Downloading files from OPenn

Researchers may wish to download images and metadata for BiblioPhilly manuscripts in advance for study over the holiday break while OPenn is offline (December 26, 2018 - January 1, 2019). Here are instructions for three ways to do so.

1) Downloading files for an individual manuscript via anonymous FTP:

From the main Bibliophilly page **(http://openn.library.upenn.edu/html/bibliophilly_contents.html**), select the “data” link below the manuscript you want to download. The “data” subdirectory will provide you with three options for images -- master (600 dpi tiffs), web (300 dpi jpegs), and thumbnails -- as well as the TEI XML file. Copy the link(s) for the folders and files of interest and use an ftp client such as **FileZilla** to retrieve the data (remembering to remove the http:// protocol!)

2) Downloading files with wget:

This section provides instructions for using wget to download files from OPenn. Wget is a command-line utility available for Linux, Mac OS, and Windows.

Installing wget
First, you’ll need to install wget on your computer.

Mac OS
On a Mac you can install wget directly -- Install and configure wget on OS X -- or if you already have the Homebrew package installer you can use it.

Windows
Download the appropriate setup*.exe file from http://cygwin.com/install.html. Double-click setup*.exe and choose "Install from Internet". Follow the prompts until you are asked to choose a download site for cygwin. Choose any site and continue. Follow the prompts again, until you get to the "Select Packages" page. Click the + next to Web (you may need to scroll down), then click directly on "Skip" and select the first box next to "wget: Utility to retrieve files from the WWW via HTTP and FTP". Click next, accept any dependencies. Download and installation may take a few minutes.

Navigating the command line
Cygwin will install its own folders. Wget will download files into these folders, and you can move the files later.

On a Mac, open your Terminal program. It will probably open in your Documents directory. On Windows, open the Cygwin terminal.

Your command prompt will look something like this, ending with a $:

```
   abc123:Documents user$
```

To move into a different directory, use the cd command:

```
   $ cd openn
```

Your command prompt will reflect your new location:

```
   abc123:openn user$
```

To see all the files and folders available to you, use the ls command:

```
   $ ls
```

To create a new folder, use the mkdir command:

```
   $ mkdir LJSManuscripts
```

More information about these commands and others can be found on this OS X command line cheat sheet.

Now on to wget.

Using wget

The basic wget command will download a single file into the directory you are in. So

```
   $ wget http://openn.library.upenn.edu/
```

will download the index.html page at that address. However, this is probably not what you want. You want to download image and metadata files, either for the entire repository or for specific manuscripts. There are a number of different commands that will allow you to control what exactly gets downloaded, and where those files are placed on your computer.

wget Recipes

Download a single file

I want to download a single image for a specific manuscript:

```
   $ wget http://openn.library.upenn.edu/Data/0001/ljs16/data/web/0284_0000_web.jpg
```

This will bring down only that image that you specify. You can use the same command to download the XML manuscript description:

```
   $ wget http://openn.library.upenn.edu/Data/0001/ljs16/data/ljs16_TEI.xml
```

Download multiple files

You can also use wget to bulk-download files.

I want to download all of the LJS Manuscript data, including master, thumbnail, and web images, and XML manuscript descriptions, in the directory structure used on the OPenn site:

```
   $ wget -np -r http://openn.library.upenn.edu/Data/0001/
```

wget = use the wget program
-np = "no parent", this means do not download any files that are in the folders containing the 0001 folder
-r = "recursive", this means download files directly in the 0001 folder, and also download any files that are in folders inside that folder (without this command, you would only get those files directly inside the 0001 folder)
http://openn.library.upenn.edu/Data/0001/ = start download from this location
I want to download only the XML manuscript descriptions and jpeg files (thumbnails and web images) for a single manuscript. All files are saved in a folder named ljs225

```
   $ wget -nd -np -r -A.jpg -A.xml -P openn/ljs225 \
      http://openn.library.upenn.edu/Data/0001/ljs225/
```

wget = use the wget program
-nd = "no directory", this means do not use the directory structure from OPenn, put all the files into a folder specified by me
-np = "no parent", see above
-r = "recursive", see above
-A.jpg = "accept list", accept only .jpg files
-A.xml = "accept list", accept only .xml files
-P openn/ljs225 = "directory prefix", the folder to which all the files will be downloaded
http://openn.library.upenn.edu/Data/0001/ljs225/ = start download from this location
I want to download all the web JPEGs for all the manuscripts in OPenn to a folder called data/web.

```
   $ wget -nd -np -r -A _web.jpg -P data/web http://openn.library.upenn.edu/Data
```

wget = use the wget program
-nd = "no directory", see above
-np = "no parent", see above

-r = "recursive", see above
-A.xml = "accept list", accept only .xml files
-P openn/msDesc = "directory prefix", see above
http://openn.library.upenn.edu/Data/ = start download from this location
You can combine the different commands to specify exactly what you want to download.

3) Downloading files with rsync:

Rsync is a command-line Remote SYNChronization designed to maintain duplicate copies of data on remote machines. It is also a very powerful tool for the bulk downloading of files. The instructions below show how to install rsync and use it to download files from OPenn.

One advantage rsync has over other tools is that it does, by default, synchronize two directories, usually one a remote server and one on a local computer. This means that rsync can be run multiple times on the same two directories and it will only copy new and changed files from the source to the destination. It can also be set up not just to copy new and changed files, but also to remove files from the destination that are no longer on the target, and thus keep two file systems truly synchronized.

The Linux manual page for rsync is here: http://linux.die.net/man/1/rsync . Note that rsync is different for each operating sytem. For complete rsync documentation for your system view the rsync man page (man rsync).

Rsync commands can be quite complex and tricky to get working just right. There are ample resources on the web for answering particular rsync questions. The samples below show basic usage of rsync for copying data.

Installing rsync
First, you'll need to install wget on your computer.

Mac OS & Linux
Mac OS ships with rsync installed.

If your Linux system does not have rsync installed, you can install with your package management software.

Windows
Download the appropriate setup*.exe file from http://cygwin.com/install.html. Double-click setup*.exe and choose "Install from Internet". Follow the prompts until you are asked to choose a download site for cygwin. Choose any site and continue. Follow the prompts again, until you get to the "Select Packages" page. Click the + next to Net (you may need to scroll down), then click directly on "Skip" and select the first box next to "rysnc". Click next, accept any dependencies. Download and installation may take a few minutes.

Navigating the command line
Cygwin will install its own folders. Wget will download files into these folders, and you can move the files later.

On a Mac, open your Terminal program. It will probably open in your Documents directory. On Windows, open the Cygwin terminal.

Your command prompt will look something like this, ending with a $:

```
abc123:Documents user$
```

To move into a different directory, use the cd command:

```
$ cd openn
```

Your command prompt will reflect your new location:

```
    abc123:openn user$
```

To see all the files and folders available to you, use the ls command:

```
    $ ls
```

To create a new folder, use the mkdir command:

```
    $ mkdir LJSManuscripts
```

More information about these commands and others can be found on this OS X command line cheat sheet.

Now on to rsync.

Using rsync

The basic rsync command, when issued on a site providing anonymous rsync like OPenn will list a directory's contents:

```
    $ rsync rsync://openn.library.upenn.edu/OPenn

    drwxrwxr-x           120 2015/04/29 14:52:07 .
    -rw-rw-r--          1857 2015/04/29 14:53:19 CuratedCollections.html
    -rw-rw-r--         10526 2015/04/29 14:53:19 ReadMe.html
    -rw-rw-r--          2220 2015/05/29 16:34:11 Repositories.html
    -rw-rw-r--         52220 2015/04/29 10:37:08 TechnicalReadMe.html
    drwxrwxr-x            70 2015/04/29 10:36:59 Data
    drwxrwxr-x          4096 2015/04/29 15:13:13 html
```

Adding a subfolder to the above command will give a list of items in that folder:

```
    $ rsync rsync://openn.library.upenn.edu/OPenn/Data/

    drwxrwxr-x            70 2015/04/29 10:36:59 .
    drwxrwxr-x          8192 2015/04/29 10:26:53 0001
    drwxrwxr-x          4096 2015/04/29 10:36:59 0002

    $ rsync rsync://openn.library.upenn.edu/OPenn/Data/0001/

    drwxrwxr-x          8192 2015/04/29 10:26:53 .
    drwxrwxr-x          8192 2015/04/29 10:49:39 html
    drwxrwxr-x            75 2015/02/13 11:26:05 ljs101
```