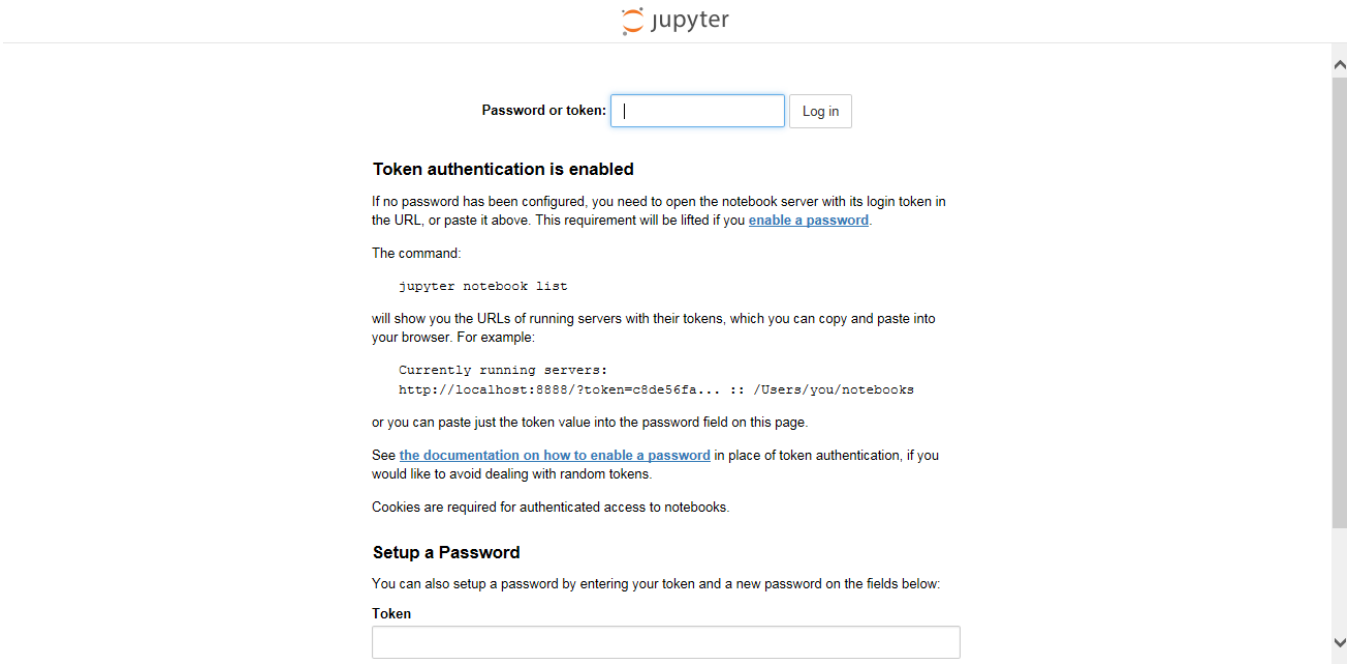


Jupyter will launch in your default web browser. If your default web browser is Internet Explorer, you may be required to enter a password or token. See below for instructions.



Instructions for IE users

Internet Explorer users may be prompted to enter a password or token before seeing the directory structure or before opening the first notebook.



Open the command-line window for the jupyter.exe that is running. You will see a path similar to "http://localhost888/?token=c8de56fa..." Copy the alphanumeric string following "?token=", paste it into the text box in your browser, and click "Log in." (To copy the string in the command-line window, you will need to use an editor. In the Windows .exe window, the editing menu with commands "Mark" and "Copy" can be accessed by right-clicking the top pathname bar.)

Alternatively, you can copy the entire path "http://localhost888/?token=c8de56fa..." into a different browser (e.g. Chrome) or choose a different browser as your default when running notebooks.

Additional Install for Macs

The Mac version of the Metadata Wizard does not currently come with Beautiful Soup, which is a necessary package for some of the Pandas methods that we will be using in Part 3. Follow the instructions below to add Beautiful Soup to your Metadata Wizard libraries:

- Download the 3 Beautiful Soup packages [bs4.zip] [bs4-0.0.1.dist-info.zip] [beautifulsoup4-4.6.0.dist-info.zip]
- Unzip the packages (Please make sure there are NOT two nested folders with the same name. e.g., bs4 > bs4 > [Folder Contents])
- Copy and Paste the bs4, bs4-0.0.1.dist-info, and beautifulsoup4-4.6.0.dist-info folders in your Metadata Wizard site-packages folder. The path for Mac users should be /Applications/MetadataWizard.app/Contents/Frameworks/Python36_64/lib/python3.6/site-packages
- If Jupyter Notebooks was already open on your machine, you will need to restart that instance of Jupyter Notebooks for the additional libraries to be recognized.

For those who have another instance of Anaconda (besides the Metadata Wizard version) on their Macs, the following command should do the same thing when run in terminal:

pip install bs4 --target /Applications/MetadataWizard.app/Contents/Frameworks/Python36_64/lib/python3.6/site-packages

Course Resources

Course resources will be available by the Thursday before the training session, if not sooner. Recordings for each session should be posted by the Tuesday following the session.

Part 1 - Working with Local Files

Download Python Training Module 1 - Course Materials [.zip]

Instructions:

Download the zip file and unzip it on your machine. The zip file includes the Jupyter notebook ("...\PythonTraining_01\Python_Training_01_Working_With_Local_Files.ipynb") and the data that we will be using for Part 1.

We are encouraging all participants to download the Metadata Wizard and use the Anaconda instance that comes with it, even if you already have Anaconda on your machine. Using the Metadata Wizard instance will ensure that you have all of the necessary libraries to complete the modules.

To test that necessary materials are installed and running properly, users should download the “Python Training 01” bundle locally, unzip, and start the Metadata Wizard. Launch Jupyter notebook from the Metadata Wizard (follow instructions on this wiki page) and navigate to unzipped folder of the downloaded training materials. When Jupyter opens in the browser, double click the ‘TEST_JUPYTER_NOTEBOOK.ipynb’ file and follow the instructions to run the test cell.

Recordings:

Part 1 - Working with Local Files Recording Download - Medium Resolution with chat, Q&A, and polls [.mp4]

Part 1 - Working with Local Files Recording Download - High Resolution without chat, Q&A, and polls [.mp4]

Part 2 - Batch Creating & Editing Metadata

Download Python Training Module 2 - Course Materials [.zip]

Download the zip file and unzip it on your machine. The zip file includes the Jupyter notebook ("...\PythonTraining_02\Python_Training_02_Batch_Generating_And_Updating_Metadata.ipynb") and the data that we will be using for Part 2.

Recordings:

Part 2 - Batch Creating & Editing Metadata [.mp4]

Part 2 Exercise Solution:

```
for fl in xmls:
    metd = XMLRecord(fl)
    updated_origin = "Madison L. Langseth"
    metd.metadata.idinfo.citation.citeinfo.origin.text = updated_origin
    metd.save()
    print("Metadata Record Updated: ", fl)
print("Script Complete")
```

Part 3 - Automation with PySB

Download Python Training Module 3 - Course Materials [.zip]

Download the zip file and unzip it on your machine. The zip file includes the Jupyter notebook ("...\PythonTraining_03\Python_Training_03_Automation_With_PySB.ipynb") and the data that we will be using for Part 3.

The training for Monday 6/25 will use ScienceBase (www.sciencebase.gov). Before the training, we recommend that USGS and DOI attendees sign into the web application using their AD credentials to make sure that they are able to access the system. USGS collaborators (partners / co-PIs / etc.) who have a ScienceBase account will use these credentials for access to the system during the training.

Credentials can be checked by logging in here:
https://my.usgs.gov/josso/signon/login.do?josso_back_to=https%3A%2F%2Fwww.sciencebase.gov%2Fcatalog%2Fjosso_security_check

NOTE: Part of this training will focus on using Python to query ScienceBase, read from public items, and download public resources. Attendees will not need an account to perform any of these steps. The second half of the training will be tailored to USGS researchers and USGS collaborators and will focus on writing to the ScienceBase system. Attendees without a ScienceBase account are welcome to follow along for the second half of the training but should note that the programming steps in this part of the training will only be possible with ScienceBase credentials.

Presentation: ScienceBase Overview

Recording:

Part 3 - Automation with PySB [.mp4]

Questions from the training:

How do you return the next page of search results using PySB?

```
items = sb.find_items_by_any_text(username)
while items and 'items' in items:
    for item in items['items']:
        print item['title']
    items = sb.next(items)
```

How do you change the maximum number of search results returned using PySB?

```
response = sb.find_items({
    'q': 'water',
    'offset': 3,
    'max': 3,
    'fields': 'title'
})
print("Found %s items" % response['total'])
response
```

WebEx Connection Information

Topic: Python For Data Management
Host: Event Host
Date: Every Monday, from Monday, June 11, 2018 to Monday, June 25, 2018
Time: 1:00 pm, Mountain Daylight Time (Denver, GMT-06:00)

Session number: 904 176 326
Session password: python

To join the training session
---
1. Go to https://doilearn2.webex.com/doilearn2/k2/j.php?MTID=t063db95f591187a24420f1f9710b9ce1
2. Enter your name and email address.
3. Enter the session password: python.
4. Click "Join Now".

5. Follow the instructions that appear on your screen.
To view in other time zones or languages, please click the link
https://doilearn2.webex.com/doilearn2/k2/j.php?MTID=t6741ed7c74fbb0f5418231213f960ffa

Post-Training Survey

If you attended any of these training events live or watched one or more of the recordings, please take a few minutes to let us know how the training series went by completing this short survey. Your feedback will help us improve future training opportunities.