

Data Archive

(Infrastructure and GUI Access)

Narges Zarrabi

Data Archive - Long-term storage

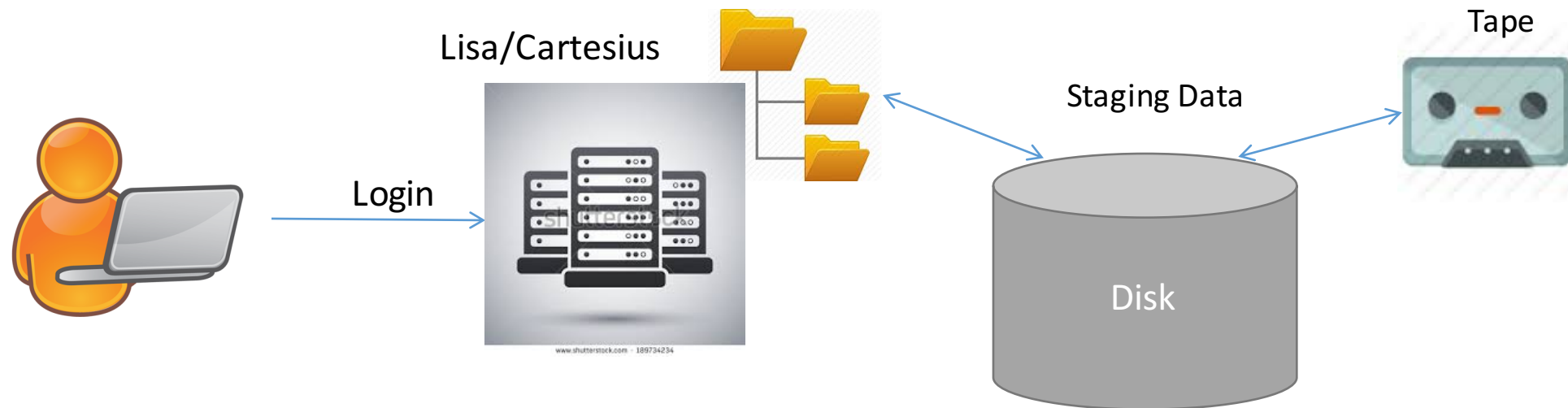
- Long-term storage of data
- Storage medium: Tape → high latency
- Powerful transfer protocols (gridftp, rsync, scp)
- Easy access from HPC services lisa and cartesius via NFS mounts → use archive as yet another directory



Data Archive Infrastructure

Workflow employing Archive from compute clusters at SURFsara:

- User logs in to Lisa/Cartesi
- Archive is mounted via NFS → User sees the archive as another folder
- Copy data to HPC (with which commands?)
- Do your computations on the data
- Copy data back to archive (tape)

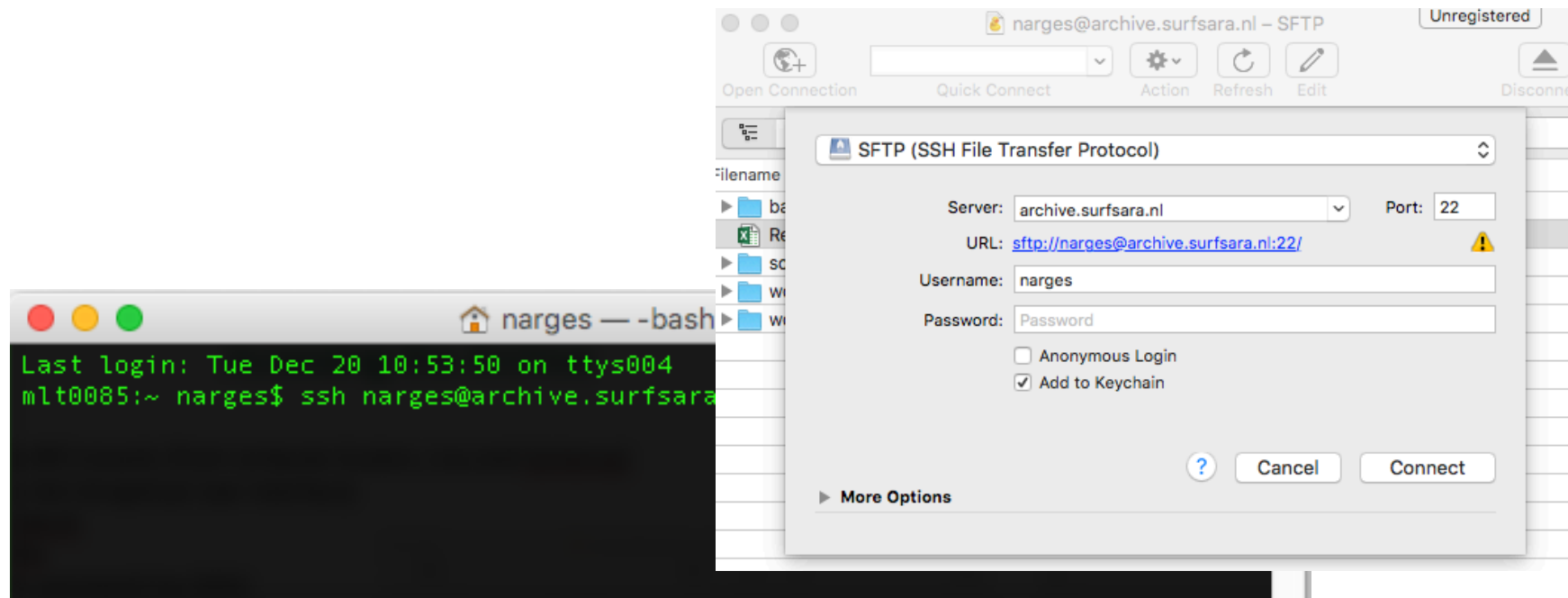


Archive Usage – Best practices

- Try to store files of significant size (> 1 GB) as much as possible. Smaller files will always be accepted, but will lower the performance of restoring your files from tape.
- If you have many small files, make sure to pack them using a file archiving tool like tar or [dmftar](#).
- Try to pack your files before uploading them to the archive, possibly by using dmftar which allows remote tarring.
- Organise your files in such a way that in case the files are needed again only parts of the data set need to be restored from tape.
- Avoid storing unpacked software packages, these usually contain a lot of small files. Instead pack these as well, or refer to a specific software repository.

Accessing the Archive

- Access via GUI (Graphical User Interface)
- Access via command line Direct access
- Access via NFS mounts (from compute clusters, Lisa and Cartesius)



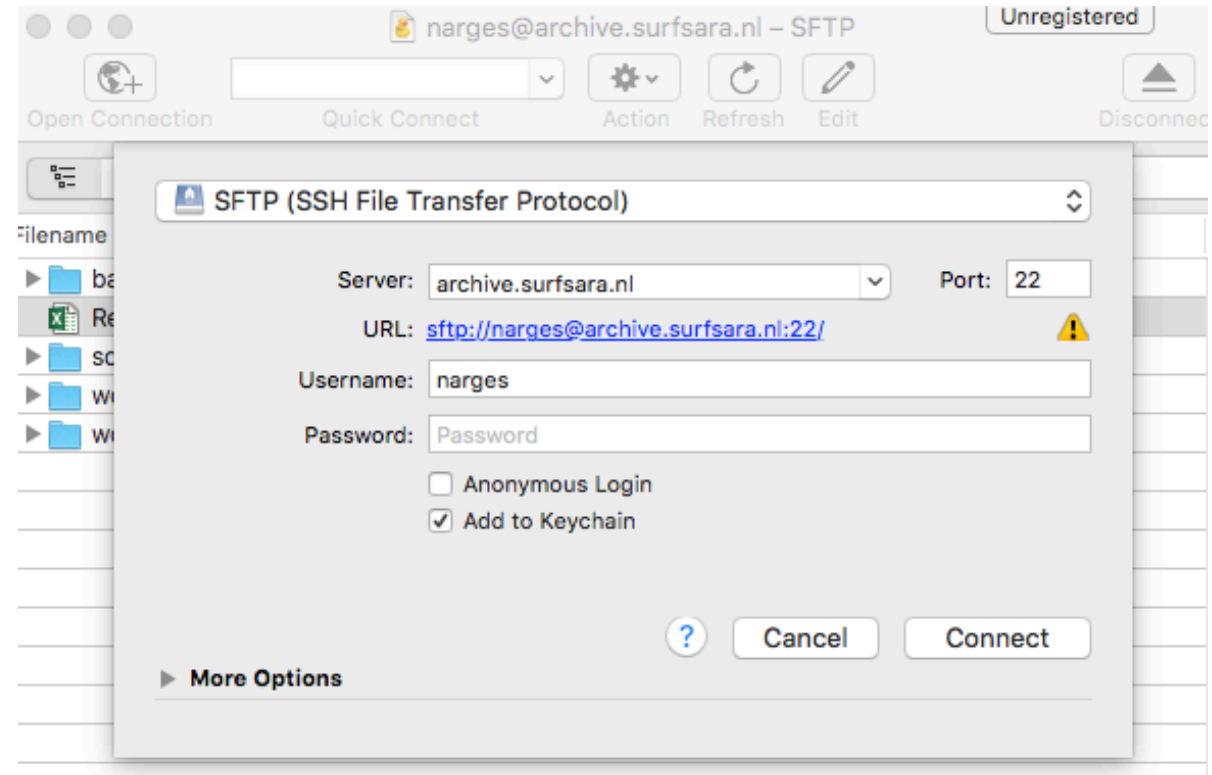
Access Archive via GUI

- Tools to access the Archive via GUI:
 - **Cyberduck** (Mac and Windows) → <http://cyberduck.ch/>
 - **MobaXterm** (Windows) → <http://mobaxterm.mobatek.net/>
 - **Filezilla** (Linux) → <https://filezilla-project.org/>



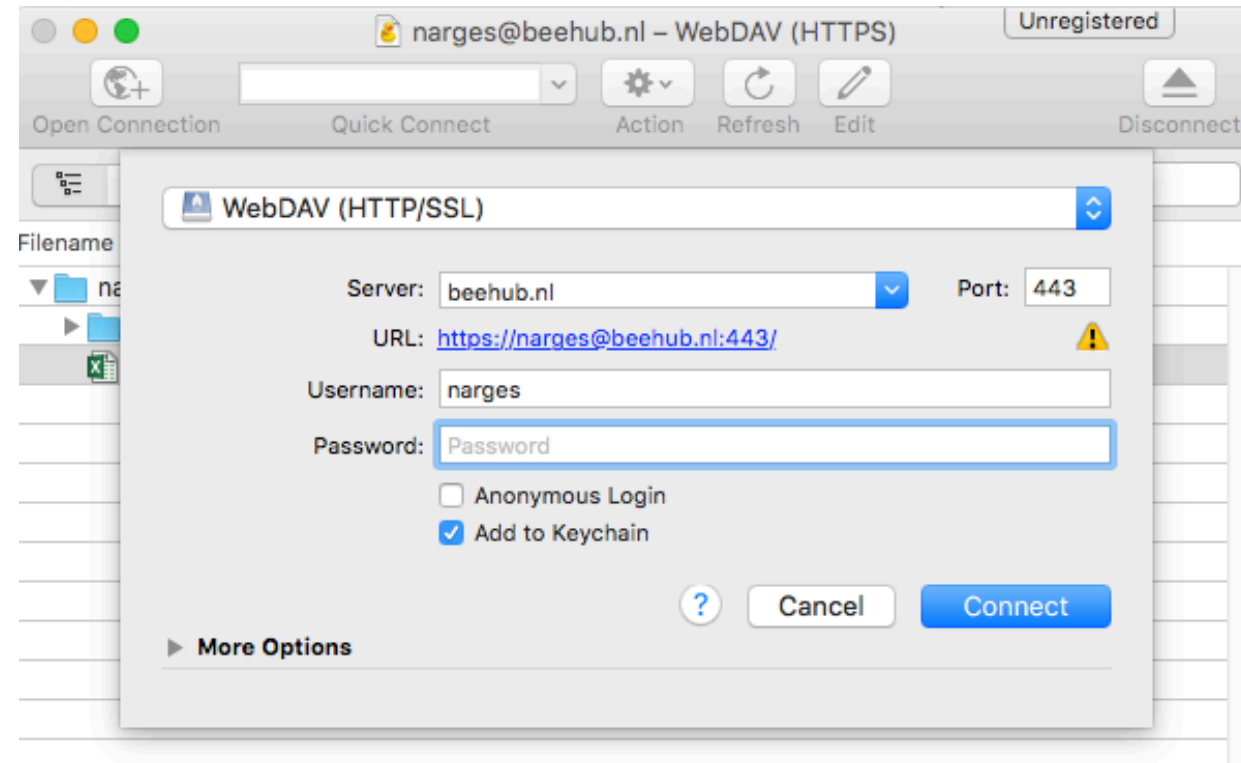
Access Archive via Cyberduck

- Cyberduck is a standalone client that runs on Windows and Mac OSX
 - Download and install: <http://cyberduck.ch/>
- To start an Archive session with Cyberduck:
 - Start Cyberduck
 - Click on 'Open connection'
 - You now see this screen
 - Choose the following options:
 - Connection type: SFTP (SSH File Transfer Protocol)
 - Server: archive.surfsara.nl
 - port: 22
 - Login with your credentials (sdemo<xxx>)



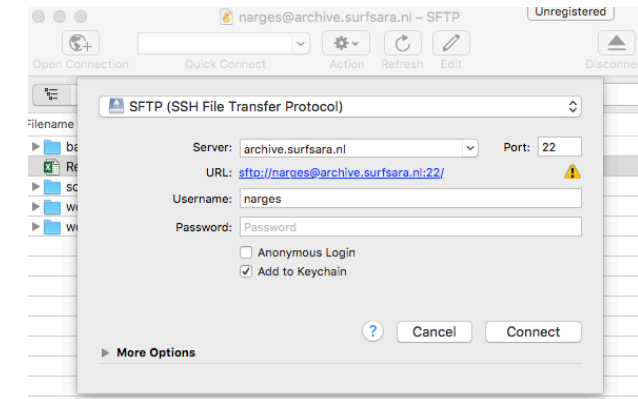
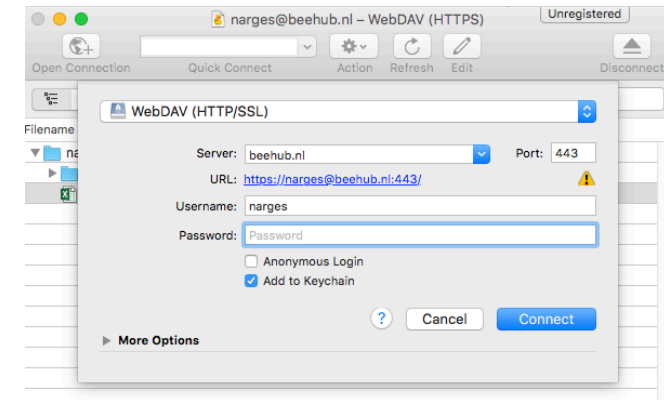
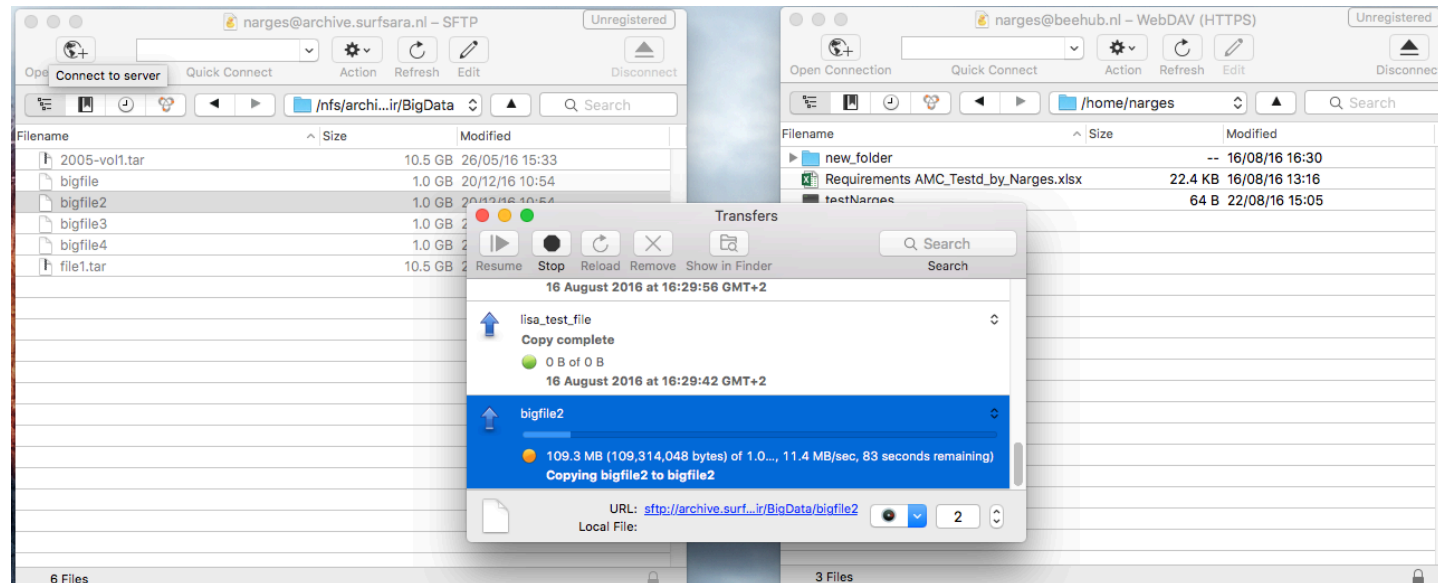
Access BeeHub via Cyberduck

- To start a BeeHub session with Cyberduck:
 - Start Cyberduck
 - Click on 'Open connection'
 - You now see this screen
 - Choose the following options:
 - Connection type: WebDAV (HTTP/SSL)
 - Server: beehub.nl
 - port: 443
 - Enter your BeeHub username and password as you use them on the website (not your sdemo credentials!)



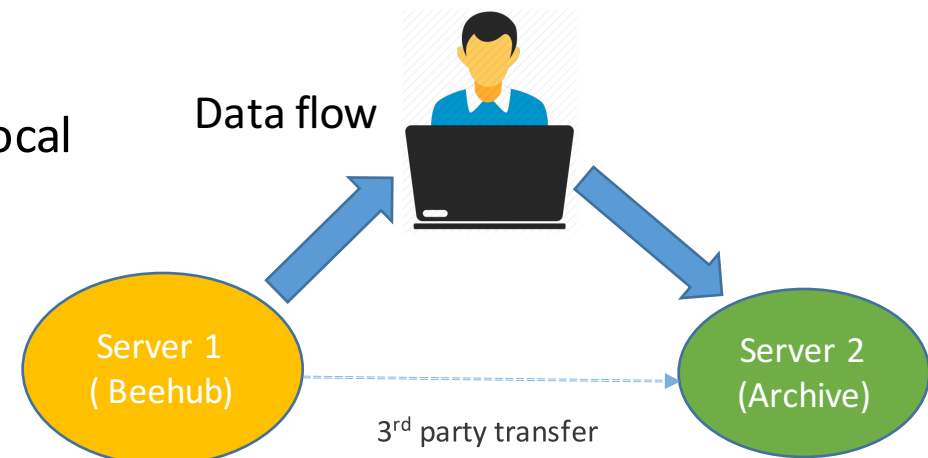
Transfer Data using Cyberduck

- To transfer data between services using Cyberduck:
 - Start Cyberduck
 - Establish a connection to the Archive
 - Establish another connection to BeeHub
 - Simply drag and drop files to transfer data



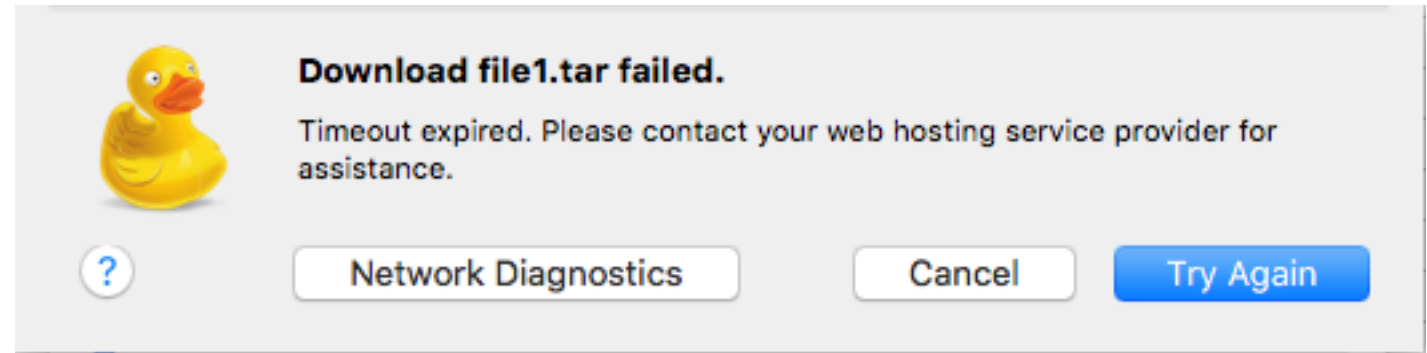
Advantages & Limitations

- Advantages:
 - Easy data transfer (to the archive)
 - Good for dumping data to the archive, and not fetching data
 - Transfer data between services (Only possible for small data)
 - Can be accessed from Windows, Mac and Linux machines
- Drawbacks
 - Does not support 3rd party transfer and the data flows via the user laptop. Therefore the transfer depends on your local storage and connectivity (If the connection is lost, the transfer is lost).
 - Only for small data files
 - Does not always work for fetching data (data needs to be staged first)
 - You can't see the status of the data (i.e. whether the data is on disk or on tape).

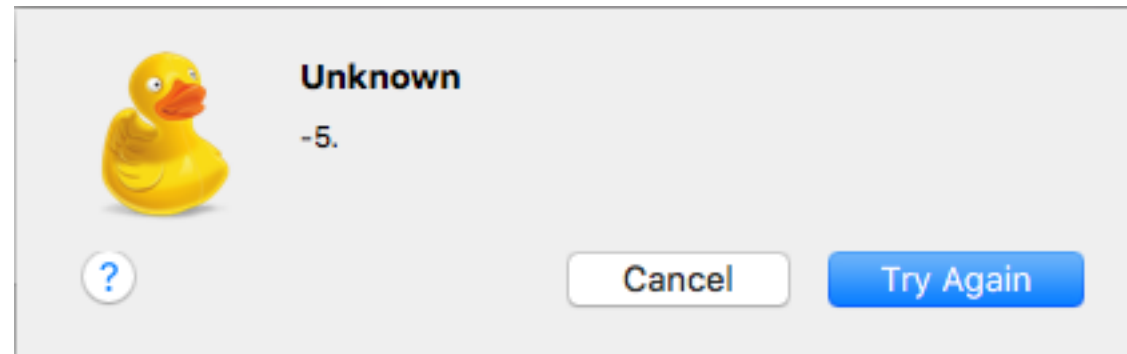


Transfer Data using Cyberduck

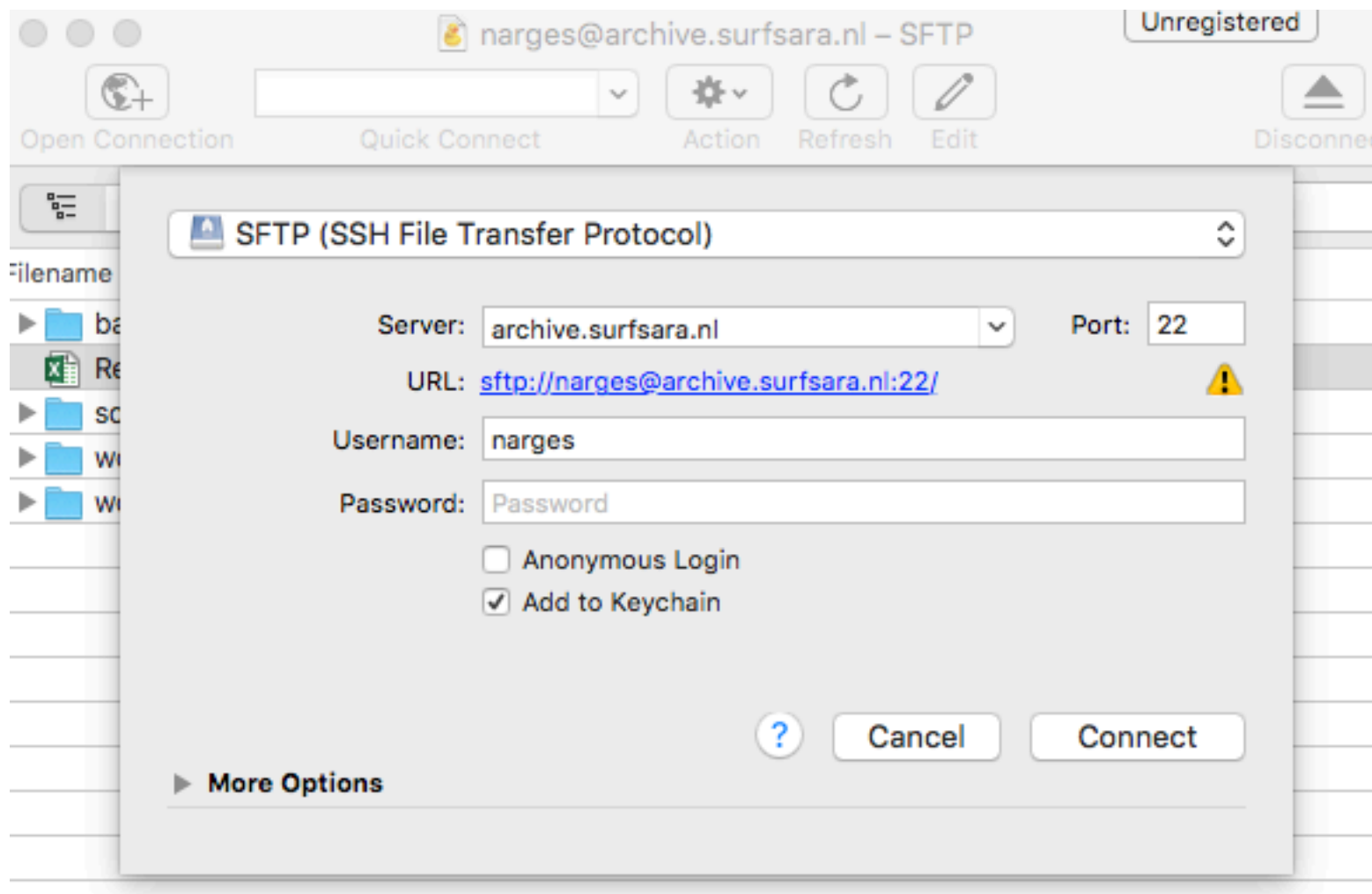
- Error: If the file is on tape, and not on disk. The files needs to be stages first.



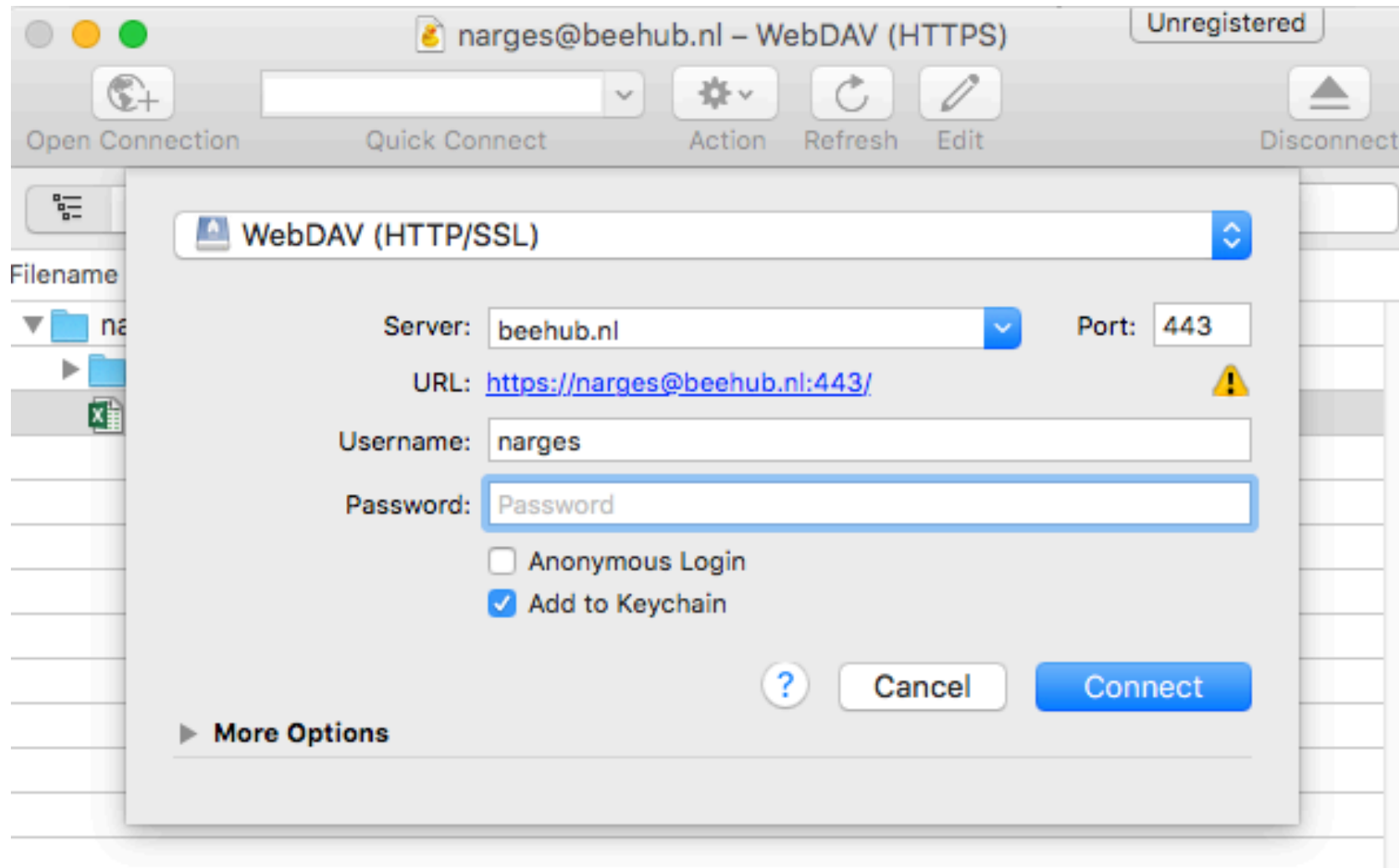
- Error: If the internet connection is lost.



Access Archive via Cyberduck



Access BeeHub via Cyberduck



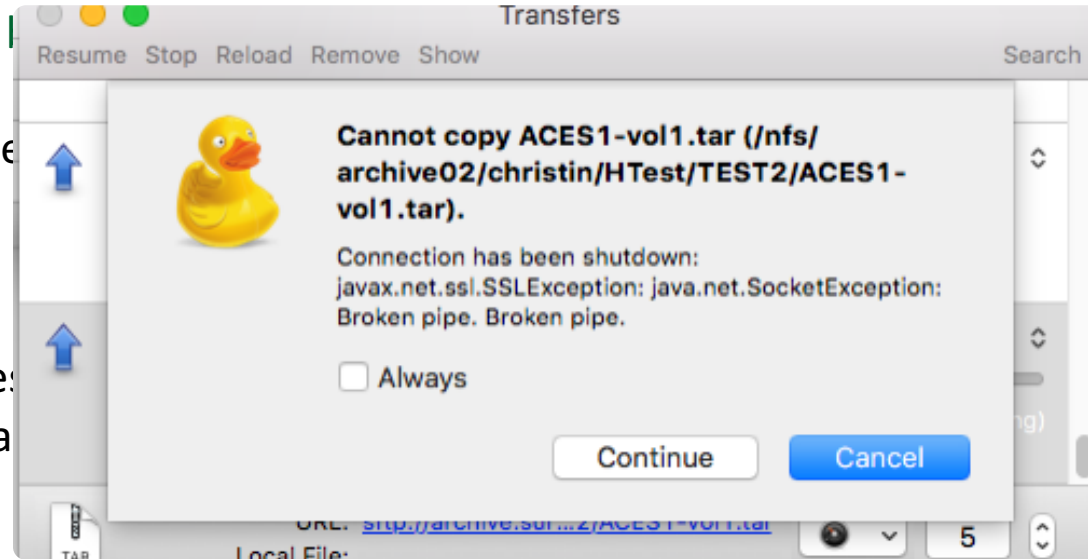
Data copied to the archive will remain on disk for a period of 4 weeks if there is enough disk space. After that the data will be removed from Disk.

Before using data on Tape, it need to first l

To directly transfer files from disk to a tape commands:

```
dmput -r -w <filename> [..]
```

The following example command migrates occupied by filename, and will wait until a



```
narges$ dmls -l
total 12288000
```

-rw-r--r--	1	narges	narges	10485760000	2016-05-26 15:33 (DUL)	2005-vol1.tar
-rw-r--r--	1	narges	narges	1048575001	2016-12-20 10:54 (DUL)	bigfile
-rw-r--r--	1	narges	narges	1048575001	2016-12-20 10:54 (DUL)	bigfile2
-rw-r--r--	1	narges	narges	1048575001	2016-12-20 10:54 (OFL)	bigfile3
-rw-r--r--	1	narges	narges	1048575001	2016-12-20 10:55 (OFL)	bigfile4
-rw-r--r--	1	narges	narges	10485760000	2016-12-20 11:09 (OFL)	file1.tar