LOFAR data sets for imaging pipeline tests

Stefan J. Wijnholds

March 1, 2021

Introduction

In the development of imaging pipelines for SKA, its SRCs and LOFAR 2.0, various approaches to direction-dependent calibration, gridding and deconvolution are being considered. These alternatives need to be tested to assess their merits in terms of scientific quality of their output and computational performance. Initially, such assessments can be made using simulated data. However, simulated data will always be a simplification of reality as only anticipated issues will be included in the underlying simulations. This makes it desirable to have a few data sets available for testing of imaging pipelines, each addressing specific challenges. In the context of SKA-LOW and LOFAR, such challenges could be varying ionospheric conditions or source structures with varying complexity. Given the architectural similarities of LOFAR and SKA-LOW, LOFAR data sets are potentially highly suitable for this purpose. The aim of this document is to propose a set if boundary conditions that need to be satisfied to make a data set suitable for imaging pipeline testing, and a sketch of the steps needed to create / obtain suitable LOFAR test data sets, so that a starting point is created for a, hopefully short / focused, discussion amongst a number of experts involved in the development of imaging pipelines for SKA, its SRCs and LOFAR 2.0. These discussions should result in an updated version of the document defining the initial test data sets.

Constraints

- Maximum size of 10 GByte in MS format
  Rationale: this constraint is chosen to be the same as for the simulated test data sets. Meeting this limit will require a significant reduction in data volume for a typical LOFAR data set. For example, according to the LOFAR data size and processing time calculator[1], a typical HBA observation with the LOFAR-NL array with 48 MHz bandwidth at 3 kHz resolution will produce 1855 GByte per hour of observing when data is stored at the correlator dump rate (1 s). Based on experience in the LOFAR EoR project, the time and frequency resolution can be brought down to 10 s and ~60 kHz after RFI flagging and direction-independent calibration at full data resolution. This reduces the data volume to 8.7 GByte per hour of observing.
- Frequency coverage: ideally the full available frequency range.
  Rationale: one of the main calibration challenge at low frequencies is ionospheric calibration. A wider frequency band is better to disentangle ionospheric effects from other direction-(in)dependent effects. The frequency coverage need not be consecutive. The frequency coverage may not only be limited by the passband of the analog filters, but also by the presence of RFI. For example, observing below 30 MHz at daytime is nearly impossible.
- Time coverage: 4 to 6 hours
  Rationale: 4 hours may already be sufficient for a test observation. Much shorter observing times are not recommended due to the poor (u, v)-coverage on the long baselines. If the purpose of a specific test set is to assess imaging performance of diffuse emission, however, a significantly shorter observing time may be feasible as most of the actual imaging will likely be done with only the core.
- Availability of “ground truth”
  Rationale: this is needed as reference for the scientific output produced by the imaging pipeline being tested. Assuming that the data volume reduction required for the test data set (see above) is achieved by decimating in time and frequency instead of aggressive averaging, reduction of the full original data will produce a deeper image than reduction of the test data set and can therefore, in principle, produce this reference. This reference can either be an image (cube) or a source model (list of extracted source components).

[1] https://lofar.astron.nl/service/pages/storageCalculator/calculate.jsp

Other considerations

- Permissions from PI if for use of intermediate data products.
  The LOFAR long-term archive contains a lot of data that is past the proprietary period and can thus be requested freely. However, one has to process these data oneself to obtain, e.g., the desired reduction in raw data volume or a sky model that can be considered as the ground truth for that data set. This can be a significant amount of work, so it may be attractive to ask PIs for permission to use their intermediate products or extracted sky model.
- Hybrid data set
  For some tests, it may be attractive to add one or more simulated test sources to the data at appropriate flux levels and locations.
- Multiple data sets
  It is likely that multiple data sets are required to cover various scenarios, for example mild ionospheric conditions versus severe ionospheric conditions or different types of source structures.

Required work

To produce similar data products as defined for the simulation tests[2], the following steps need to be taken

1. Once a suitable observation is selected, the raw data from this observation needs to be pre-processed, including direction-independent calibration and RFI flagging, so that averaging in time and frequency can be done without loss of signal quality. The output of this step can be stored in MS format. As this is a standardized step for both the LOFAR surveys Key Science Project (KSP) and the LOFAR EoR KSP, such data may be readily available if permission from the respective PIs can be obtained.
2. The MS produced at the end of the previous step needs to be decimated further in time and frequency to produce a MS for pipeline testing with a size smaller than 10 GByte. This decimation requires care to ensure that the impact on the (u,v) coverage of the observation is minimized. The output of this step should be stored in MS format.
3. The MS produced at the end of step 1 needs to be imaged with one of the current pipelines to produce a reference image or source component catalogue that can be used to assess the scientific performance of the proposed workflows. As this is done using the full data set (as opposed to the decimated data set produced in step 2), the images or source component catalogue produced should be able to reach a lower flux limit than will be achievable with the MS produced in step 2.
4. The instrument model for LOFAR can be made available via the EveryBeam package.

[2] https://confluence.skatelescope.org/display/SE/Testing+WG