Advanced ways to find and retrieve data in the LTA

There are some useful ways to find and retrieve your data in the LTA that might not be immediately obvious. This page explains some of the more advanced options you have.

Queries

You can use colons in numeric queries, to select ranges. This will for example give all
observations and pipelines that have a SAS/Observation ID in the range from 432000 to
432190:



In textual entries, wildcards can be used.



• You can put a list of SAS/Observation IDs in the query:

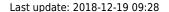


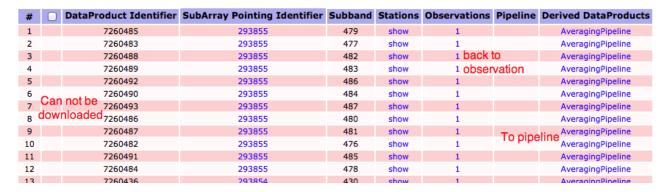
Viewing data

When you are looking at the results of a query you might see something like this:



This means that the observation is known in the LTA, it knows what data was produced, the produced data was not archived, but further processing happened on the raw data and the results of some of those pipelines were archived. If you click on the zero, you will see something like this:





This allows you to navigate from a pipeline back to the original observation, or from the observation to any pipelines that have run on the raw data.

Retrieving data

 You can retrieve data on the Observation and Pipeline level, you don't have to select all files individually.

#		Observation Id	Observing Mode	Antenna Set	Instrun Filte
1	⋖	146448	Interferometer	HBA Dual Inner	110-190
2		146447	Interferometer	HBA Dual Inner	110-190
3	⋖	146446	Interferometer	HBA Dual Inner	110-190
4		146445	Interferometer	HBA Dual Inner	110-190
5	⋖	146444	Interferometer	HBA Dual Inner	110-190
6	\checkmark	146443	Interferometer	HBA Dual Inner	110-190
7		146442	Interferometer	HBA Dual Inner	110-190
8	\checkmark	146441	Interferometer	HBA Dual Inner	110-190
9	⋖	146456	Interferometer	HBA Dual Inner	110-190
10	\checkmark	146455	Interferometer	HBA Dual Inner	110-190
11		146454	Interferometer	HBA Dual Inner	110-190
12		146453	Interferometer	HBA Dual Inner	110-190
13		146452	Interferometer	HBA Dual Inner	110-190
	_				

• If you have a query with more than 1000 results, you can open the multiple pages each in a separate tab/window.

Observation 1001 to 1100 (showing 100 of total 1156) -

r Of SubArray Start Time Duration Nr Stations Nr Stations

With the small triangle next to a list, you can fold or unfold the list to get a better overview.

Folded entries

Observation 1 to 100 (showing 100 of total 1156)

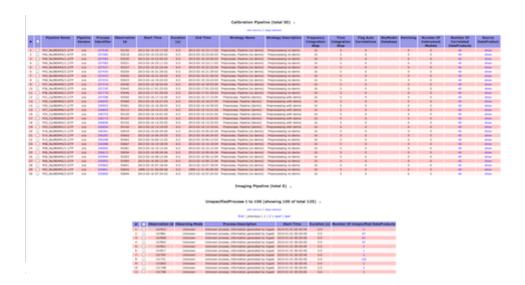
Averaging Pipeline 1 to 100 (showing 100 of total 4060)

Calibration Pipeline (total 30)

Imaging Pipeline (total 0)

UnspecifiedProcess 1 to 100 (showing 100 of total 125)

Unfolded entries



DBView

There is a server that gives the option to run your own queries on the database http://lofar-dbview.target.rug.nl/

A useful query might be this one, that gives you all files for a certain Obs Id (SAS VIC tree ID).

```
SELECT fo.URI, dp."dataProductType", dp."dataProductIdentifier",
    dp."processIdentifier"
FROM AWOPER."DataProduct+" dp,
        AWOPER.FileObject fo,
        AWOPER."Process+" pr
WHERE dp."processIdentifier" = pr."processIdentifier"
    AND pr."observationId" = '123456'
    AND fo.data_object = dp."object_id"
```

```
AND dp."isValid" > 0
```

In this '123456' should be replaced with the Obs Id of an Observation/Pipeline you're looking for. Pipelines also have an "observationId" == the SAS Id, even though that's a but confusing. To be able to run this query, you have to go to the link above, login as the right user, select the right project, and then put this query into the "Manual SQL".

Example You can also modify these queries. for example if you want to also know the MD5 checksum, you can run:

```
SELECT fo.URI, fo.hash_md5, dp."dataProductType",
dp."dataProductIdentifier",
dp."processIdentifier"
FROM AWOPER."DataProduct+" dp,
         AWOPER.FileObject fo,
         AWOPER."Process+" pr
WHERE dp."processIdentifier" = pr."processIdentifier"
   AND pr."observationId" = '123456'
AND fo.data_object = dp."object_id"
AND dp."isValid" > 0
```

AstroWise Python Interface

There is a Python client library for accessing the LTA. With this library, you can script your own queries. The installation description can be found here: LTA Client installation. Be sure to have the latest version installed. Note that since January 2018 this library uses python3, python2 is no longer supported.

Once you have installed the client, set up your user name and password. These are the same as for MoM. Remember that this is just a different interface to the LTA catalogue: you will need the same credentials as for the web interface.

After installing the LTA client, the file .awe/Environment.cfg will appear in your home directory (if not, then create one). Make sure the file at least contains the following lines:

```
[global]
database_user : <your username>
database_password : <your password>
```

The following script can be used to test your installation:

```
# Python3 code
from pprint import pprint
from awlofar.main.aweimports import Observation, Pointing, SubArrayPointing
from common.database.Context import context
result = {}
for project in sorted(context.get_projects()) :
    print("Project %(project)s" % vars())
    ok = context.set_project(project)
```

```
# do your query
   obs ids = set()
    query = (Pointing.rightAscension > 95) & \
            (Pointing.rightAscension < 105) & \
            (Pointing.declination > 20)
                                            & \
            (Pointing.declination < 30)
    print("Total Pointings %d" % len(query))
    for pointing in query :
        print("Pointing found RA %f DEC %f" % (pointing.rightAscension,
pointing.declination))
        query subarr = SubArrayPointing.pointing == pointing
        for subarr in query subarr:
            query obs = Observation.subArrayPointings.contains(subarr)
            for obs in query obs :
                obs ids.add(obs.observationId)
    result[project] = sorted(list(obs ids))
    print(result[project])
pprint(result)
```

It should print out a list of pointings (note that in this example the library was installed in \$HOME/tmp):

```
$ env PYTHONPATH=$HOME/tmp/lib/python3.5/site-packages python3 lta_test.py
Project ALL
Total Pointings 202
Pointing found RA 95.003499 DEC 24.838742
Pointing found RA 95.174754 DEC 28.660087
Pointing found RA 95.220000 DEC 29.140000
Pointing found RA 95.546250 DEC 23.331750
Pointing found RA 95.561458 DEC 24.584056
..etc..
```

You may need to kill the script, because it will print out all the observations in a certain patch of the sky archived in the LTA.

In case of errors, there may be the need to open some port on the firewall at your institution. Specifically, port 1521 should be open. Also make sure that the LTA client library can be found in your PYTHONPATH (see LTA Client installation for more details). In case of trouble, get in contact with Science Operations and Support.

Examples

Once you have tested that your connection to the catalogue is working, you are ready to browse the archive and stage the data you need. Here we will list a few examples of python scripts that can be used to access the LTA. All of them will need to import some modules:

```
from datetime import datetime
from awlofar.database.Context import context
```

```
from awlofar.main.aweimports import CorrelatedDataProduct, \
   FileObject, \
   Observation
from awlofar.toolbox.LtaStager import LtaStager, LtaStagerError
```

The lines above must be added to each of the scripts below for these to work.

This simple script will allow you to find all data within a single project, for example LC2_035. Please change the project name to the code of a project of yours. If you also want to stage the data you found, just set the do_stage variable to True. Be careful with how many files you stage and what size they have: the same limits as for the web interface apply here.

```
# Should the found files be staged ?
do stage = False
# The project to query, LC2_035 has public data
project = 'LC2 035'
# The class of data to query
cls = CorrelatedDataProduct
# Query for private data of the project, you must be member of the project
private data = False
# To see private data of this project, you must be member of this project
if private data :
    context.set project(project)
    if project != context.get current project().name:
        raise Exception("You are not member of project %s" % project)
query_observations = Observation.select_all().project_only(project)
uris = set() # All URIS to stage
for observation in query observations :
    print("Querying ObservationID %s" % observation.observationId)
    # Instead of querying on the Observations of the DataProduct, all
DataProducts could have been queried
    dataproduct query = cls.observations.contains(observation)
    # isValid = 1 means there should be an associated URI
    dataproduct query &= cls.isValid == 1
    for dataproduct in dataproduct query :
        # This DataProduct should have an associated URL
        fileobject = ((FileObject.data object == dataproduct) &
(FileObject.isValid > 0)).max('creation date')
        if fileobject :
            print("URI found %s" % fileobject.URI)
            uris.add(fileobject.URI)
        else :
            print("No URI found for %s with dataProductIdentifier %d" %
(dataproduct. class . name , dataproduct.dataProductIdentifier))
print("Total URI's found %d" % len(uris))
if do stage :
    stager = LtaStager()
```

```
stager.stage_uris(uris)
```

The following script will find subbands 301 and 302 for all targets within two different projects.

Pay attention to the difference between the keys subband and stationSubband; the former is a sequential number assigned to each subband in an observation, while the latter is linked to the frequency at which the observation was performed. Example: an observation was set up covering the range 30-77.3 MHz with two simultaneous beams using 244 subbands each. In this case, subband will range from 0 to 487, while stationSubband from 153 to 396. The stationSubband information is stored in the observation, but not in the pipeline products (which instead contain the frequency). If you want to search on stationSubband, you must perform your search on observations first, then fetch the pipelines linked to those observations. If you use frequency, you can search directly on pipelines.

As a general advise, before performing a search, you need to **understand thoroughly the meaning of the keywords that you are using and where their values are stored**, otherwise you may not find the data you are looking for.

```
do stage = False
project1 = 'LC2 016'
project2 = 'LC2 012'
subband1 = 301
subband2 = 302
cls = CorrelatedDataProduct
# Query for private data of the project, you must be member of the project
private data = False
# All URIS to stage
uris = {
   project1: set(),
   project2: set(),
for project in (project1, project2) :
   print("Using project %s" % project)
    if private data :
        context.set_project(project)
        if project != context.get current project().name:
            raise Exception("You are not member of project %s" % project)
   query observations = Observation.select all().project only(project)
    for observation in query observations :
        print("Querying ObservationID %s" % observation.observationId)
        dataproduct_query = cls.observations.contains(observation)
        # isValid = 1 means there should be an associated URI
        dataproduct query &= cls.isValid == 1
       dataproduct_query &= ((cls.subband == subband1) | (cls.subband ==
subband2))
       # Or for stationSubband do :
       #dataproduct query &= ((cls.stationSubband == subband1) |
(cls.stationSubband == subband2))
        for dataproduct in dataproduct query :
            # This DataProduct should have an associated URL
```

Here, we find data between freg1 and freg2 taken within one project between day1 and day2

```
do stage = False
project = 'LC2 033'
freq1 = 172.0
freg2 = 178.0
day1 = datetime(2014,8,26) # this could include time; ie hours, minutes,
secondes
day2 = datetime(2014,8,29) # idem
# DataProduct class to query; CorrelatedDataProduct, SkyImageDataProduct,
etc ...
cls = CorrelatedDataProduct
# Query for private data of the project, you must be member of the project
private data = False
# To see private data of this project, you must be member of this project
if private data :
    context.set project(project)
   if project != context.get current project().name:
        raise Exception("You are not member of project %s" % project)
query observations = (
    (Observation.startTime >= day1) &
    (Observation.endTime < day2) ).project only(project)
uris = set()
for observation in query observations :
   print("Querying ObservationID %s" % observation.observationId)
   dataproduct query = cls.observations.contains(observation)
   # isValid = 1 means there should be an associated URI
   dataproduct query &= cls.isValid == 1
   dataproduct query &= cls.minimumFrequency >= freq1
    dataproduct query &= cls.maximumFrequency < freq2</pre>
```

```
for dataproduct in dataproduct_query :
    # This DataProduct should have an associated URL
    fileobject = ((FileObject.data_object == dataproduct) &
(FileObject.isValid > 0)).max('creation_date')
    if fileobject :
        print("URI found %s" % fileobject.URI)
        uris.add(fileobject.URI)
    else :
        print("No URI found for %s with dataProductIdentifier %d" %
(dataproduct.__class__.__name__, dataproduct.dataProductIdentifier))

print("Total URI's found %d" % len(uris))

if do_stage :
    stager = LtaStager()
    stager.stage_uris(uris)
```

Example; query public data

Querying public data in projects you are not member of. First set project ALL, then construct a query and optionally limit the query to a certain project :

```
context.set_project('ALL')
query = CorrelatedDataProduct.select_all()
query &= query.project_only('LCO_017')
print(len(query))
# 1800
```

Python Module for Staging

The python interaction with the LTA catalog can be complemented with the use of a specific module developed to give users more control over their staging requests.

The module is made available **here** and its functions are self-explanatory.

For a description of what the user can do, we list here the functions that are available.

stage(surls)

It takes in a list of surls, queues a staging request for those urls, and outputs the ID of the request.

get status(stageid)

It tells the user if a request is gueued, in progress or finished (success).

abort(stageid)

It allows users to end a staging request.

get surls online(stageid)

It gives a list of the surls that have been staged for the relative request. The list is updated whenever a new surl comes on line.

get srm token(stageid)

The srm token is useful to interact directly with the SRM site through GRID/SRM tools.

reschedule(stageid)

If a request failed, it can be rescheduled.

get progress()

No input needed. It returns the statuses of all the requests owned by the user.

Below is an example of how to use this:

```
> python
Python 2.7.10 (default, Oct 23 2015, 19:19:21)
[GCC 4.2.1 Compatible Apple LLVM 7.0.0 (clang-700.0.59.5)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import stager access as sa
2016-11-24 16:39:55.865000 stager_access: Parsing user credentials from
/Users/renting/.stagingrc
2016-11-24 16:39:55.865111 stager access: Creating proxy
>>> sa.prettyprint(sa.get progress())
+ 12227
  - File count
                        100
  - Files done
                        40
                 ->

    Flagged abort

                     ->
                            false
  - Location
                        fz-juelich
               ->
  - Percent done
                            40
                     ->
  - Status
                    on hold
  - User id
                        1919
                 ->
```

From:

https://www.astron.nl/lofarwiki/ - LOFAR Wiki

Permanent link:

https://www.astron.nl/lofarwiki/doku.php?id=public:lta tricks&rev=154521172



