Author:Corina Vogt Michael Wise	Date of issue: 2008-04-04 Kind of issue: public	Scope: wp/science/site/… Doc.id: LOFAR-ASTRON-doctype-ddd	LOFAR
	Status: draft Revision nr: 0.9		

LOFAR Archive and Reprocessing Requirements

Verified:			
Name	Signature	Date	Rev.nr.

Accepted:		
Work Package Manager	System Engineering Manager	Program Manager

Distribution list:

Group:	For Information:
ASTRON	

Document history:

Revision	Date	Section	Page(s)	Modification
0.1	2006-09	all	all	Creation
0.9	2008-04	all	all	Revision after the de-scope

Table of contents:

1 Introduction

1.1 Purpose of this document

This document is written to facilitate the design of a LOFAR archive. Although LOFAR consists of radio astronomical, geophysical and agricultural applications, this document is focused on the description of the archive requirements of the astronomical application. The latter application is by far the most demanding in archival (storage and processing capacities) requirements. This document assumes that the other applications can use the framework of the archive described below but that it does not demand any other functionality than described.

1.2 Executive summary

The archive of LOFAR is not a classical static archive that stores all data in one place and the user downloads the data to its local machine where he/she will process the data further. The requirements on storage space and processing capabilities are high. In view of funding, choices will have to be made concerning how much (UV) data to store and what level of processing capabilities to provide.

1.3 Current version

The current version is a complete revision of the requirements which takes full account of the re-scope of LOFAR. Furthermore use cases and time lines have been added and/or updated.

2 LOFAR Observatory operation model (needs changing)

The Radio Observatory (RO) at ASTRON is responsible for the astronomical exploitation of LOFAR, and is focused on allowing the full user community to maximise the scientific output of LOFAR in the long term. The RO will take an integrated approach to user support, which will be offered in the form of online portals, visitor facilities, documentation, and advice from science support staff. It will cover the beginning-to-end chain that runs from planning projects and preparing proposals, setting up observations, observing and processing, through to supporting access to the data in the archive. For these functions, a properly designed, operated, and maintained archive is a crucial part of the infrastructure that the RO will present to the users.

All LOFAR users, irrespective of their project, can use the unified LOFAR portal presented by the RO to access all general facilities. RO staff members are available for advice, and they also define and safeguard appropriate and consistent data and metadata content in the overall archive.

The storage and computing facilities that are required to process all data immediately following observation are likely to be co-located with the central processor, and are thus part of what has been termed Tier 0.

Longer term storage and (re)processing capacity are likely to be physically distributed over several Data Centres, and may involve facilities in the Netherlands as well as centres located abroad using EGEE GRID or other functionality. These resources, which are likely to expand in volume as time goes on, have collectively been termed Tier 1.

Operationally, distinctions in the use between Tier 0 and Tier 1 are not sharp. In order to be responsible for an optimal technical and scientific performance, RO support staff need access to the integrated archive, in which the operational (meta)data is kept together with the astronomical (meta)data. Equally importantly, RO support staff may need to redistribute data flows and (re)processing streams, as needed as to maximise the long-term scientific output of LOFAR and to allow optimal resource allocations for individual user research projects.

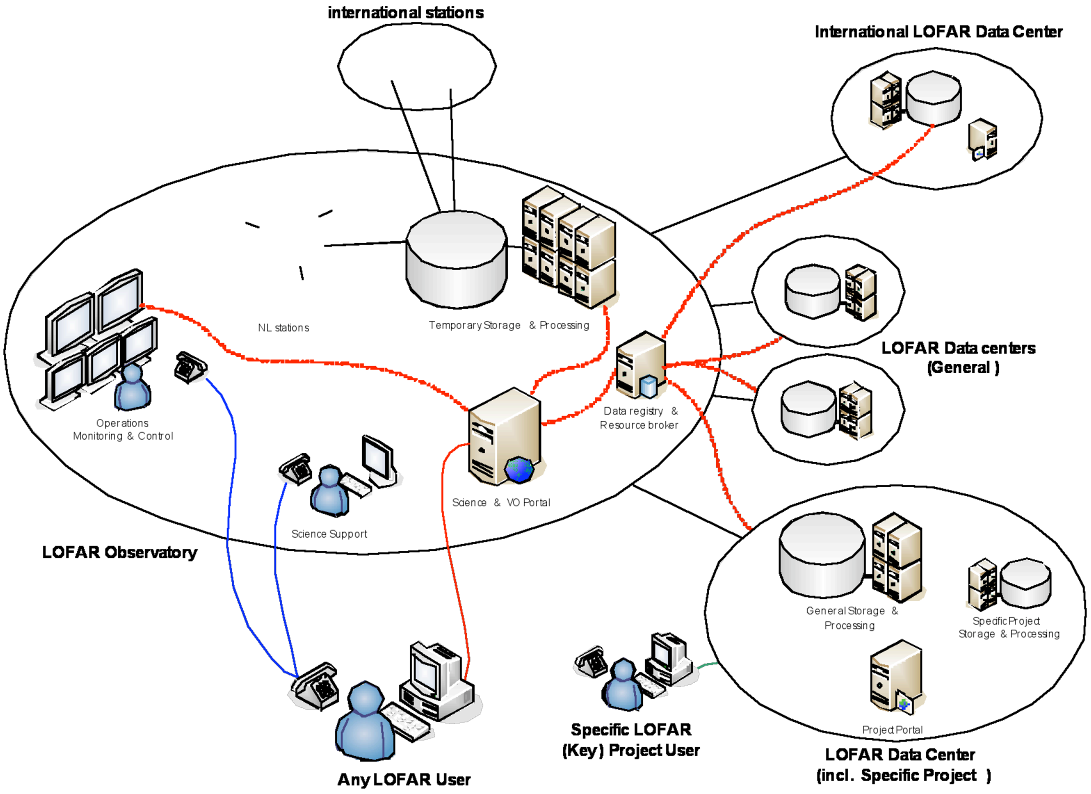
Thus the data and metadata stored in all Tier 0 and 1 facilities are part of the integral LOFAR archive, operated under the data access and distribution policy of the LOFAR Observatory, which retains the long-term ownership and responsibility for this data. The policy provides for public access, subsequent to, or next to, specific exclusive usage of particular data during a limited period, as granted to research groups in response to observing proposals and archive processing requests, assessed by the Programme Committee. Likewise, a uniform allocation policy governs the shared use of the (presumably limited) storage and processing capacity available in both Tier 0 and Tier 1, and is safeguarded by the RO.

Data registry and resource brokerage facilities for Tier 0 and 1 are implemented under central responsibility of the RO, and distributed down the chain as needed, to ensure the integrity of the archive, and to allow efficient operation. All LOFAR users have access to Tier 0 and 1 storage and processing via the RO portal, which also facilitates VO publishing.

A particular Data Centre may also be associated with a research group engaged in a specific LOFAR (Key) Science Project. Such a centre can then, additionally, offer processing and storage facilities dedicated to the members of that science group. This function has been termed Tier 2. It may contain elements outside the unified structure of LOFAR user support and archive, such as derived data products pertaining specifically to the individual research project. Some additional data products may, after some time, be made accessible by the Data Centre to general users as part of the Tier 1 common LOFAR archive; indeed, in some cases such product delivery may be one of the conditions under which the original rights to observe and to use the data were granted to the research group.

Figure 1 sketches the implications for the way access to the data and (re)processing capacity are organised, as they follow from the general model for LOFAR user interaction and support. Blue: human interaction with RO support scientists and operators is available to all users by e-mail, telephone, or visits to the RO. Red: network interaction with the integral LOFAR archive is available to all users via the LOFAR science portal of the RO. Green:

interaction with the facilities of a specific LOFAR Project Centre is available to members of that particular research group. Solid lines depict direct connections; dotted lines depict indirect connections.



international stations
International LOFAR Data Center
NL stations
Temporary Storage & Processing
Operations Monitoring & Control
LOFAR Data centers (General)
Data registry & Resource broler
Science & VO Portal
Science Support
LOFAR Observatory
General Storage & Processing
Specific Project Storage & Processing
Project Portal
Specific LOFAR (Key) Project User
LOFAR Data Center (incl. Specific Project)
Any LOFAR User

Figure 1: LOFAR (data) access from the user perspective

3 Users of the archive

In the following the potential users of a LOFAR archive and their expectations from the LOFAR archive are summarised.

3.1 Guest observer

The guest observer requests through a proposal an observation that - if the observation time is granted - will be performed, data will be recorded and processed by on-line pipelines. The data that are recorded for the guest observer will be either uv-visibilities or antenna time series. After on-line processing a final data product - which still needs to be defined but in general - will be something like a UV-data cube, a time series or an image will be generated. This data product will be transferred to a storage facility from which the guest observer can download or further process the data. The transfer to the long-term storage facility will happen within a week or two after the observation has been done. This observer will have privileged access to the data and only the specific observer associated with that particular observation will be granted access to and process the data for a proprietary period whose length still needs to be defined by the LOFAR Time Allocation Committee.

For data quality assessment, the guest observer will most likely require data concerning the specific observation settings and environmental data.

3.1.1 Key science projects

The key science projects are specific examples of guest observers. They will perform dedicated observation campaigns and have defined data products which they intend to store in an archive. They are special in that they will be the main user of the astronomical facility and produce the largest data sets in the first phase of the telescope operation. In that they are the first user of the archive and to some extent will set the requirements for a LOFAR archive.

Currently, there are six key science project: Epoch of Reionisation, Survey, Transient, Cosmic Ray, Cosmic Magnetism and Solar physics.

3.2 Operations

The performance of the telescope is monitored by the *operators*. The operator is the person who will set the observation setting and will be most interested in operation data such as what stations/antennas have been active and which not. The operator might be also interested in weather information.

There will also be maintenance specific test observations that need archiving at least for a specific period of time. Since these test observation could contain also interesting data – such as a transient – it might be useful to archive them. These test data – if not requested otherwise – should be archived and open for the general archive user from the point of transfer to the long term storage facility.

The operator would be more concerned with the short term monitoring of performance and thus most likely interested in performance (meta) data from the last couple of months. The operator might not need access to all the data cube but the operator might want to perform quality checks on the recorded data itself, i.e. check the quality of data once they are put on storage permanently.

The *System Engineer* investigates technical issues found during the operation of the instrument. Defines and realizes solutions to these issues. He/she carries out improvement programs to enhance the capabilities of the instrument. By doing this, it might be required to trace back certain problem over a length of time and the system engineer would be interested to use the archival data for that purpose. If a problem has been found the system engineer might has to reprocess the data having the problem to correct it and might also want to inform the guest observer.

3.3 Calibration/Support scientist

The calibration/support scientist will be responsible for the quality of the calibration of the data. If a problem is identified the calibration scientist might need to check previous observations for similar problems.

3.4 Archival user

The archival user is a user who will access the data of observations after the proprietary period of these archived data has expired. Such a user will browse the content of the archive, select, retrieve and/or process data thereby it will be essential for the archival user that a reconstruction of the observational setting, the calibration procedure and other operation information are available.

3.5 Management

The management will require certain archive statistics about the usage of the archive facilities.

4 Components of the archive

Starting from the potential users, their potential interests and their usage of the archive, the components that the archive should consist of are described in the following.

4.1 Interfaces

The user needs interfaces to the various components, i.e. observation catalogues, processing pipelines, retrieving data. These interfaces should possibly be available through the internet via web interface. The user should be able to retrieve all information from one site without the user actually knowing where what is stored and processed.

4.2 Observation Catalogue

As there will be a lot of observations, the various users will need to search an observation catalogue to find a specific observation. Thus, the observation catalogue will need at least the following entrances:

- Observation title
- Observation date
- Observer (PI)
- Observation (principle) target (RA, DEC)
- Observation duration and integration time
- Observation type and configuration (core, full, E-LOFAR)
- Observation frequency and spectral mode
- Type of data product/format

4.3 Instrument/Health database

While an observation, the system continuously monitors its functionality and delivers data about the performance of the telescope such as which antennas/stations are working and where problems are. These are the so-called instrument/health data. These data should be archived if possible with the associated actual scientific data and made searchable separately as well for the operators so that specific problems with single antennas or stations can be monitored and short and long-term problems can be detected if necessary.

4.4 Associated Metadata databases

Apart from the system health information, there are other information important to the user in order to interpret the data of a specific observation correctly. These data will be summarised in the Metadata base which contains:

- Instrumental configuration & observation setting
- Observation logbook containing:
  - weather information from station
  - information from the operator
  - information from the correlator
- Processing configuration (information on how the final data product was derived)
- Calibration parameters & strategy

4.5 Data storage

A significant component in size of the LOFAR archive will be the actual data storage facility. This data storage facility will have to archive a significant amount of data (about Petabyte/yr) permanently. This data will be the scientific data provided to the scientific community following the observation and a first reprocessing on the central reprocessing facility. The user can retrieve the data sets but might also choose to reprocess it further with the archive/project provided. Due to the large data sets it might be advisable to provide the user with a certain working space where intermediate products of reprocessing can be stored for a certain amount of time. This working space should be only available for this particular user and for a certain amount of time. This disk space might have to be requested in the initial observation proposal of the guest observer.

The general guest observer might not want and will not have to archive its final data products after re-processing the archived data off-line. However, the key science projects have committed themselves to make their final data products available to the general user – the archival user. These final data products contain various databases (such as source catalogues, event catalogues, image catalogues, etc.) that will need to be accessed by the archival user and need to be made searchable.

4.6 Processing software

The data which are delivered from the telescope following a (dedicated) observation will be to some extend already processed. The guest observer and/or the archival user might find that the provided quality of the data is not sufficient enough for his purposes and would like to process the data further. As the data cubes are rather larger it becomes increasingly difficult to re-process the data on a home computer. Furthermore, the user might only want to inspect the data cube and transferring huge amount of data for that purpose might not be desirable.

The processing software will need to support at least the following:

- Visualise data
- Add/subtract data
- Running the software modules of the general LOFAR pipeline (flag data; calibrate data, Imaging data)
- Running KSP specific pipelines, ie. data mining
- Retrieve only part of the observation/data

Within the framework of the User Software group a software package is developed which contains part of this processing software. However, this software then needs to be compatible and integrated with the archive environment.

4.7 Computing resources

Since the data sets will be rather large, the user should have the possibility to process the archive data on-line or view the full dimensionality of the data. Especially the re-processing of uv-visibilities, ie. the processing steps from the on-line pipeline should be possible on the provided computer resources.

5 Archive data content

The data products for the general guest observer that will be transferred and stored in the archive are uv data, n-dimensional images and time series.

The **uv data** are in table format. Most likely it will be in HDF5 or FITS format. It might be advisable to support both data formats.

The **images** produced by the on-line processing pipeline can be n-dimensional. The dimensions possible are the x,y locations, the time steps, the frequency channels and the four Stokes parameters (I, Q, U, V) for polarisation work. This gives maximum n=8. It depends from the request of the guest observer what size the image should have.

Antenna Time series will most likely be in a binary format.

The meta data associated with each observation data cube should be stored in databases.

Table 1: A summary of the potential archive user and the LOFAR data product which they will access and if they require to re-process the data.

User	Archived data product/information	Re-processing desired	Access
Guest observer (incl. KSP)	- UV-visibilties - spectral images - antenna time series - source lists - light curve database - event catalogue	- visualise data - add/subtract data - re-processing - retrieve part of the data	Privileged access for a proprietary time to data from his/her observation
Operator	- instrumental configuration - observation log - instrumental health information - processing configuration	No	Access to all data and databases for the purpose of inspection
System Engineer	-	Yes	Access to all data and databases for the purpose of correction of problems
Calibration scientist	- calibration parameters/solutions	No	Access to all data and databases for the purpose of inspection
Archival user	-	Yes	Access to all data after a proprietary time has passed
Management	-	No	Access to all data and databases for the purpose of inspection

6 Archive services

As already mentioned the general LOFAR data product will be large, the LOFAR archive should offer more than the general data searching and retrieving services of any static archive. In particular, it should allow the user to visualise data cubes, split data before downloading and re-process the data if necessary.

6.1 Data Re-processing

As some of the planned observations reach the limits of the telescope and its calibration capabilities, it is advisable to store the data for those particular observations in a format that is as raw as possible (ie. as uv-visibilities). These raw data can then at a later stage with improved algorithm and procedures offline re-processed to improve the image quality. For some applications (ie. for the transient KSP), storing uv-visibilities might actually be the most cost effective way to keep the information required. However, the scientific user is most likely interested in a good quality image and thus, re-processing the uv-visibilities and as a result produce an image is essential.

Furthermore, a user might want to subtract sources from a data cube, be it uv-visibilities or an image, compare data, add and subtract data from different times and might want to perform other data manipulation on them.

As a conclusion the archive should at least allow the user to use software packages that perform these manipulation on the data.

6.2 Data Examination

The user should be able to visualise the data requested. That could be the data in an image or in the uv-visibilities to inspect them for quality. Operators and calibration scientist will be interested to visualise database data and statistic thereof.

6.3 Data searches

As only the guest observer for a specific observation will know exactly which data to get it is absolutely crucial that the observation catalogue is searchable for various keywords:

- Observation title
- Observation date
- Observer (PI)
- Observation target (RA, Dec)
- Observation duration and integration time
- Observation type and configuration (core, full, E-LOFAR)
- Observation frequency and spectral mode
- Type of data product/format

The operator or the support scientist might need to search the content of the meta-data and search for observations that were done under specific weather conditions, done with specific calibration parameters and so on.

6.4 Data mining

7 Archive resource estimate

7.1 Storage space

In the following, an attempt is done to estimate the amount of storage space for a standard data product and for the total storage space required over the first couple of years.

As discussed earlier, the following three data products will be archived: uv-visibilities, n-dimensional images and antenna time series. In the following estimates are given for the amount of storage space required for various modes and scenarios.

In general, there is a minimum number of stations namely 18 core, 18 remote and 7 E-LOFAR stations. The maximum number of stations is 25 core, 25 remote and 20 E-LOFAR stations.

UV-visibilities

In general the uv-data size for any observation can be calculated as follows[1] for 160 MHz sampling rate:

$N_{stat} * (N_{stat}+1)/2 * 4\ \mathrm{pol} * 2*32\mathrm{bit} * N_{beams} * 6{,}656\ \mathrm{channels} * (\Delta\nu_{channel}/0.61\mathrm{kHz}) * t_{obs} / t_{int}$,

and for 200MHz sampling rate:

$N_{stat} * (N_{stat}+1)/2 * 4\ \mathrm{pol} * 2*32\mathrm{bit} * N_{beams} * 5{,}248\ \mathrm{channels} * (\Delta\nu_{channel}/0.76\mathrm{kHz}) * t_{obs} / t_{int}$

where N_{stat} is the number of stations, t_{int} is the integration time, t_{obs} is the duration of the full observation in s, 6,656 or 5,248 channels representing 4 MHz bandwidth and N_{beams} is the number of beams can vary between 1 and 8 for one observation. One beam does not need to be 4 MHz but it is assumed here for simplicity. In the following, a sampling rate of 160 MHz rather than 200 MHz is assumed, as this rate needs more storage capacities. It might be possible that a limited amount of stations (most likely in stations in the core) could also observe with 4 beams each having 32 MHz. That introduces a factor of four in the above equation. One should keep this factor in mind in the discussion.

LOFAR will not always observe with all stations at the same time but with the core, with LOFAR NL and E-LOFAR for certain amounts of time.

In Table 2, a summary is given over the uv-data sizes for various core station scenarios that would have to be stored. Core observations do not need frequency and time resolution of 0.61/0.76kHz and 1s, respectively. Thus, the final data product might be averaged to 5s-5kHz data that would reduce the numbers by quite already as can be seen in the table below. Some observations plans might require higher frequency and time resolution and some request lower resolutions.

[1] The autocorrelations are written as well, so that the output scales with N(N+1)/2 rather than N(N-1)/2.

Scenario	N_{stat}	t_{int}/s	t_{obs}/h	N_{beams}/4MHz	Delta f / kHz	UV-data size / Gbyte
Core: 18 LBA	18	1	4	1	0.61/0.76	516/415
Core: 18 LBA	18	1	4	8	0.61/0.76	4130/3300
Core: 18 LBA	18	5	4	1	5/5	12
Core: 22 LBA	22	1	4	1	0.61/0.76	764/610
Core: 25 LBA	25	1	4	1	0.61/0.76	981/785
Core: 25 LBA	25	5	4	1	5/5	23
Core: 36 HBA	18	1	4	1	0.61/0.76	1900/1500
Core: 36 HBA	18	1	4	8	0.61/0.76	15200/12200
Core: 36 HBA	18	5	4	1	5	45
Core: 44 HBA	22	1	4	1	0.61/0.76	2860/2290
Core: 50 HBA	25	1	4	1	0.61/0.76	3700/2960
Core: 50 HBA	25	5	4	1	5	87

Table 2: UV-data sizes for a couple of core station scenarios. The channel bandwidth is 0.61 kHz for 160 MHz sampling rate and 0.76 kHz for 200 MHZ sampling rate. A single beam is assumed to be 4 MHz. As can be seen from the numbers above average the data reducing the data sizes quite drastically.

In , a summary is given for uv-data sizes that will be produced. Please note that one of the remote station is actually a core station from the layout. Again for some observations it might be possible to reduce the time and frequency resolution to 3s-3kHz, which would reduce the numbers considerably.

Scenario	**N_{stat}**	**t_{int}/s**	**t_{obs}/h**	**N_{beams}/4MHz**	**$\Delta\nu$ / kHz**	**UV-data size / Gbyte**
LOFAR NL: 36 LBA	36	1	4	1	0.61/0.76	1900/1500
LOFAR NL: 36 LBA	36	1	4	8	0.61/0.76	15200/12200
LOFAR NL: 36 LBA	36	3	4	1	3	130
LOFAR NL: 44 LBA	44	1	4	1	0.61/0.76	2860/2290
LOFAR NL: 50 LBA	50	1	4	1	0.61/0.76	3700/2960
LOFAR NL: 50 LBA	50	3	4	1	3	250
LOFAR NL: 55 HBA	55	1	4	1	0.61/0.76	4500/3600
LOFAR NL: 55 HBA	55	1	4	8	0.61/0.76	36000/28000
LOFAR NL: 55 HBA	55	3	4	1	3	300
LOFAR NL: 67 HBA	67	1	4	1	0.61/0.76	6680/5340
LOFAR NL: 76 HBA	76	1	4	1	0.61/0.76	8600/6900
LOFAR NL: 76 HBA	76	3	4	1	3	574

Table 3: UV-data sizes for a couple of LOFAR NL scenarios. The channel bandwidth is 0.61 kHz for 160 MHz sampling rate and 0.76 kHz for 200 MHZ sampling rate. A single beam is assumed to be 4 MHz. As can be seen from the numbers above average the data reducing the data sizes quite drastically

For the European LOFAR, the frequency cannot be averaged and integration times of 0.25s are necessary. This mode produces quite a lot of data. It is not clear yet how the output data rate could be reduced, ie. the data that would need to be stored permanently. For the moment, only one scenario is considered in. 18 core, 18 remote and 7 E-LOFAR stations.

Scenario	N_{stat}	t_{int}/s	t_{obs}/h	N_{beams}/4MHz	Δν / kHz	UV-data size / Gbyte
E-LOFAR: 42 LBA	42	0.25	4	1	0.61/0.76	10400/8300
E-LOFAR: 63 HBA	63	0.25	4	1	0.61/0.76	23600/18900

Table 4: UV-data sizes for the case of a LOFAR with 18 core, 18 remote and 7 European stations. LOFAR will most likely grow beyond that but the data output rates would have to be reduced in the first place due to the BlueGene/L capabilities. This can be thus seen as the maximum E-LOFAR scenario.

N-dimensional images

A full spectral images: (4096x4096)pixels*1beam*$N_{channel}$*4hours*4pol*4byte
= 14,400 TBytes

This is the maximum size of an image data cube containing all information from the observation. A typical observation requesting an image as data product would in general require no time (or at least a drastically reduced) resolution and a reduced frequency resolution, ie. 10KHz channels instead of 0.61/0.76kHz. This reduces the data size already to 110 GBytes. Due to the decreased number of antennas and the associated increased field of view, the image size of 4096x4096 might not be sufficient to image the full beam anymore thus increasing the image size by a factor 2 to 4.

With the above calculation it becomes also clear that if information with full time resolution is required it is less costly to store the UV-visibilities. Furthermore, archiving core data as uv-visibilities might be less storage demanding if the averaging is possible. The ultimate question then is if the available re-processing resources needed to reproduce the requested images are available from the LOFAR archive or at the user site.

Antenna Time series

Antenna time series will be recorded when a trigger is initiated that sends a certain amount of antenna time series samples to the central processing system from which are certain amount of triggered data will be transferred to the storage facility.

LOFAR core: 25 LBA stations x 48 antennas x 2pol x 2^{17} samples x 2bytes/sample = 0.6 Gbyte

LOFAR NL: 50 LBA stations x 48 antennas x 2pol x 2^{17} samples x 2bytes/sample = 1.2 Gbyte

E-LOFAR: (50 LBA stat x 48 ant + 10 stat x 96 ant) x 2pol x 2^{17} samples x 2bytes/sample = 1.7 Gbyte

Assuming an event rate of 1 event per 10 minutes this amounts to 15 Gbyte per 4 hours for core observations and 30 GByte for LOFAR NL observations.

TBD: pulsar modes and the various CR modes

Total storage Requirements

The challenge is now to estimate the total storage space required for the next years and especially the amount of storage space required once LOFAR is fully operational.

It is not clear yet which data product the guest observer will eventually request. Considering the special case of the KSPs, one could expect a mixture between uv-visibilities, images and antenna time series to be requested to be archived in a long-term storage facility. Thus, a possible scenario could be the following:

- 1 year consists of 360 days
- 60% of that can be used to observe, ie. on 216 days can be observed or 5184 hours
- Observations are split as: 45% core, 45% NL and 10% E-LOFAR
- Observations 50% dedicated to HBA observations and 50% to LBA observations

- Data format requested:
  - 30% of observers request UV-visibilities,
  - 60% of observers request images
  - 10% of observers request antenna time series
  - For long baseline experiments 30% visibilities and 70% images
- LOFAR scenario: 18 core, 18 remote and 7 E-LOFAR stations

As standard products are the following assumed:

- UV-data core: 5s-5kHz averaged data
- UV-data LOFAR NL: 3s-3kHz averaged data
- UV-data E-LOFAR no averaging
- Image: 10 kHz channel bandwidth, 4 polarisation
- Image size: 5333x5333 pixel
- Antenna times series are as defined above

The above assumption result into about 4.4 Petabyte. In the 4.4 Petabyte the most demanding item are the long baseline observations with about 2 Petabyte, if archived in full resolution as uv-visibilities. The excel sheet for this example calculation can be found in Appendix A: Storage Requirements Calculation. The total storage requirement can already be reduced to 3.5 Petabyte by just setting the average channel width for an image from 10kHz to 20 kHz. If one requests only one polarisation per image then the total storage requirements are about 2.7 Petabyte. Thus, it seems difficult to give an accurate estimate as there are too many unknowns. The required amount of storage space will most likely be between 1-4 Petabyte per year.

There is an additional difficulty in the above calculation that enters through the so-called piggy-backing mode. Observers can piggy-back on observation of other observers and might store the “same data” in a different format thereby adding to the data storage need for a single observation, ie. increasing the total storage required.

Considering that in radio astronomy, it is traditional that the raw data, ie. uv-visibilities, are archived and transferred to the observer who then reprocesses the data and generates an image another scenario should be discussed. In addition storing uv-visibilities is – in some circumstances – more effective in terms of required storage space versus information content. On top of that manipulating data (ie. adding data) is more effective in the uv-plane. Furthermore, improved calibration algorithms might improve the generated images from the raw data even years after the observations have been done. The case for archiving all data as uv-visibilities is discussed in more detail in Appendix B. In conclusion, it might be feasible to archive all LOFAR data as uv-visibility under the constraint that the data can be averaged in a suitable fashion and the re-processing of the data is not too demanding.

LOFAR will be fully operational 2010/2011. Before then the required amount of data storage will be less. The number of stations will be smaller. On the other hand it will be required to store UV visibilities rather than images. A summary of the required storage space for the next years is given in Table 2 and a more detailed discussion on timelines can be found in Appendix A.

Table 5: THESE ESTIMATES NEED TO BE UPDATED AND CORRECTED. MORE STORAGE WILL BE NEEDED. Timeline of the required storage space before LOFAR will be fully operational in 2010/2011.

Year	Storage requirement/yr	LOFAR phase
2007	10 TByte	CS1 completed
2008	35 TByte	LOFAR 20 completed 1-2 E_LOFAR stations
2009	300 TByte	LOFAR 40 completed more E-LOFAR stations
2010	800 TByte	LOFAR completed
2011	1-1.5 Petabyte	LOFAR fully operational

7.1.1 Access pattern

The access pattern will depend on the archive user.

The guest observer has proprietary access to his/her data for inspection, further reprocessing and scientific analysis. The guest observer will:

1) Request the data
2) Inspect the data
3) Upload the data onto a working area
4) Manipulate the data
5) Visualise the data
6) Release the data

Some of the KSP – as special guest observer – require to access one data set multiple times at various times. In some cases it goes even so far that a whole stack of data for one particular point at the sky observed over a year or two (or even more) is re-processed at a time.

The archive user will first search the data base – the observation catalogue, select the data, retrieve the data if no privileged access is required, process them and analysis them.

Further access patterns can be found in the use case in Appendix D: Use cases.

7.2 Computing resources

At the moment, it is difficult to judge the size of a cluster required for reprocessing LOFAR data. A reasonable assumption would be to require a cluster that is comparable in size with the on-line re-processing cluster of LOFAR. However as the size of this cluster is under investigation right now nothing more specific can be said about the computing resources required can be made.

It should be noted that providing computing resources to the user is a choice that one has to make consciously given the data volumes involved and funding available.

7.3 Software

Software which is developed for LOFAR, ie. visualisation tools, flagging tools, the software modules from the LOFAR on-line pipeline, tools for adding and subtracting data, selecting and extracting data content should be compatible and useable with the archive content on the potential re-processing facility.

For the archive three main datastreams can be distinguished:

1. Red: The primary observation (or scientific) datastream
2. Green: The metadata datastream
3. Blue: The data replication & relocation datastreams

The data streams are associated with operational use cases for the LOFAR archive and will be used to identify components and interfaces that are required for the LOFAR archive.

NB Solid lines represent the datastreams themselves while dotted lines represent the associated control connections.

8 LOFAR Archive Topology Model

8.1.1 Summary of Tier Model

In order to conform to established terminology, the LOFAR archive can be viewed as consisting of the following multi-tiered structure:

- Tier 0 is associated with LOFAR central processing. It provides temporary storage and reprocessing facilities for data produced by the LOFAR telescope system. It also provides a central portal for LOFAR users. The portal gives access to both telescope facilities and archive storage and (re)processing resources. The latter are managed through the LOFAR Data registry & Resource broker, which keeps track of data governed by the usage policy of the LOFAR Observatory, which therefore retains long-term ownership.
- Tier 1 consists of a number of sites that provide long term storage for LOFAR data. Tier 1 sites implement the LOFAR Data Policies. In general, Tier 1 sites will also provide computing resources. Resource usage (storage & computational) is managed through the LOFAR resource broker.
- Tier 2 provides services offered by groups associated with LOFAR. These services are not managed by the LOFAR Data registry and Resource broker. Tier 2 may use Tier 1 resources (i.e. stored data or re-processing facilities) in a, for the user, transparent way as long as the Tier 1 resources are controlled by the Tier 1 services (and thus managed through the Data registry & Resource broker).
- The LOFAR user may be viewed as a fourth Tier. It is important to realize that user interfaces are required through which users can access the services provided at all (other) Tiers. These include two way data connections interfaces (for retrieval of stored data and submission of derived data). However, there is no control from the other Tiers of what happens at this level. Also, Tier 0 and Tier 1 services make no distinction between data access and resource usage by LOFAR users or by Tier 2 processes.

A single geographical or organizational site may provide both Tier 1 and Tier 2 functionality. The distinction between Tier 1 and Tier 2 is in the way in which the services are managed, and in the scope of the functionality and the user community serviced. On the other hand, Tier 1 facilities can have overlapping functionality with Tier 0, but typically differ in proximity to the central processing, both physically, and in time post-observation.

8.1.2 LOFAR archive configuration and data streams

The high level architecture of the LOFAR archive is presented in **Error! Reference source not found.** including the operations data streams.

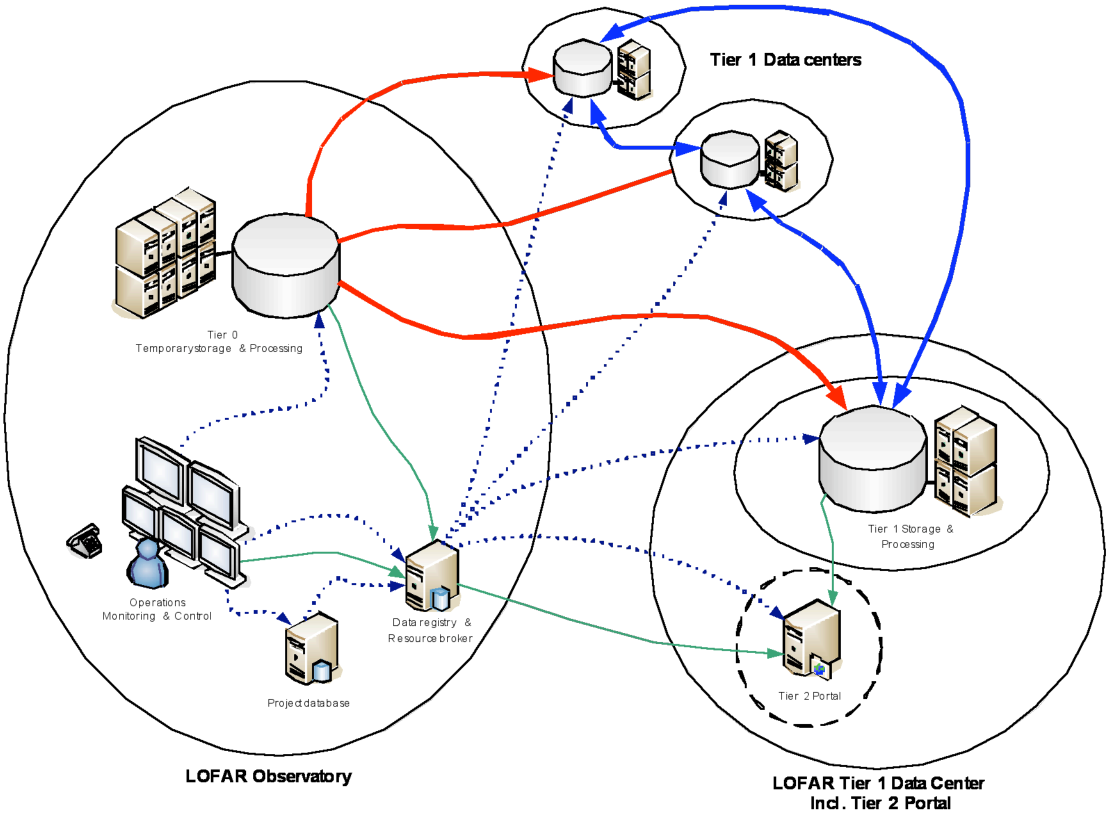


Figure 2: The LOFAR Archive

Metadata related to archived datasets, i.e. datasets that have been transferred from the temporary central (Tier 0) storage, is stored in a central database. This database, the LOFAR Data Registry, is used for querying the LOFAR archive and returns full metadata and location references for resulting datasets.

8.1.3 LOFAR Observatory (Tier 0)

For the archive user, the LOFAR Observatory is associated with Tier 0. The LOFAR Observatory provides the central services and facilities and a central portal for all LOFAR users. It also provides temporary storage and reprocessing facilities for data produced by the LOFAR telescope system.

The telescope facilities are controlled through the Monitoring and Control (MAC) applications. The SAS application provides the complete administration used for carrying out observations and processing on the Tier 0 facilities.

The project database provides the long term administration of project related processes and data products. It facilitates observation project processes (LOFAR beginning to end support) For this, it associates users with projects and stores the allocation and usage of resources for projects. Conceptually, the Project database also provides the user administration although this may be implemented as a separate (linked) application.

The science & VO portal gives access to telescope facilities, archive storage, and (re)processing resources. It provides applications to interact with the LOFAR Project database and other telescope facilities, to query the LOFAR Data Registry, and links through to Tier 1 facilities. The VO services provide query, retrieval and visualisation tools for general use. LOFAR data visualisation is provided through the Science portal (principally for visualisation of data stored at Tier 0 and possibly for data stored at Tier 1 if there are no bandwidth limitations) but Tier 1 sites may provide local analysis and visualisation services.

LOFAR Tier 1 archive and (re)processing resources are managed through the LOFAR Data registry & Resource broker. It is linked to the Project database to determine data access authorization and access to processing resources.

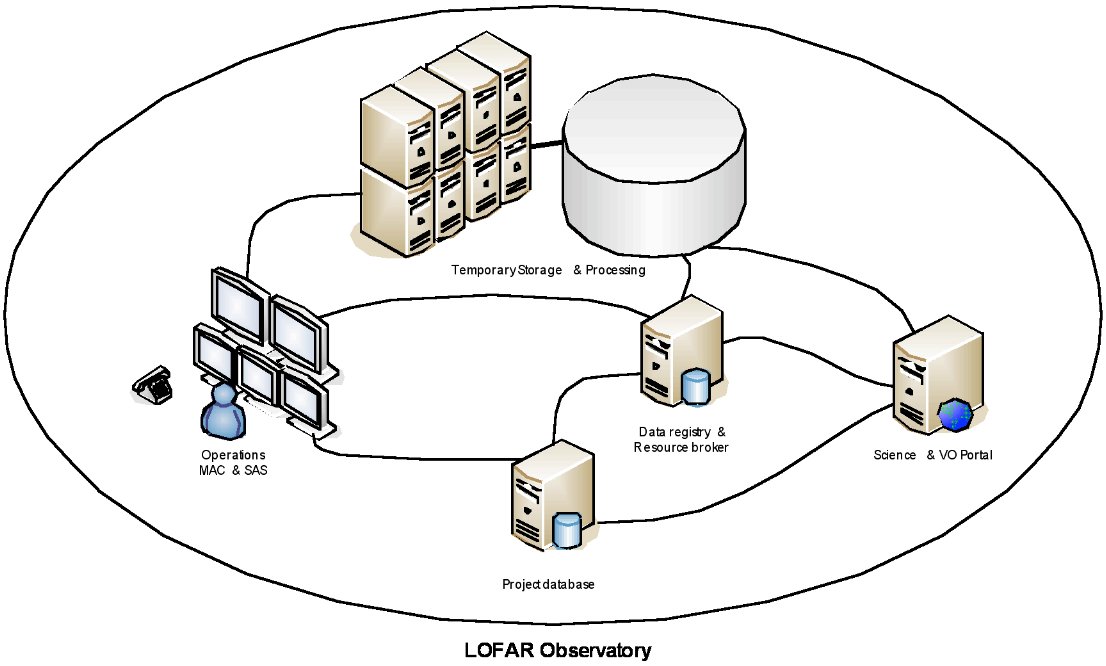


Figure 3: LOFAR Tier 0

8.1.4 LOFAR Tier 1 archive sites

It is essential that LOFAR Tier 1 sites implement the interfaces and policies defined for LOFAR data retrieval, storage, and processing. Any site implementing these can be a LOFAR Tier 1 site. A possible architecture for a LOFAR Tier 1 site is presented in figure 4. It includes a local storage manager component that communicates with the LOFAR Data registry & Resource broker and implements the LOFAR data interfaces and policies.

The presented architecture includes components and functionality that add to the compulsory functionality. The Tier 1 (re-) processing functionality (including a local processing manager) provides (re) processing resources and e.g. data analysis and visualisation services. It is managed through the (central) Resource Broker.

Tier 2 functionality may be provided as well and would be independent of the LOFAR Observatory. Integration with the local Tier 1 facilities may include using shared hardware as long as there is a clear separation between Tier 1 and Tier 2 data and processes and the Tier 1 resources are managed through the LOFAR Data registry & Resource broker.

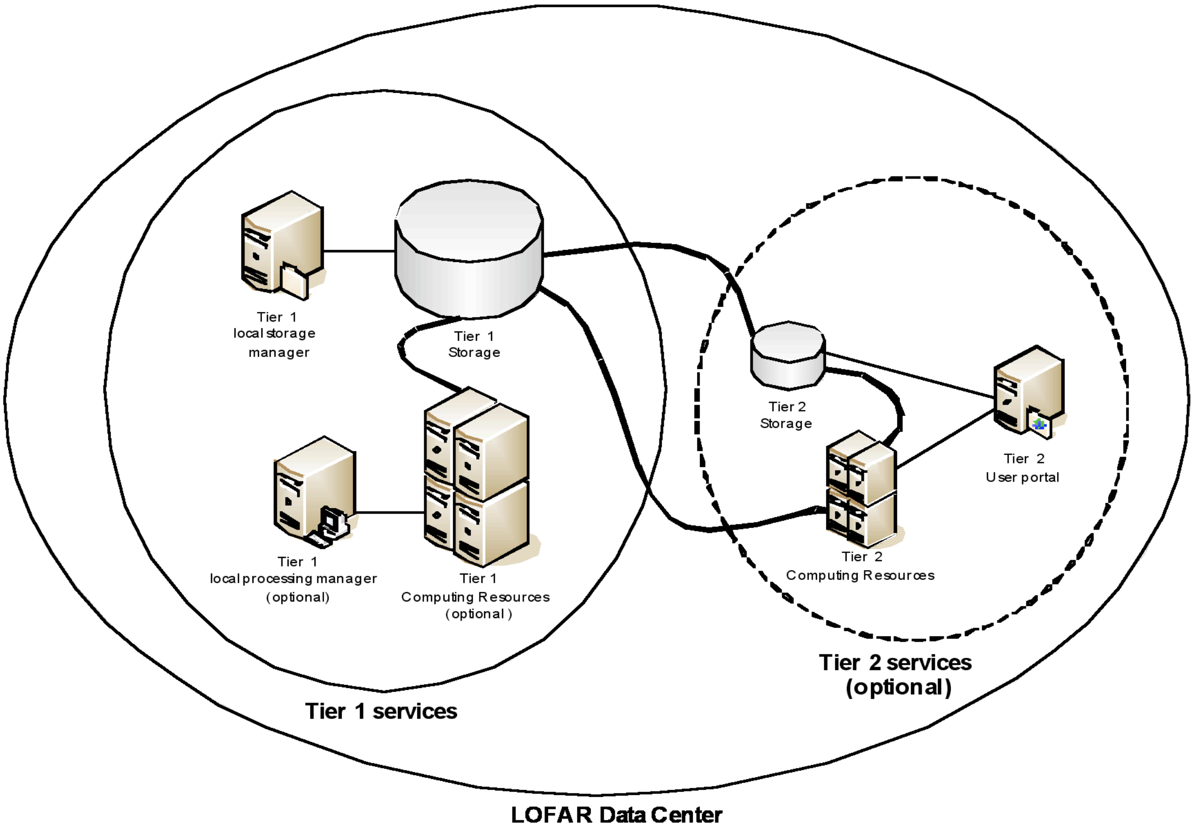


Figure 4: LOFAR Tier 1 (& optionally also Tier 2)

Appendix A: Storage Requirements Calculation

For an idea of what the various contributions are from what data product, a table to calculate the total amount of storage space required.

Total Data to Store		
Total potential Obs time/days	360	
% of actual scientific obs. time	60	
Hours to observe	5184	
observation unit/h	4	should coincide with obs duration
no. of beams	8	should be for 8 for 4 MHz standard obs
Total obs units	10368	

Core

obs in %	45		obs in %	total obs units		requ in %	total obs units	Data Volume	
Total obs units	4665.6	LBA	50	2332.8	UV-visibilities	30	699.84	15479.34106	
					Images	60	1399.68	150289.4956	
					Antenna Time Ser.	10	233.28	2799.36	
		HBA	50	2332.8	UV-visibilities	30	699.84	63207.30931	
					Images	60	1399.68	150289.4956	
					Antenna Time Ser.	10	233.28	5598.72	
							Total	387.6637216	TB

NL

obs in %	45		obs in %	total obs units		requ in %	total obs units	Data Volume	
Total obs units	4665.6	LBA	50	2332.8	UV-visibilities	30	699.84	175575.8592	
					Images	60	1399.68	150289.4956	
					Antenna Time Ser.	10	233.28	2799.36	
		HBA	50	2332.8	UV-visibilities	30	699.84	397733.0688	
					Images	60	1399.68	150289.4956	
					Antenna Time Ser.	10	233.28	2799.36	
							Total	879.4866392	TB

E-LOFAR

obs in %	10	total obs units		requ in %	total obs units	Data Volume	
Total obs units	1036.8	1036.8	UV-visibilities	20	207.36	909650.1658	
			Images	80	829.44	89060.44185	
			Antenna Time Ser.	0	0	0	
					Total	998.7106076	TB
					Total	2.265860968	PB

Appendix B: Full storage of UV-visibilities

The classic radio observatory case is that the raw data are archived and the observer re-processes the data and generates an image. In the following an estimate is derived for the required storage space of such a scenario.

In this approach, frequency- and time-averaged, calibrated visibilities (UV data) as exported by LOFAR are archived continuously. As mentioned one can observe using the core, the full or E-LOFAR with 8 beams each 4MHz. At a later stage there might be the possibility to observe using 4x8 beams.

If all the LOFAR (core and NL) uv-data that are continuously produced by the correlator are archived (in full time and frequency resolution) this will amount to about 30 to 40 Pbyte of data per year. This number is rather large and will be too expensive. However, if it is possible to average the data in time and frequency in a sensible way then the scenario of archiving UV visibilities might become attractive again.
Frequency averaging will be dependent on the observing frequency and the longest baselines to avoid bandwidth smearing. To avoid bandwidth smearing the fractional bandwidth must be smaller than the ratio of the station size (65m) to the longest baseline (100km for the full array, 3km for the core). This is most demanding at the lowest frequencies and at an observing frequency of 30MHz equates to channel widths of <650 kHz for the core and <20 kHz for the full array. Rounding to power of two gives 512 kHz channels for the core and 8 kHz for the full array.

To avoid time smearing, the angular shift caused by the Earth's rotation should be less than the ratio of station size to longest baseline within on integration. That means integration times below 9s for the full array and below 300s for the core.

However these numbers might be too optimistic and does not take into account that LOFAR is actually a full sky telescope and strong sources even though far away from the phase centre need to be accounted for. These sources (the A-Team) could be subtracted from the raw visibilities and then averaged for archiving. To assume somewhat more realistic numbers, in the following it is assumed that a 5s-5kHz averaging for the core and a 3s-3kHz averaging can be done for the full array.

For the available actual observation time, a percentage of 60% is assumed. This reflects that night time will be almost always observed but at daytime LOFAR might not be able to observe all the time due to a difficult ionosphere. Furthermore, one should allow for downtime and maintenance of the telescope.

Another point to consider is what will actually be transferred to the storage facility. At the moment, the best guess is that the raw visibilities with some calibration corrections for various directions should be archived. This seems the most effective way and it does not increase by multiplication of the original data but the actual archived data will be somewhat higher.

In summary if LOFAR runs continuously and observes 60% of the available time (50%HBA, 50%LBA and 45%core, 45%NL and 10% E-LOFAR), core data are stored on a 5s-5kHz basis and full array data are stored on a 3s-3kHz basis and one allows for a 10% overhead, a data volume of 2.5 Petabyte/year is produced. In addition, 1 Petabyte of data is produced for the E-LOFAR observations. On the contrary, if one is able to archive core data on 10s-10kHz basis and full array on 5s-5kHz basis, then the required amount of storage decreases to 850 TByte/year with an additional Petabyte/year for E-LOFAR.

Appendix C: Timelines

LOFAR CS1 has produced about 10TB of data over the last year, 2007. It is expected that this rate will start to increase by the middle of this year when roll-out starts and station will be added.

It is expected that by the end of the year 2008, 20 stations will be operational. This will be:

32000MHz x (20x19/2)stations x 4crosscor x 2pol x 4bytes x 4hours =   46Gbytes (for integration t_{int}=1min)

This amount will be produced by 20 LOFAR stations observing at full spectral resolution and a time resolution of 1min integration time. If one runs this LOFAR with 20 stations for half a year for 60% of available observation time then this amounts to 30TB of data. The first half year of archiving CS1 data should be about 5TB given past experience. That number should be seen as an upper limit as stations will be added to LOFAR CS1 from summer on the data rate will increase and the data rate given above will only be achieved at the end of 2008 but the HBA data output rate will be higher than the LBA data output rate.

In 2009, LOFAR will be completed which means that another 20 to 30 stations will be added. As commissioning proceeds in 2009 it is likely that uv-visibilities will be stored. The data rate will also gradually increase and might reach 184Gbytes for a 4 hour observation with full spectral resolution (assuming integration times of 1min). This might increase for the HBA observations since there are technically more HBA stations. Also integration times will decrease as the instrument is brought to its full capabilities increasing the data output rate. In addition observation with E-LOFAR stations will add to the data. From that the estimated required storage capacity is about 300TByte for the year 2009.

In 2010, the roll-out of LOFAR is completed. That does not necessarily imply that the LOFAR will be fully operational. Commissioning will be continued and only by the end of the year it can be expected that LOFAR is fully operational.

Year	Storage requirement	LOFAR phase
2007	10 TByte	CS1 completed
2008	35 TByte	LOFAR 20 completed 1-2 E_LOFAR stations
2009	300 TByte	LOFAR 40 completed more E-LOFAR stations
2010	800 TByte	LOFAR completed
2011	1-1.5 Petabyte	LOFAR fully operational

Appendix D: Use cases

LOFAR Archive Use Case (guest observer)

Overview

This use case describes the steps a typical user might take in order to make use of data stored in the LOFAR archive. As described here, a remote user might utilize the archive to identify existing archived LOFAR observations, examine the standard data products and the processing used to produce them, and finally initiate a reprocessing of the data using different criteria.

Identify Existing Observations

(a) Open archive query tool (web-based or otherwise)

(b) Type in common name of object (or coordinates)

(c) Resolve object name using SImbad query

(d) Search archive for available LOFAR datasets within specified radius

(e) Receive summary list of observations which match criteria

(f) Select desired datasets from list for further examination

A minimum amount of information about each dataset should be available in the summary list. These might include a list of files associated with the observation including standard data products as well as associated calibration files, system health metadata, etc. If image cubes are available then snapshot images of the entire field and center of the field should be present. Some general summary information of any processing problems should also be available as well as an indication of whether the dataset is public yet. If not yet public, the date when the data will be public should be given.

Examine Standard Data Products

(a) Select desired dataset(s) from summary list

(b) View list of data products associated with the observation

(c) Select subset of available data products for examination and/or download

(d) Display datacube

(e) Page through frequency channels in the image cube

(f) Read off coordinates for sources in the image plane

(g) Examine flux intensity at points in a selected frequency plane

(h) Extract and display spectrum for a given spatial location

(i) Examine polarization information for positions in the image

The actions described here refer specifically to a standard image cube consisting of axes such as ra, dec, frequency, and polarization. For a given observation, other related data products might also be available including source lists, beam-formed data files, or time series data. In some cases, the uv data may also be available. All available data products should be summarized and available for inspection.

The user interaction with the data products described here might be accomplished using off-line analysis tools. However for larger data sizes, such off-line analysis may not be feasible in which case web or network-based tools will be needed.

Examine Data Quality

(a) Select desired data product from summary list

(b) Select data quality report on datacube

(c) View list of data quality checks performed and results

(d) View cube statistics such as mean intensity, rms , etc.

(e) Display associated error cube

(f) Read RMS values from positions in the error cube

(g) Examine histograms of values in entire error cube

(h) Examine histograms of values in selected spatial regions

Examine Processing Logs

(a) Select processing status for given observation

(b) View list of processing pipelines which were run (may be more than one)

(c) Select a given processing chain

(d) Examine processing logfile for entire pipeline and associated products

(e) View list of processing modules used in pipeline

(f) Examine parameter file used by given module

(g) Examine processing log for given module

Examine System and Calibration Information

(a) Select desired data product from summary list

(b) View list of associated calibration data files

(c) Select self-cal solutions from among list

(d) Display phase and gain solutions for observation

(e) Display chi-squared values for solution as a function of time

(f) Select time-range based on chi-square limits

(g) Apply time filter to solutions and display

(h) Apply time filter to ionospheric solution and display

(i) Apply time filter to system temperature data and display

This description assumes that a set of calibration related products associated with the given observation have also been archived. At a minimum, users will likely wish to examine the output of the BBS self-calibration step. Additional information about the health and status of LOFAR during the time of the observations will also be available. Information such as system temperature or observing conditions like rain, may be available in the form of databases which can be queried as necessary by users on the fly as in the example.

Submit Reprocessing Request

(a) Select desired observation from summary list

(b) Select raw data (if available)

(c) Select option to submit processing request

(d) View list of possible pre-defined processing requests

(e) Select standard imaging pipeline reprocessing

(f) View default modules in processing pipeline

(g) Select BBS module

(h) Edit module parameters

(i) Select flagging module

(j) Edit module parameters

(k) Submit reprocessing request

(l) View status of processing queue

(m) Receive email notification of completed reprocessing

(n) Repeat from Step 3 above

This step assumes that processing chains may be initiated by users. The types and amount of such processing requests available to a given user will be a function of their access privileges. For example a guest observer might have privileges which allow him to initiate a reprocessing request for his own proprietary dataset, but not other datasets and only within some allotted time frame. Limits on the availability of processing resources will similarly need to take into account user access privileges. Archive scientists or LOFAR calibration scientists will necessarily have higher access privileges.

We have also assumed that certain standard processing chains will be available. These might include the default pipelines used to create the data products in the first place as well as an array of basic scientific analyses such as extracting and fitting spectra, source detection and characterization, or pulse detection and parameterization in the case of time series data.

Use case for the LOFAR Operators

LOFAR Observatory Science Support and Portal

Observatory Science support staff are available for advice to all users at many stages of their project. There is also one main online entry point for all LOFAR related information and (central) processing. An incomplete list of functions provided via the portal:

- Project proposal preparation
- Observation setup preparation
- Observation (re-) processing setup preparation
- Observation inspection
- Project administration
- Data reduction & analysis
- Dataset location retrieval
- Publication of the LOFAR Global Sky Model catalogue
- Archive queries including reprocessing requests

Dedicated project portals

Groups and organizations may set up their own portals for users that wish to access project specific resources. In general these are set up to publish project related information and data products (reduced data, catalogues, etc.). Full metadata and references to datasets of the associated observations may be provided as well. If applicable, such a Tier 2 data centre may provide (re-) processing resources that can be accessed through the portal or through some other route.

Operations Use Case 1: Primary observation data stream

The primary observation data stream is driven by the need to offload data from the temporary (Tier 0) storage facilities. It is triggered by LOFAR operations, either automatically as a scheduled transfer or manually by an operator. The process requires the following steps:

1. Check available storage capacity and supported data policies at Tier 1 site
    a. If insufficient or not appropriate storage, try alternative site
    b. If non available abort (later retry or discarding of data possible)
2. Initialize Tier 1 storage (apply authorization)
3. Initiate transfer of observation data
4. Check transferred data
5. Register data in central metadata archive (initiates metadata stream)
6. Update LOFAR project database

Operations Use Case 2: Metadata stream

The metadata stream provides the functionality to store and retrieve metadata (data about data) associated with LOFAR observations. There is a single central metadata archive that stores all metadata associated with LOFAR Registered Data on Tier 1 sites. It also provides the registry that stores the references to Registered Data. The metadata archive is linked to the LOFAR project database. The association of Registered Data with a project, and thus to project members, determines the access rules that apply to the Registered Data. For Tier 1 sites, this link should be transparent (i.e. Tier 1 sites are managed by the Data registry alone and require no knowledge about the project database).

The process taking care of the metadata stream is as follows:

1. Initiated by “Register new data” request
    a. In general by primary data stream process “Register data”
    b. Optionally from Tier 1site (e.g. corrections, reduced data)
    c. Request includes:
        i. Observation ID
        ii. Tier 1 reference
        iii. Metadata (from SAS database; optional)
2. Register data
    a. Store Observation ID
    b. Validate Tier 1 reference
    c. Store Tier 1 reference
3. Retrieve metadata
    a. From provided metadata (if available)
    b. From SAS database (using Observation ID)
    c. From observation data (if metadata incomplete)
        i. In Tier 0 storage (using Observation ID; optional)
        ii. In Tier 1 storage (optional)
4. Validate metadata integrity & completeness
5. Store metadata
6. Update project database: Observation ID “archived”
7. On failure (steps 2 – 7)
    a. Roll back
    b. Return “failed” & Reason
8. (Optional) Tier 2 sites may subscribe to the LOFAR Data registry. Triggers will than be provided from the LOFAR Data registry to the Tier 2 site when data for certain projects is registered or updated. The Tier 2 site may use the trigger to import metadata stored in the LOFAR Data registry.

Interfaces for extraction of metadata from observation data are optional and only required if metadata stored in the SAS database is incomplete. Note that metadata will be stored together with the observation data at Tier 1 as well.

Operations Use Case 3: Data replication & relocation streams

TBD

Science Uses Cases

Error! Reference source not found. shows the main (generic) science use case. It considers Tier 1 access only. Tier 2 may access Tier 1 in a similar way. User interaction with Tier 2 left to the Tier 2 site. More detailed specific science cases are described elsewhere.

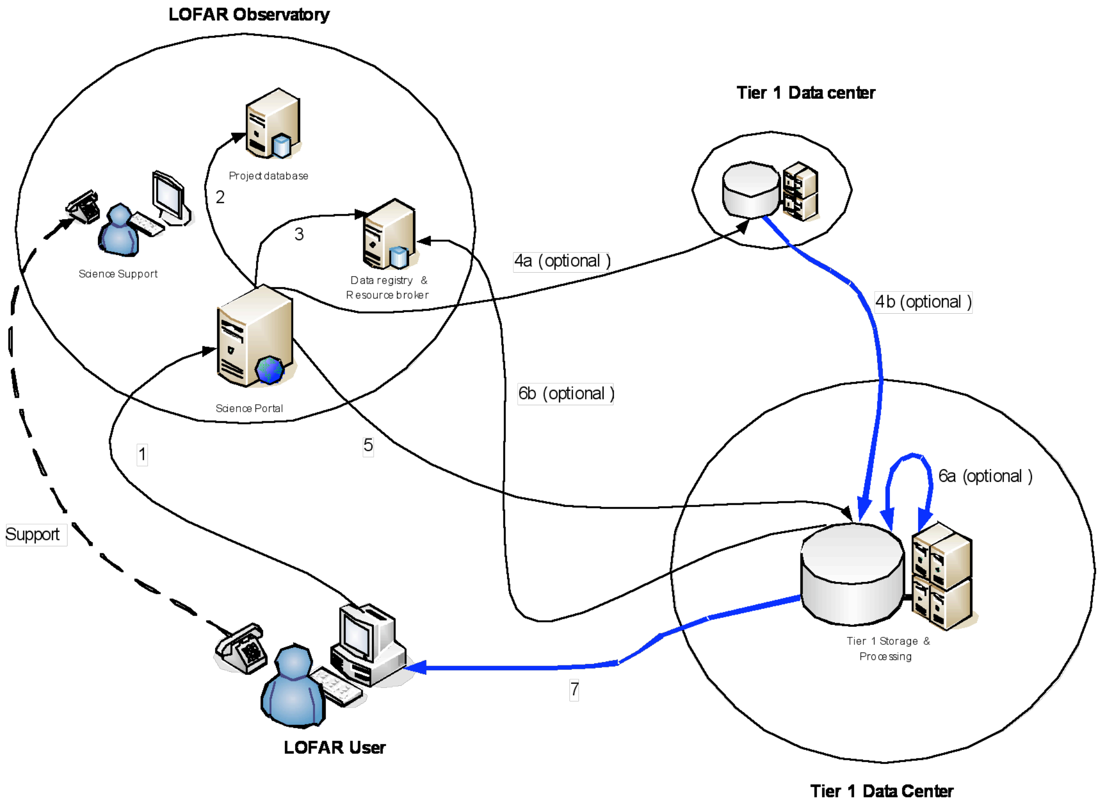


Figure 5: Science use case

1. The user accesses the LOFAR Science Portal
2. The user is authenticated against the LOFAR project database and authorization attributes (e.g. project memberships) are associated with the user. Anonymous access will be supported as well.
3. The user requests access to data, e.g. through a query or VO service. The data registry is queried for references to the Tier 1 storage locations.
4. (Optional) If data is stored at a location where no proper processing resources are available, data relocation may be requested. (Note that data relocation involves interaction with the LOFAR Data registry & Resource broker. This is implicit to the Tier 1 services and not shown explicitly)
    a. Through the portal, a relocation request is sent to the Tier 1 site where the data is stored
    b. The data is sent to the desired destination Tier 1 storage facility.
5. The user is redirected to the Tier 1 site where the data is stored.
6. (Optional) Data (re-) processing (including analysis & visualisation)
    a. User initiated data (re-) processing may be carried out at Tier 1. The stored data is accessed by the processing system. Stored Registered data can not be overwritten so read-only access is provided. The process may produce data that is to be stored at Tier 1, in which case the resulting data will have to be registered. Alternatively, resulting data is stored in a workspace and removed

once the User has retrieved it. A final alternative is that the resulting data is stored in Tier 2 storage.

b. Resource usage is updated by the Resource Broker.

7. The user retrieves data from the Tier 1 storage and stores it locally. Alternatively, data is stored in Tier 2 storage.

Use case for the LOFAR Transient KSP

Data base mining

One of the final data products of the Transient KSP which is stored in an archive is a data base of light curves containing infos on variability, fluxes and class of transient objects. These databases should be searchable and information retrievable:

(a) Open archive/data base query tool (web-based or otherwise)

(b) Type in common name of object (or coordinates)

(c) Resolve object name using SImbad query

(d) Search database for available LOFAR datasets within specified radius

(e) Receive summary list of observations which match criteria

(f) Select desired datasets from list for further examination

(g) Request data with specific time resolution from a specific time onwards

A different approach:

(a) Open archive/data base query tool (web-based or otherwise)

(b) Type in a number for a spectral index or flux density

(c) Indicate smaller/greater

(d) Search database for sources having spectral index/flux density smaller or greater than indicated

(e) Receive summary list of sources which match criteria

(f) Select desired datasets from list for further examination

(g) Retrieve and examine light curves (visualise?)

A different search option:

(a) Open archive/data base query tool (web-based or otherwise)

(b) Type in number of variability (in flux), frequency and time scale

(c) Indicate greater or smaller

(d) Search database for sources having the greater or smaller variability at requested frequency and on indicated time scale

(e) Receive summary list of sources which match criteria

(f) Give number of sources that match criteria to total number of sources in database

(g) Select desired datasets from list for further examination

(h) Retrieve and examine light curves (visualise?)

If a trigger is received from other telescopes than it is not only desirable to point the telescope in real time into the direction of the source but also start a query in the archive database for known Transients or sources. This information could be send directly to the scientists/operators who receive the trigger message as well.

(a) Trigger is received by database query tool

(b) Reads in from the trigger the location of source and the error circle

(c) Search Transient database and the survey source catalogue for known sources within error circle of indicated location

- Combine this search with optical databases as well

(d) Receive summary list of sources from databases which match criteria

(e) If search successful: send source names, location and fluxes to observer/astronomer

(f) If search unsuccessful: send info to astronomer no source found in the requested area

Data Analysis

This has the purpose to look at the data to find something not produced by the standard processing.

A new transient source has been found. The archive is searched to determine whether that is correct or if the source might have been detected before but not within the acceptable detection limit.

(a) Open archive/data base query tool (web-based or otherwise)

(b) Type in location of new transient source

(c) Search Transient database for questionable detection of sources

(d) Receive summary list of sources which match criteria

Here it is possible to either stop the request or try to look at original data. The Transient KSP plans to store also uv – data in a compressed form for possible reprocessing.

(e) Request all available uv-data around the location of new transient

(f) Stack all the data

(g) Run off-line version of BBS/Imaging pipeline on the data cube

(h) Visulise resulting data cube

Use case for the LOFAR EoR KSP

Before proceeding, a number of distinct phases of data-access can be identified. Each have their own mode of data access.

(i) *Standard Image Processing:* During this phase the EoR-KSP is flagged, calibrated and processed (at 1sec, 1Khz level) through the standard LOFAR imaging and calibration pipe-line to a quality level sufficient to allow compression to 10 sec, 10KHz. All data-access modes here will also be required in the next phase.

(j) *Reprocessing phase:* The full 10-10 data-set of ~1 Pb of raw visibilities plus ME parameters, as delivered by LOFAR, has to be stored permanently and accessible at ANY moment in time for improved calibration, imaging and flagging purposes.

1. At this point each visibility for all dipole cross correlations should be accessible as function of (u,v,w) coordinates, frequency, time, telescope/baseline, and all other ME calibration parameters stored together with the 10-10 data-set, and a combination or range thereof. The reason being that the calibration unknowns can be functions of each of these parameters (or combination). This prohibits the data to be split in smaller sub-sets and processed serially and/or independently.

   2. When calibrated and flagged to a sufficient level, the data will be compressed further to 10 sec, 100KHz and stored together with all its ME parameters.

(k) *Signal Processing Phase:* During this phase the 100 Tbyte compressed data-set is initially treated a "perfectly" calibrated, which allows each EoR window, each field within the latter and each frequency channel to be treated independently. The processing will be embarrassingly parallel. Hence all these sub-sets of data can be processed simultaneously and again have to be accessible all the time (at this stage the data will be sorted to improve processing speed). During this phase the end-result will be an considerably compressed data-cube (several GBytes at most), which will be used in the signal extraction phase.

   1. If at this point issues arise with the calibration, one can use the best UV/image data-cube to improve the reprocessing of the data. Hence iteration between (1) and (2) can not be excluded. Iteration between (1) and (0), however, will not be possible due to storage issues.

(l) *Signal Extraction Phase:* This process can be done on workstations and requires little storage capacity (see EoR KP plan).

(m) *Storage of Calibrated UV Data:* After a full calibration, the ME parameters and the 10-10 UV data-set can be stored for other purposes (e.g. time-lapse movies of variable sources, etc). This requires the same data-access as in the reprocessing phase. The relatively low information content in the final data-product, however, would probably mean that most users are satisfied with the final compressed calibrated data-cube (either in UV or image space). The latter would require no specific access modes, since these are standard product of moderate size.

Assumptions: As basic stored data set we assume 1 Pbyte of raw visibilities (10s-10Khz; all dipole cross correlations) including all ME equation parameters.

Each visibilities should be uniquely identifiable through: the phase centre, (u,v,w) coordinates, frequency, time,telescopes, baseline, etc (MORE?). Each visibilities should also have a correction associated with it, calculable from the initial LOFAR calibration/imaging ME.

IMPORTANT: It should be possible to search or create data-subsets based on a combination of ranges of the above parameters (e.g. time and freq. range, baseline and/or telescope, etc) or combinations thereof. Possibly more complex selection functions might be required at later stages.

These data-sets should be sortable in any order as function of the above (incomplete) list of parameters.

To accomplish the above we require:

(g) A tool is required to select subsets of data either by hand or automatically (steered by the reprocessing software) based on the parameters/criteria outline above.

(h) The subset of data can be resorted in different order(s) (time, freq., UV coordinates, etc) if required for different aspects of the reprocessing/analysis.

(i) The selected data-sets can be compressed (in time, freq, UV, space, etc). The compressed data-set can also be visualized.

(j) The data can be inspected by simple statistical tools to assess initial data-quality.

(k) The best ME corrections can be applied to the raw data in any of the above phases (visualization, compression, etc).

(l) The data-set can be delivered to the reprocessing software (via GRID, or otherwise) for calibration, imaging and flagging (see above). Improved reprocessing calibration parameters and sky models can be stored (e.g. in the ME files of the raw UV data; old ME parameters are retained for backup in case of problems).

(m) etc.

1) Select, sort and compress data sub-sets

(i) Open storage tool to access raw UV visibilities, and currently best ME parameters and best global/local sky models.

(j) Select data-subset(s) plus ME pars and sky models for processing based on range of, or combination of, UVW, freq, time, ME pars, etc (see above)

(k) Obtain simple statistics of selected data-set and history of previous data manipulations.

(l) Sort data in any required order if necessary.

(m) Compress data if necessary (e.g. 10-10 to 10-100) with or w/o ME corrections applied and/or with or w/o sky model(s) subtracted

   1. This step might require a BBS core as part of this tools to interpret the ME parameters and apply them before compression.

(n) Transport (sorted, manipulated, compressed) data-sets to reprocessing cluster or select for further inspection.

(o) Keep logs of all processing and allow reversal of every step at all times.

2) Examine Data & Data-quality

(a) Select the data-subset to examine (UV or images).

(b) Select averaging scheme (if necessary).

(c) Apply correction and subtract sky model if necessary.

(d) 1,2 & 3-D visualisation/displaying of (sub-sets) of selected UV data, ME parameters and images against different parameters (e.g. uv, freq, time, line-of-sight, etc).

(e) Infer statistics from selected (residual/corrected) data-sets (e.g. chi^2, likelihood, Bayesian evidence), etc.

(f) Select data-regions based on residual statistics for further analysis. Apply pattern-recognition software to identify areas of low-level RFI or/and bad calibration.

(g) Apply cross-correlations on residual data-cubes to asses level of calibration as function of ME parameters (hence these can be correlation between UV points which are far in UV space, time and frequency, but close in ME-parameter space).

(h) Select regions of data-that require more processing. Store selected regions for input in storage/selection tool.

(i) Keep logs of all processing and allow reversal of every step at all times.

3) Reprocessing Phase:

(a) Select UV data-set to calibrate (from -1- or -2- above).

(b) Select best calibration pars for this data-set (ME pars) and best local/sky model.

(c) Select parameters to be calibrated.

(d) Calibrate parameters using BBS/Meqtree (or other codes), using best local/global sky model, starting from best calibration so far. Allow inspection of results (tools under -2- ?)

(e) Store new calibration parameters.

(f) Apply new calibration to data.

(g) Create improved local/global sky model.

(h) Store new local/global sky model.

(i) Examine residual data-set (see -2-). Allow flagging and selection of sub data-sets for further calibration.

(j) Allow compression of data.

Use case for the LOFAR Cosmic Ray KSP

The CR-KSP data is stored as single events. The archive is thus accessed to retrieve data associated with selected events. The first step of every access is to generate (or reuse) a list of selected events. This selection is usually based on the event-parameters stored with each event, from the trivial case of selecting all events, via selecting events according to data taking parameters (like timestamp, number of triggered stations, ID of triggered stations etc.) to parameters reconstructed with the processing pipeline (like arrival direction, reconstructed energy etc.).

1) Retrieve event list → also offer the possibility to use a previously created selection

(a) Choose fields on the sky for event selection (from CR event list, from meteorological database, from system health database, etc.)

(b) Select events (Make cuts on the selected fields.)

(c) Display selected events; this should include both a simple list as well as available data products previously created

(d) Repeat from step “Event selection” or

(e) Export event-parameters for selected events as

- ASCII-table (csv)
- Root-Table (native format of the program “Root”)
- SQL-Database?
- HDF5 dataset (would enable more complex collection of data, but format would need to be defined)

2) Retrieve selected events

(a) Generate event list (see use case 1)

(b) Download raw data to local disk

(c) Download additional available data products (optional)

3) Retrieve image cubes

(a) Generate event list (see use case 1)

(b) Set pipeline parameters (incl. version of pipeline) → this actually might require retrieving and building a specific version

(c) Run pipeline with selected parameters to run image cubes

(d) Download image cubes to local disk

4) Update selected events (restricted to selected users, e.g. KSP members)

(a) Generate event list (see use case 1)

(b) Set pipeline parameters (incl. version of pipeline)

(c) Run pipeline with selected parameters (only on events not processed in the same way already)

(d) Add results to CR event list

(e) Export updated event list

5) Database Update: (restricted to selected users, e.g. KSP members)

Like “Update of selected events” but done on all events after a new version of the pipeline becomes available.

Use Case for the LOFAR Solar KSP

Solar observations, including those with LOFAR, are usually addressed by the observation time. This is both true for studies of the long-term development of e.g. solar active regions, as well as studies of certain flare or CME events. The latter are looked up in event lists, and the primary search criterion is the time of the event. Information on the observing mode, e.g. routine monitoring or a joint campaign with other ground- and space-based instruments, might also be relevant. Further information, like the number of the active region that caused a flare, is less important since LOFAR provides full-Sun imaging.

So archive users looking for solar data will perform the following steps:

1. Data retrieval

(a) Open archive/data base query tool (web-based or otherwise)

(b) Search database for available LOFAR datasets within the desired time interval. Some users would like to limit the search on certain observation modes, e.g. joint observation campaigns

(c) Receive summary list of observations within this interval, including information on frequencies, image cadence, and the observation mode used: routine monitoring, solar radio burst mode, observation campaign

(d) Refine search with further criteria: Low/High Band, user-specified frequency interval, image or uv datasets, observation cadence, ...

(e) Select desired datasets from list for further examination

(f) Request data

2. Data analysis

A user who wants to combine LOFAR data with other observations, e.g. X-ray or radiospectrograph data, will be interested in locating the source of long-wave radio emission on the Sun. This requires both ionospheric corrections, i.e. standard calibration, and corrections for diffraction and scattering of the radio waves in the solar corona. The latter requires information that is not part of the LOFAR data processing pipeline, e.g. a solar coronal density model. This part usually will be done on the user's local workstation.

Data reprocessing steps relevant for the archive are:

(a) apply improved calibration sky model on raw uv visibilities, if such data are available

(b) if calibrated uv visibilities are available and images desired, run standard image processing pipeline, possibly with solar extensions

(c) if images are available, or calibrated uv datasets are desired by the user, download the data