ASTRON

Netherlands Institute for Radio Astronomy

Data retrieval from the Long Term Archive

L. Cerrigone

LOFAR Data School 2016

LOFAR data flow

International LOFAR Telescope (ILT)

Data from every station is transferred to COBALT and written on CEP2/4

Dutch stations

Onsala

Norderstedt

Chilbolton

LOFAR Core (NL)

Potsdam

Bałdy

Borówiec

Jülich

Effelsberg

Tautenburg

Łazy

Unterweilenbach

Nançay

Netherlands Institute for Radio Astronomy

The LOFAR Long Term Archive

What is the LTA?

A collection of data centres offering computing and storage facilities to many big scientific projects, among which is LOFAR

The LOFAR Long Term Archive

INGEST

Data and metadata flow from the LOFAR cluster to the LTA sites at 1.5 GB/s

Size of stored data growing at ~6 PB/yr

Whenever a data set needs to be archived, a series of checks is performed. For example, if the same data file has already been archived or if the checksum of the file in the archive matches that of the file in the LOFAR cluster.

Data files missing metadata are archived as *Unspecified products.* These cannot be searched for unless the proper ID is known.

The LOFAR Long Term Archive

The data reaching the LTA is written on disks, but its final storage is on tapes. Tapes are cheap and reliable devices for data storage.

LTA tapes are not reachable through the internet.

Whenever the data is needed by a user, the files on tape must first be copied to disks reachable through the net.

The LOFAR Long Term Archive

STAGING

The tape containing the data is reached by a robotic arm, then its content is read and copied to disk

Tape space optimized

↓

Fragmentation

↓

Overheads

The LOFAR Long Term Archive

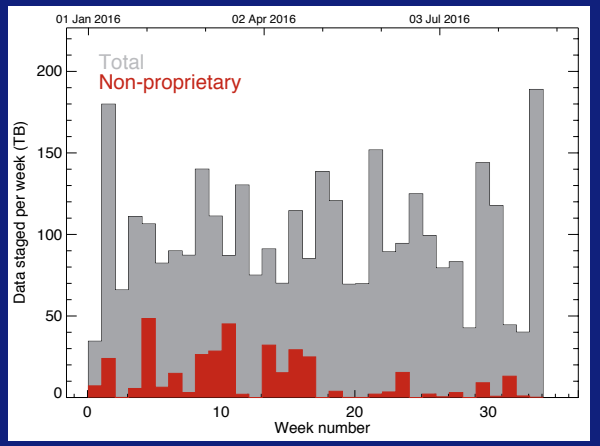
STAGING Limitations

Each staging request cannot contain more than a few thousand files and it cannot exceed 5 TB

What if I need to stage 20k files?

Split them into 10 consecutive requests for 2k each

Consecutive: a new request is issued after the previous one has completed (i.e., all files staged)



Accessing the archive

The necessary information can be found in the *LTA How To* at

http://www.lofar.org/operations/doku.php?id=public:lta_howto

The web interface to the archive is located at http://lofar.target.rug.nl/

The data in the archive have a *proprietary* status for 1 year: only the members of the project are allowed to download them

All the metadata is public as soon as the data are archived

Staging is possible only if you are a registered user: you need a MoM account with *LTA user* privileges. Whenever you create a MoM account, send Science Support a request for *Scientist* and *LTA user* privileges.

The web interface

The web interface is the main location for data browsing.

Home Help Login Project Search Show Latest
AWWORLD All public data

Lofar Long Term Archive

Welcome to the Lofar Long Term Archive (LTA) web service.

On the top are links to: help pages, user login, project overview, seach form and most recently added data.

Before you can query and stage proprietary data make sure that :

1. you are logged in; see **username** below login link
2. selected the correct project; see the **project** name below project link

---

From March 1 2015 onwards, cycle data which have passed the proprietary period will be publicly available. All metadata in the Archive can be queried anonymously at anytime, but downloading public data can only be done by registered users (follow the "Create account" link). Non-public data can only be downloaded by project members.

Note: for the first cycles of LOFAR operations, part of the data were ingested in the archive without metadata. Theses data cannot be found using the standard search parameters, except for the appropriate Observation ID. When unspecified data are present in a project, this is listed in the "Unspecified" column on the Projects page. Data can still be requested using the "Project" pulldown in the various Search options.

A list of all LOFAR observing cycles and approved projects can be found here.

For more information on this web service see the Lofar wiki.

This system was developed as part of the Target project. Target was supported by Samenwerkingsverband Noord-Nederland (SNN) and the Groningen Municipality. The project was financially supported by the European Fund for Regional Development and the Dutch Ministry of Economic Affairs, Agriculture and Innovation (Pieken in de Delta), the Province of Groningen and the Province of Drenthe.

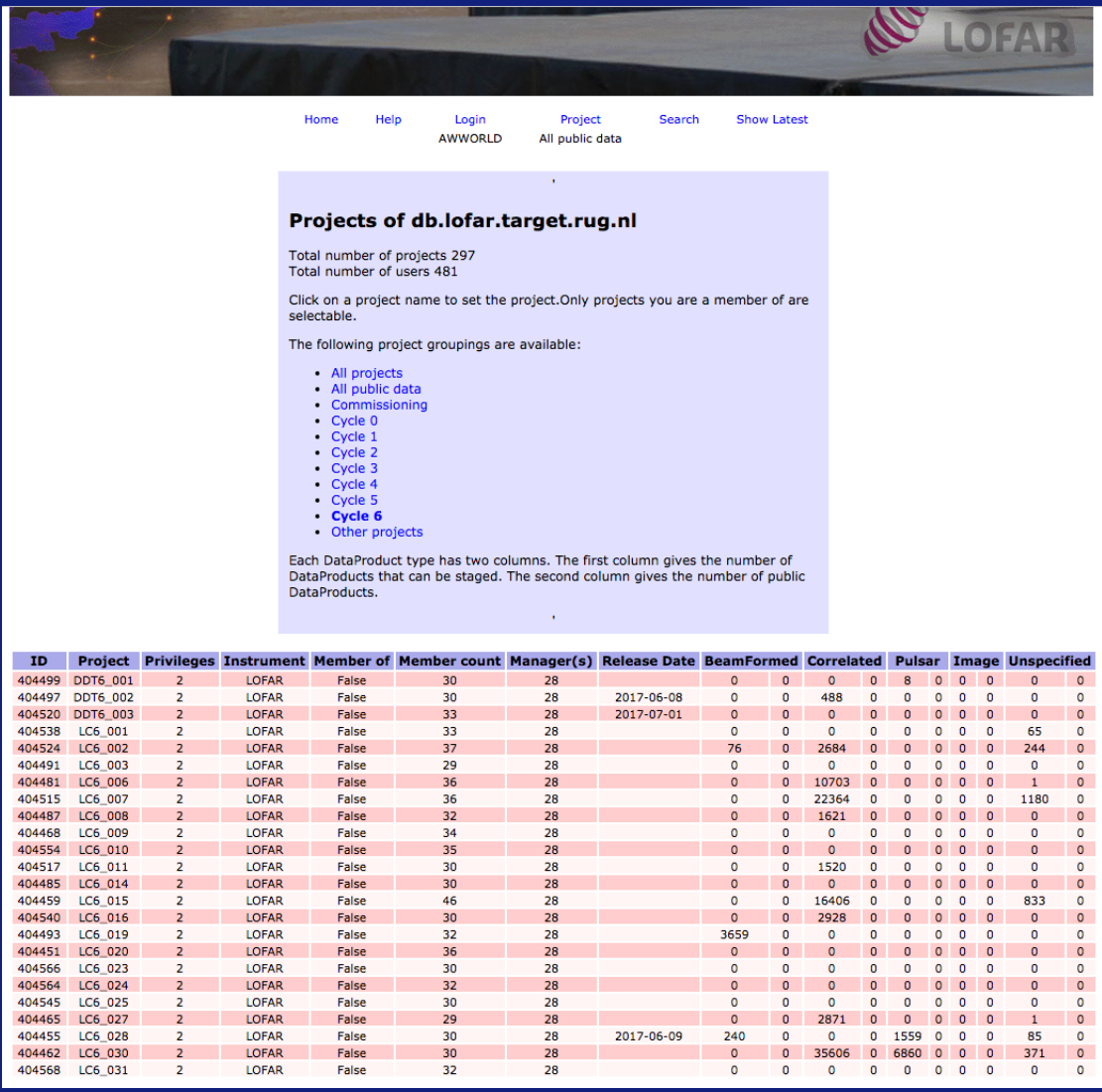
powered by

ASTRO WISE

The web interface

You can only see the list of projects, if you don’t login.

The account you need is the same you use with MoM.



Home Help Login Project Search Show Latest
AWWORLD All public data

Projects of db.lofar.target.rug.nl

Total number of projects 297
Total number of users 481

Click on a project name to set the project.Only projects you are a member of are selectable.

The following project groupings are available:

- All projects
- All public data
- Commissioning
- Cycle 0
- Cycle 1
- Cycle 2
- Cycle 3
- Cycle 4
- Cycle 5
- **Cycle 6**
- Other projects

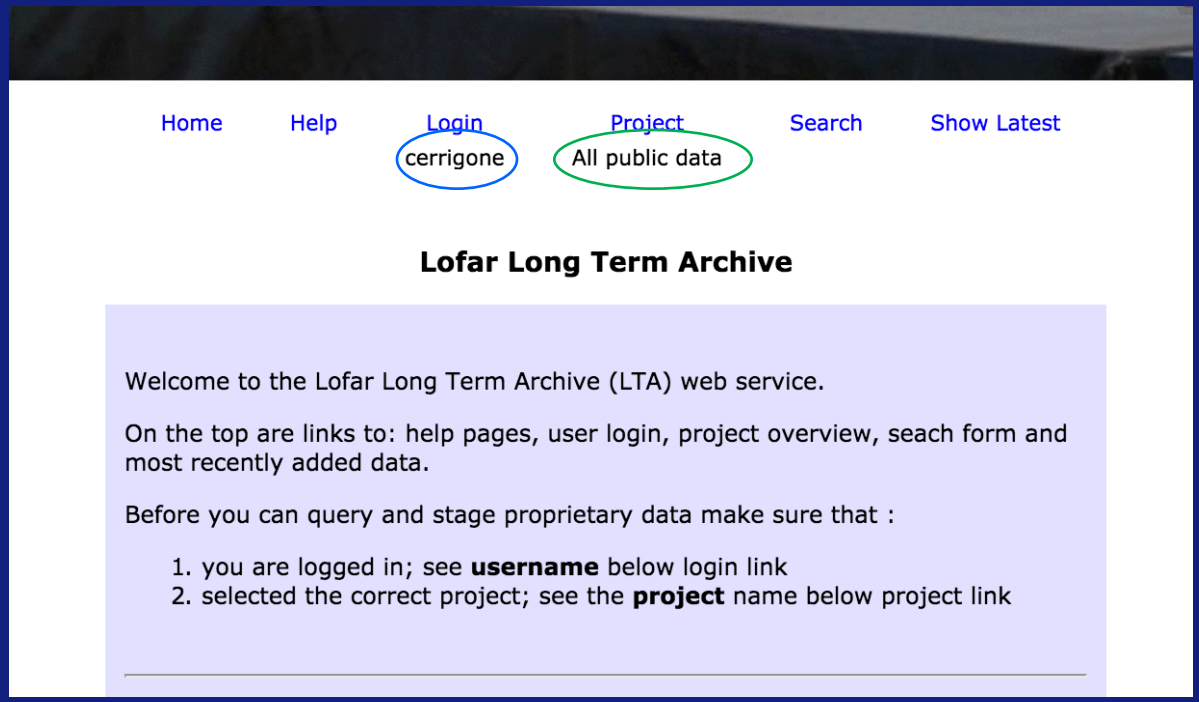
Each DataProduct type has two columns. The first column gives the number of DataProducts that can be staged. The second column gives the number of public DataProducts.

ID	Project	Privileges	Instrument	Member of	Member count	Manager(s)	Release Date	BeamFormed		Correlated		Pulsar		Image		Unspecified	
404499	DDT6_001	2	LOFAR	False	30	28		0	0	0	0	8	0	0	0	0	0
404497	DDT6_002	2	LOFAR	False	30	28	2017-06-08	0	0	488	0	0	0	0	0	0	0
404520	DDT6_003	2	LOFAR	False	33	28	2017-07-01	0	0	0	0	0	0	0	0	0	0
404538	LC6_001	2	LOFAR	False	33	28		0	0	0	0	0	0	0	0	65	0
404524	LC6_002	2	LOFAR	False	37	28		76	0	2684	0	0	0	0	0	244	0
404491	LC6_003	2	LOFAR	False	29	28		0	0	0	0	0	0	0	0	0	0
404481	LC6_006	2	LOFAR	False	36	28		0	0	10703	0	0	0	0	0	1	0
404515	LC6_007	2	LOFAR	False	36	28		0	0	22364	0	0	0	0	0	1180	0
404487	LC6_008	2	LOFAR	False	32	28		0	0	1621	0	0	0	0	0	0	0
404468	LC6_009	2	LOFAR	False	34	28		0	0	0	0	0	0	0	0	0	0
404554	LC6_010	2	LOFAR	False	35	28		0	0	0	0	0	0	0	0	0	0
404517	LC6_011	2	LOFAR	False	30	28		0	0	1520	0	0	0	0	0	0	0
404485	LC6_014	2	LOFAR	False	30	28		0	0	0	0	0	0	0	0	0	0
404459	LC6_015	2	LOFAR	False	46	28		0	0	16406	0	0	0	0	0	833	0
404540	LC6_016	2	LOFAR	False	30	28		0	0	2928	0	0	0	0	0	0	0
404493	LC6_019	2	LOFAR	False	32	28		3659	0	0	0	0	0	0	0	0	0
404451	LC6_020	2	LOFAR	False	36	28		0	0	0	0	0	0	0	0	0	0
404566	LC6_023	2	LOFAR	False	30	28		0	0	0	0	0	0	0	0	0	0
404564	LC6_024	2	LOFAR	False	32	28		0	0	0	0	0	0	0	0	0	0
404545	LC6_025	2	LOFAR	False	30	28		0	0	0	0	0	0	0	0	0	0
404465	LC6_027	2	LOFAR	False	29	28		0	0	2871	0	0	0	0	0	1	0
404455	LC6_028	2	LOFAR	False	30	28	2017-06-09	240	0	0	0	1559	0	0	0	85	0
404462	LC6_030	2	LOFAR	False	30	28		0	0	35606	0	6860	0	0	0	371	0
404568	LC6_031	2	LOFAR	False	32	28		0	0	0	0	0	0	0	0	0	0

The web interface

If you have successfully logged in, your user name will appear in the top bar of the page, under Login.

Pay attention to the Project field: you always need to choose a project to browse.

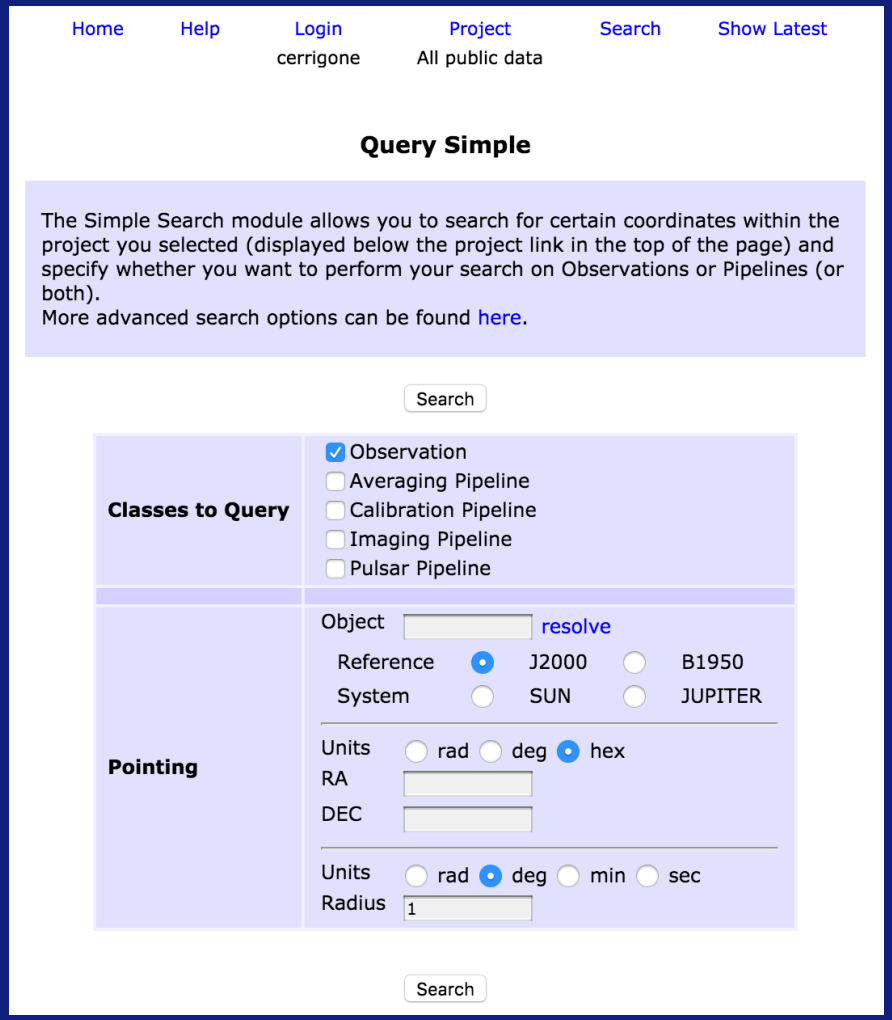


The web interface

Once you click on the Search field in the top bar, you are shown the Simple query page.

Important:
Pay attention to the Classes and their meaning.

Your search will be performed within the parameters of the classes that you have selected.



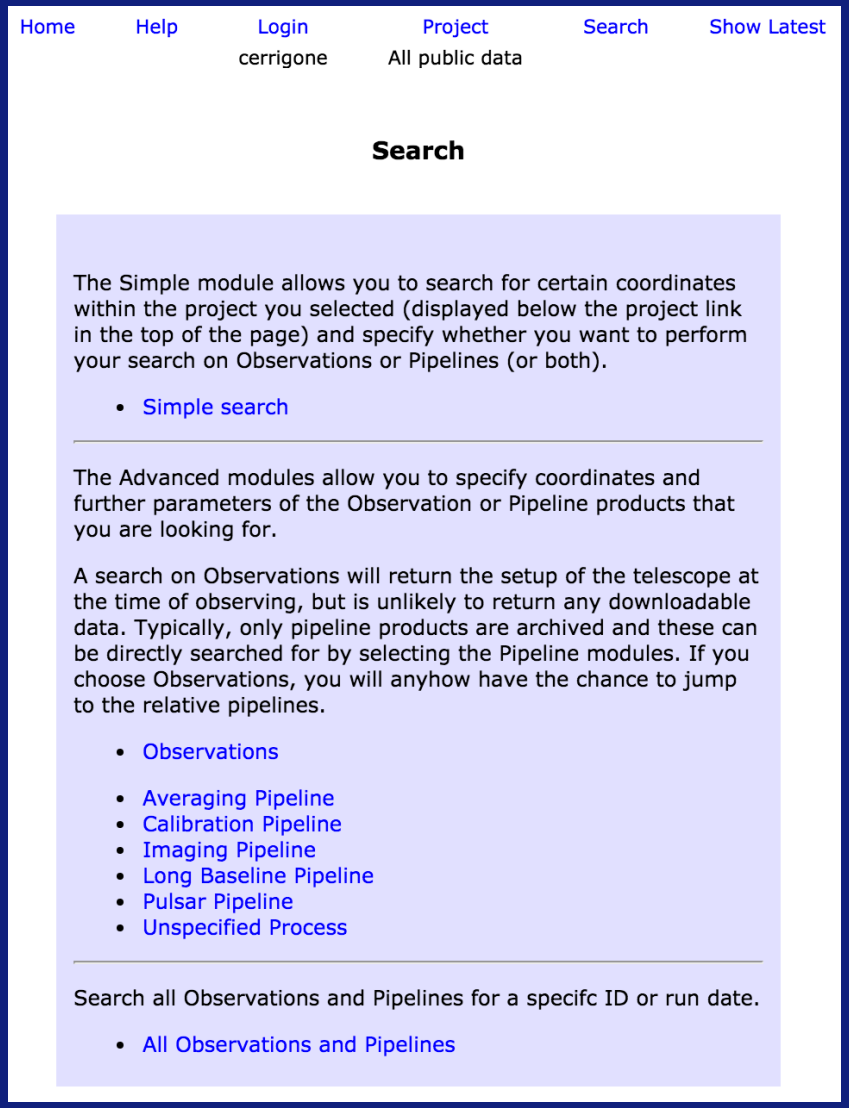
The web interface

If you follow the link to the advanced search boxes, you reach this page.

You have links to several classes, each with its own set of parameters (although many can be shared).

Observations almost **never** returns retrievable data. Searching on observations means searching over the products of an observation, which are raw data. With some exceptions, **raw data are not archived.**

You can use *Observations*, when you don't know what processing was performed on your data.

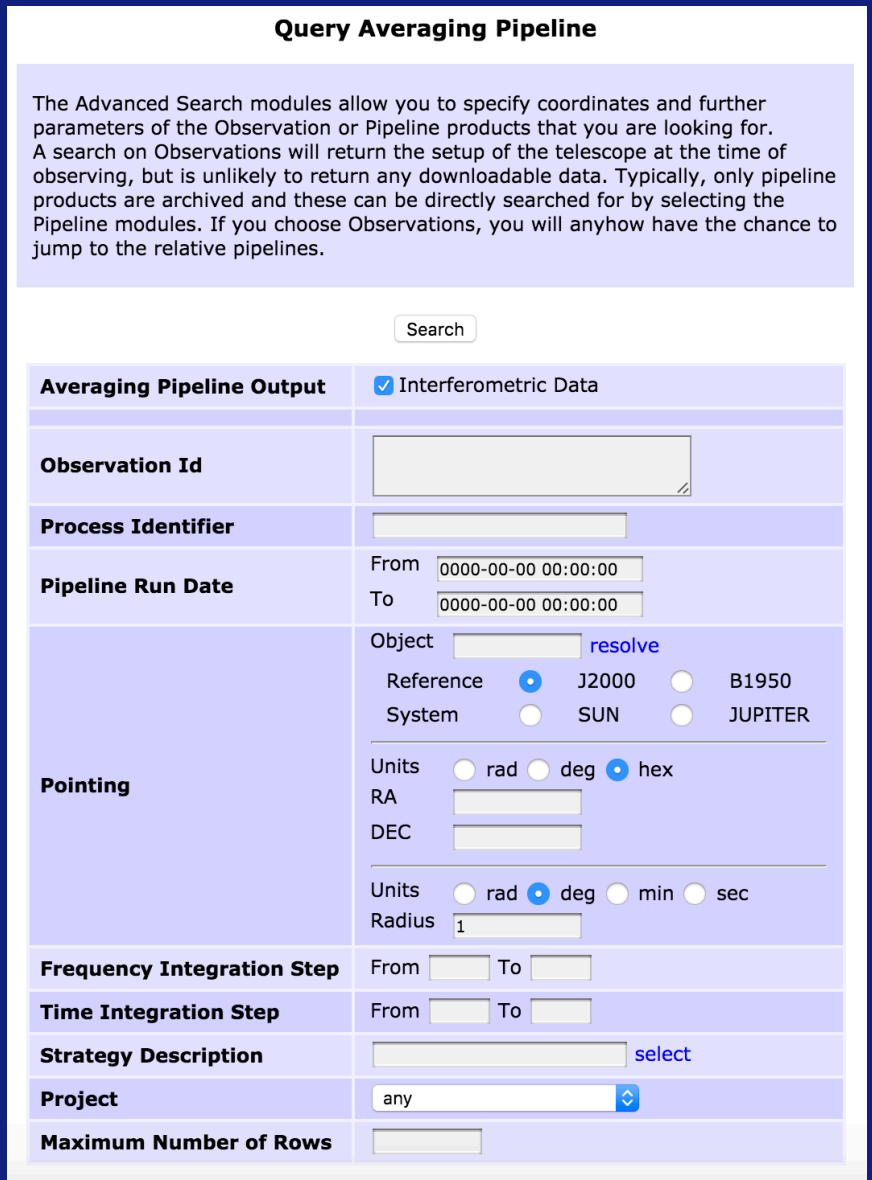


The web interface

Most data go simply through the averaging (pre-processing) pipeline.

Several parameters are very technical. In the future we will have an interface where only physical parameters are displayed by default.

You can simply search your target here. Your search will take you immediately to the products of the pipeline in the archive.



The web interface

Here is the output of a search for M 31 over All public data with the Averaging pipeline.

You can either click on the link in *Number of Correlated Data Products* or select the pipeline and click on *show data products*, for further details.

Show pipelines will tell you which pipelines were run on the product selected. In this case: none.

Source Data product will take you to the Observation setup of the selected pipeline.

Home Help Login Project Search Show Latest
cerrigone All public data

Averaging Pipeline 1 to 100 (showing 100 of total 415)

edit columns stage selected show dataproducts show pipelines

first previous 1 2 3 4 5 next last

#	Project	Release Date	Pipeline Name	Pipeline Version	Observation Id	Process Identifier	Frequency Integration Step	Time Integration Step	Flag Auto Correlations	Strategy Name	Strategy Description	Demixing	Number Of Correlated DataProducts	Start Time	Duration [s]	Source DataProduct
1	LC3_022	2016-05-15	P320410/BEAM6/2.6/TP	n/a	338038	513235	16	5	1	Preprocessing Pipeline	Preprocessing only	1	4	2015-04-10 14:33:29	290.0	show
2	LC3_022	2016-05-15	P320410/BEAM5/2.5/TP	n/a	338036	513234	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:33:29	660.0	show
3	LC3_022	2016-05-15	P320410/BEAM4/2.4/TP	n/a	338034	513233	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:33:29	310.0	show
4	LC3_022	2016-05-15	P320410/BEAM3/2.3/TP	n/a	338032	513232	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:33:29	660.0	show
5	LC3_022	2016-05-15	P320410/BEAM2/2.2/TP	n/a	338030	513231	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:21:28	561.0	show
6	LC3_022	2016-05-15	P320410/BEAM1/2.1/TP	n/a	338028	513230	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:33:29	350.0	show
7	LC3_022	2016-05-15	P320410/BEAM0/2.0/TP	n/a	338026	513229	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:33:29	640.0	show
8	LC3_022	2016-05-15	P320410/BEAM6/1.6/TP	n/a	338018	513217	16	5	1	Preprocessing Pipeline	Preprocessing only	1	4	2015-04-10 14:21:29	210.0	show
9	LC3_022	2016-05-15	P320410/BEAM5/1.5/TP	n/a	338016	513216	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:21:29	290.0	show
10	LC3_022	2016-05-15	P320410/BEAM4/1.4/TP	n/a	338014	513215	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:21:29	580.0	show
11	LC3_022	2016-05-15	P320410/BEAM3/1.3/TP	n/a	338012	513214	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:09:29	570.0	show
12	LC3_022	2016-05-15	P320410/BEAM2/1.2/TP	n/a	338010	513213	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:09:29	310.0	show
13	LC3_022	2016-05-15	P320410/BEAM1/1.1/TP	n/a	338008	513212	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:09:29	340.0	show
14	LC3_022	2016-05-15	P320410/BEAM0/1.0/TP	n/a	338006	513211	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:21:29	570.0	show
15	LC3_022	2016-05-15	P310410/BEAM6/2.6/TP	n/a	337998	513198	16	5	1	Preprocessing Pipeline	Preprocessing only	1	4	2015-04-10 13:57:28	240.0	show
16	LC3_022	2016-05-15	P310410/BEAM5/2.5/TP	n/a	337996	513197	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 13:57:28	300.0	show
17	LC3_022	2016-05-15	P310410/BEAM4/2.4/TP	n/a	337994	513196	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:09:29	610.0	show
18	LC3_022	2016-05-15	P310410/BEAM3/2.3/TP	n/a	337992	513195	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 14:09:29	620.0	show
19	LC3_022	2016-05-15	P310410/BEAM2/2.2/TP	n/a	337990	513194	16	5	1	Preprocessing Pipeline	Preprocessing only	1	40	2015-04-10 13:57:28	290.0	show

The web interface

Here we have followed the link in the Number of Correlated Data products field of the first data set

Home Help Login Project Search Show Latest
cerrigone All public data

Interferometric Data (total 4)

edit columns stage selected filter dataproducts

#	Project	Release Date	DataProduct Identifier	Target Name	Right Ascension [degrees]	Declination [degrees]	Central Frequency [MHz]	Channel Width [Hz]	Channels Per Subband	Integration Interval [s]	Start Time	Duration [s]	SubArray Pointing Identifier	Subband	Station Subband	Stations	Observations	Pipeline	Derived DataProducts
1	LC3_022	2016-05-15	15431731	BEAM6	10.6847	41.3521	149.0234	48828.125000	4	10.01390	2015-04-10 11:20:26	660.0	513228	242	0	show	1	P320410/BEAM6/2.6/TP	
2	LC3_022	2016-05-15	15431729	BEAM6	10.6847	41.3521	148.6328	48828.125000	4	10.01390	2015-04-10 11:20:26	660.0	513228	240	0	show	1	P320410/BEAM6/2.6/TP	
3	LC3_022	2016-05-15	15431732	BEAM6	10.6847	41.3521	149.2188	48828.125000	4	10.01390	2015-04-10 11:20:26	660.0	513228	243	0	show	1	P320410/BEAM6/2.6/TP	
4	LC3_022	2016-05-15	15431730	BEAM6	10.6847	41.3521	148.8281	48828.125000	4	10.01390	2015-04-10 11:20:26	660.0	513228	241	0	show	1	P320410/BEAM6/2.6/TP	

While by following the link to Source Data product, we obtain the observational setup

Home Help Login Project Search Show Latest
cerrigone All public data

Interferometric Data (total 4)

edit columns filter dataproducts

#	Project	Release Date	DataProduct Identifier	Target Name	Right Ascension [degrees]	Declination [degrees]	Central Frequency [MHz]	Channel Width [Hz]	Channels Per Subband	Integration Interval [s]	Start Time	Duration [s]	SubArray Pointing Identifier	Subband	Station Subband	Stations	Observations	Pipeline	Derived DataProducts
1	LC3_022	2016-05-15	15424951	BEAM6	10.6847	41.3521	148.6328	3051.757812	64	2.00278	2015-04-10 11:20:26	658.0	513228	240	249	show	1		AveragingPipeline
2	LC3_022	2016-05-15	15424952	BEAM6	10.6847	41.3521	148.8281	3051.757812	64	2.00278	2015-04-10 11:20:26	658.0	513228	241	250	show	1		AveragingPipeline
3	LC3_022	2016-05-15	15424953	BEAM6	10.6847	41.3521	149.0234	3051.757812	64	2.00278	2015-04-10 11:20:26	658.0	513228	242	251	show	1		AveragingPipeline
4	LC3_022	2016-05-15	15424954	BEAM6	10.6847	41.3521	149.2188	3051.757812	64	2.00278	2015-04-10 11:20:26	658.0	513228	243	252	show	1		AveragingPipeline

The web interface

Finally, when you have found the data you want, you can stage it.

Depending on the load on the system, staging can be slow. Waiting times of a week are not unheard of.

Home Help Login Project Search Show Latest

cerrigone All public data

Staging Service

The following file(s) will be requested for download (count 4, total size 131.3 MB).

Cancel Submit

Size	MD5 checksum	Filename
32.8 MB	001d604422947c515740c593ca12d67b	L338038_SB240_uv.dppp.MS_ac9f6ed5.tar
32.8 MB	231ceb9cc667c964854cea157bffe2f3	L338038_SB241_uv.dppp.MS_df2ef8b2.tar
32.8 MB	bb2eae903d578568ec1debde8cb60194	L338038_SB242_uv.dppp.MS_4c04e20a.tar
32.8 MB	0cdd738bb113730d071b24edfc2f8bfc	L338038_SB243_uv.dppp.MS_c8a99ccb.tar
131.3 MB	*Total filesize*	

The download process

When you have received an email announcing that staging has finished, you can download your data.

The download speed can vary between the different LTA sites (different lines) and it depends on how you retrieve your files.

The fastest way is a globus copy (requires grid certificate and software), which should reach 100 MB/s; the slowest is http at about 10-20 MB/s.

Some users have reported slower speeds and these need to be investigated.

Alternative interfaces

http://www.lofar.org/operations/doku.php?id=public:lta_tricks

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.

edit

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:

Observation Id	432000:432190
Observing or Pipeline Run Date	From 0000-00-00 00:00:00 To 0000-00-00 00:00:00
Project	any
Maximum Number of Rows	

In textual entries, wildcards can be used.

Target Name	3c19*

Alternative interfaces

If you are familiar (or wants to become to) with SQL, you can run your own queries on the catalog

DBView

edit

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
```

Alternative interfaces

AstroWise also has a Python interface that can be used to find your data and stage it.

This is useful for projects with many files that cannot be staged all at once: write your scripts to stage in chunks.

Examples of scripts are in the Wiki.

Attention: you must understand very well what each parameter means (for example, subband vs. station subband).

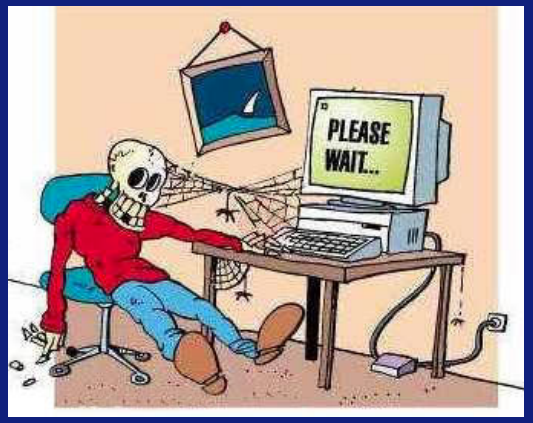
```python
# python code
from pprint import pprint
from common.database.Context import context
from awlofar.main.aweimports import Observation, Pointing, SubArrayPointing
result = {}
for project in sorted(context.get_projects()) :
    print "Project %(project)s" % vars()
    ok = context.set_project(project)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

Conclusions

Bug fixes and general improvements are necessary for the web interface. The plan is to develop from scratch a new tool for archive browsing.

Most of the time finding, staging, and downloading data is not painful, but several points need familiarization and can imply delays.



To skip manual work, searching and staging through python scripts can be a powerful alternative, but test your scripts well: do not try to stage the entire archive!