

Long Term Archive Howto

This is a short manual on how to search for and retrieve data from the Long Term Archive.

For background information and in case of problems, please refer to the [Frequently Asked Questions](#) page

User Access

Public data

The LTA catalogue can be searched using the anonymous account (AWWORLD). This account gives direct access to the project “All public data”, and search queries will return results of the entire catalogue because metadata are public for all LTA content.

Staging and subsequent downloading of public data always requires an account with LTA user access. If you do not have an account yet, you can register with “[Create account](#)”. LTA access should be automatically added, but if you encounter problems contact Science Support.

Project data

To directly access project-related data in the LTA you need to have an account in [MoM](#) that is enabled for the archive.

1. This automatically happens if you were a member of the original project proposal in Northstar/MoM.
2. Otherwise Science Support can add you to the project for which you need access upon request.

If you were not originally a member of the project in MoM and Science Support adds you to it, you might get an email asking you to set a new password in [ASTRON Web Applications Password Self Service](#). Please note that this will set a new password not just for the LTA *but for MoM (LOFAR/WSRT) and Northstar as well*.

How to find data in the archive

Once your account is set up, or as anonymous user (AWWORLD) you can navigate the catalogue.

Page navigation

The LTA menu, as shown below, gives access to the main functionality.



A “project” can be selected by registered users that have access to a specific project. Subsequent searches will then be done within that project.

Searching the catalogue is easiest by selecting one of the options in the “Search” menu. Depending on the knowledge of the data you are looking for (if any), searches can be performed on e.g., target (“Simple Search”), date and/or observation-ID and more advanced parameters per data type (“Advanced Search”).

Another option is to choose “Show Latest”, which will show the latest additions in the LTA. This is particularly useful for project related searches.

Finding Data



Depending on the search parameters, e.g., which data products were requested (observation, pipeline), lists of observations and/or pipelines will be returned (see observations example above). From this point there are several options:

1. select observations/pipelines and stage (prepare for download) all data related to the selection
2. select observations and “show pipelines” related to the observations, then select pipelines and stage
3. select observations/pipelines, “show dataproducts” related to the selection, possibly filter the data products (to have smaller selections) and then select and stage the data products

Note that observations often have no raw data in the archive, but the metadata is visible because subsequent pipelines have processed the raw data further. To get to the pipelines related to observations, use “Show Pipelines”.

To see whether observations or pipelines have data products in the LTA, look for the “Number of Correlated/BeamFormed DataProducts” column. These columns, as also a few others, can also be used to navigate to the relevant dataproducts.

Once you have a list of dataproducts on your screen, the “Release Date” will tell you when the data are available for public download. If data is public, or you are a member of the project you are looking in, the “checkbox” column will be selectable and staging can proceed. You can also hover with your mouse over the checkbox and get more information, like the size, location and checksums.

[There is a separate page with more detailed information and advanced tricks to help find and download your data](#)

Unspecified Data/Process

Some data has had problems somewhere in the automation and control part of the LOFAR software during observation or processing. Sometimes a few subbands might be affected, sometimes an entire observation. Science support will check the data, (re)run things manually or fix things if needed and then archive the data. This does mean that the automation and control sometimes loses track of the files and the archiving process has no information beyond the Observation ID and filename itself. In

such cases a few subbands or an entire observation might end up under “Unspecified Process”. We do attempt to fix things at a later date, but that's not always feasible. If the files were archived the data itself is usable. It is the information the LTA needs to properly label and query the data is missing.

If an Observation is missing, or is missing subbands, please check if it ended up under Unspecified.

Staging data (Prepare for download)

Once you have a list of dataproducts, observations or pipelines, you can use the check boxes to select which files you want to download. The first check box can be used to select or deselect all files or observations on a page.

edit columns **stage** uris

#	<input type="checkbox"/>	DataProduct Identifier	Target Name	Right Ascension [degree]
1	<input checked="" type="checkbox"/>	3400453	3C218	139.52354
2	<input checked="" type="checkbox"/>	3400408	Hercules-A	252.78416
3	<input checked="" type="checkbox"/>	3400429	3C218	139.52354
4	<input checked="" type="checkbox"/>	3400445	3C218	139.52354
5	<input type="checkbox"/>	3400450	3C218	139.52354
6	<input checked="" type="checkbox"/>	3400462	3C218	139.52354
7	<input type="checkbox"/>	3400442	3C218	139.52354
8	<input checked="" type="checkbox"/>	3400463	3C218	139.52354
9	<input checked="" type="checkbox"/>	3400465	3C218	139.52354
10	<input type="checkbox"/>	3400469	3C218	139.52354
11	<input type="checkbox"/>	3400464	3C218	139.52354

The LOFAR Archive stores data on magnetic tape. This means, that it cannot be downloaded right away, but has to be copied from tape to disk first. This process is called 'staging'.

When you have made your selection of files, you click on *stage*. This shows you the following message. It means that a request has been sent to the LTA staging service to start retrieving the requested files from tape storage and make them available. You will get an e-mail when this tape retrieval is complete.

The following file(s) are requested for download. You will receive an email when the files can be retrieved.

Size	Filename
758.2 MB	L75432_SAP004_SB215_uv.MS_f6b663ca.tar

The e-mail that you get when the tape retrieval is complete gives you a list of files and has two attachments, `html.txt` and `srn.txt`:

Your data retrieval request with id 25 has been staged and is ready for retrieval.

List of files:




srm://lofar-srm.fz-juelich.de:8443/pnfs/fz-juelich.de/data/lofar/ops/projects/commissionin_04d3730c.tar

The attached files can be used to retrieve the staged files.

For more information visit http://www.lofar.org/wiki/doku.php?id=public:lta_howto

This mail has been automatically generated by the ASTRON/LOFAR LTA staging service.

Do not reply to this message. If you have any questions or remarks, please contact sciencesupport@astron.nl and provide the id of the request in your message.

Name	Size	Type
 Message	4KB	Message Attachment
 html.txt	214 Bytes	File Attachment
 srm.txt	189 Bytes	File Attachment

There are two ways you can use this list to retrieve the files: [http](#) and [srm](#)

Please take note of the following

1. Unless you have an extremely fast connection (10 Gbit/s or more), **it is in general advisable to stage no more than 10 TB at a time** (see also point 4). At maximum efficiency a 1 Gbit/s connection will already take 24 hours to retrieve 10 TB of data, in practice it will often take quite a bit more.
2. On a 1 Gbit/s connection as a general rule of thumb, you should be able to retrieve data at about 100-500 GB/hour, especially if you try to retrieve 4-8 files concurrently. If you see speeds much lower than this, you might have some kind of network problem and should in general contact your IT staff.
3. Staging the data from tape to disk might take quite a bit of time. In the large data centres that the LTA uses, the tape drives are shared with all users and requests are queued. This is not just users of LOFAR but large data other projects like the LHC. This might mean that it takes anywhere from a few hours to a day or more to stage a copy of your data from tape to disk.
4. The amount of space available for staging data is limited although quite large. This space is however shared between all LOFAR LTA users. This includes LTA operations for buffering data from CEP to the LTA before it gets moved to tape. If many users are staging data at the same time, and/or LOFAR operations is transferring large amounts of data, the system might temporarily run low on disk space. You might then get a message that your request was only partially successful. In general the request will still finish 1-2 days later and we do monitor if requests don't get stuck and restart if needed.
5. We strive to keep a copy of data that was staged on disk for 1-2 weeks so you have some time to download it. After that it might get removed to make space for more recent requests. The the copy of the data on tape is only read and will still be available if you need to access the data again at a later stage but you might need to stage a copy to disk again.
6. We are continuously trying to improve the reliability and speed of the available services. Please contact Science Support if you have any problems or suggestions for improvement.
7. The data centres the LTA uses also have maintenance or small outages sometimes. Science Support can advice you if this is the case and when it is planned to end, if you are having trouble accessing data. In general this will not be at the same dates as the LOFAR stop days.

Download data

You can download your requested data with the files from your e-mail notification. There are different possibilities and tools to do this. If you're unsure, which one to use, please refer to the according [FAQ Answer](#).

HTTP download

If you open `html.txt` this file contains a list of http links that you can feed to a unix commandline tool like `wget` or `curl` or even use in a browser.

For `wget` you can use the following command line:

```
wget -i html.txt
```

This will download the files in `html.txt` to the current directory (option `-i` reads the urls from the specified file).

Preferrably, especially when downloading large files, you should also use option `-c`. This will continue unfinished earlier downloads instead of starting a fresh download of the whole file. (Make sure to first delete existing files that contain error messages instead of data, if you use this option):

```
wget -ci html.txt
```

Do not set the username and password on the `wget` command line because this allows other users on the system to view them in the process list. Instead you should create a file `~/.wgetrc` with two lines according to the following example:

```
user=lofaruser  
password=secret
```

Note: This is only an example, you have to edit the file and enter your own personal user name and password!

Set access authorizations of the `.wgetrc` file to user only so that the credentials are not exposed to anybody else, e.g.:

```
chmod 600 .wgetrc
```

There is no easy way to have `wget` rename the files as part of the command directly. It does not accept the `-O` flag inside a file it gets with `-i`. You can either rename files afterward, e.g. using the following command:

```
find . -name "SRMFifoGet*" | awk -F %2F '{system("mv "$0" "$NF)}'
```

or add the `-O` option to each line in `html.txt` but then feed each line to `wget` separately like this: `cat html.txt | xargs wget`. By default the `html.txt` file does not contain such options.

The following Python script will take care of renaming and untarring the downloaded files:

```
#M.C. Toribio
#toribio@astron.nl
#
#Script to untar data retrieved from the LTA by using wget
#It will DELETE the .tar file after extracting it.
#
#Notes:
#When using wget, the files are named, as an example:
#SRMFifoGet.py?surl=srm:%2F%2Fsrm.grid.sara.nl:8443%2Fpnfs%2Fgrid.sara.nl%2F
data%2Flofar%2Fops%2Fprojects%2Flofarschool%2F246403%2FL246403_SAP000_SB000_
uv.MS_7d4aa18f.tar
# This scripts will rename those files as the string after the last '%'
# If you want to change that behaviour, modify line
# outname=filename.split("%")[-1]
#
# Version:
# 2014/11/12: M.C. Toribio

import os
import glob

for filename in glob.glob("*SB*.tar*"):
    outname=filename.split("%")[-1]
    os.rename(filename, outname)
    os.system('tar -xvf '+outname)
    os.system('rm -r '+outname )

    print outname+' untarred.'
```

Note that wget does not overwrite existing files. If you use the continue option ('-c') it will append any missing parts to the existing file. If you don't use the continue option and there is a file present (e.g. from a stopped earlier download), wget creates a new file by appending a number (e.g., '.1') to the filename.

There are some small example links if you browse to <https://lofar-download.grid.sara.nl/> where you can test with for example the file1M (which is 1 MB) if your setup is correct.

SRM download

If you open the file `srm.txt` this file contains a list of srm locations which you would feed to `srmcp`. SRM is a GRID specific protocol that is currently supported for data at the SARA and Jülich locations. It is faster, especially if you have significantly more than 1 Gbit/s bandwidth. It requires a valid [GRID certificate](#) and installation of the [GRID srm software](#). NB There is an [alternative installation that does not require root privileges](#). Contact Science Support if you think you might need a GRID account but it can not be provided by your own institute. An example command line would be:

```
srmcp -server_mode=passive -copyjobfile=srm.txt
```

to retrieve all requested files contained in srm.txt or e.g.

```
srmcp -server_mode=passive srm://lofar-srm.juelich.de:8443/pnfs/fz-jeulich.de/data/lofar/ops/projects/commissioning2012/file.tar
file:///data/files/file.tar
```

to retrieve a single file. You need `--server_mode=passive` if you are behind a firewall or on an internal network. Omitting this option may result in improved transfer speed as it will attempt to use multiple streams when retrieving a file. An alternative strategy to improve the overall transfer speed is to run multiple srmcp requests in parallel, e.g. by splitting the provided srm.txt file and feeding the partial lists to separate srmcp commands.

If you do experience insufficient transfer speeds with srmcp, you may want to look into using srmcp with a [globus-url-copy](#) copy script.

Troubleshooting

- There is a [LTA FAQ page](#) that should help with the common difficulties.

From:

<https://www.astron.nl/lofarwiki/> - **LOFAR Wiki**

Permanent link:

https://www.astron.nl/lofarwiki/doku.php?id=public:lta_howto&rev=1425241650

Last update: **2015-03-01 20:27**

