LWA Data Management
LWA Memo #177
Ver. 4

Greg Taylor*
Jayce Dowell
Ylva Pihlstrom
Joe Craig
(University of New Mexico)

January 3, 2019

Contents

*E-mail: gbtaylor@unm.edu

1 Introduction

The first LWA station, LWA1, has been in operation for a several years. In addition, the recently completed second station located on the Sevilleta National Wildlife Refuge, LWA-SV, increases the capabilities available to observers. In particular it opens up the possibility of interferometery between the two stations to decrease the confusion level. This memo describes the mechanisms by which users of the LWA1, LWA-SV, and the LWA interferometer will retrieve their data for subsequent analysis. While there has been some discussion of how to move data from stations to the correlator [3, 4], prior to this memo nothing has been written about getting data to users after it is recorded at the site. Given that there is a wealth of science that we can do with these three instruments we need to establish the mechanism by which users can collect their data.

Since the LWA stations are capable of capturing large amounts of data, users will have to select between a limited number of options for data retrieval. Furthermore, since the operational budget for the LWA is constrained, some of the choices may involve modest effort and expense on the part of the user. By way of examples in this memo we assume that users are interested in the widest bandwidth beamformed DRX output at LWA1 which produces 602.8 Mbps (19.6 MSPS, 4 bit I + 4 bit Q, 2 tunings, 2 polarizations), the TBN mode which has a slightly higher data rate, 812.1 Mbps (100 KSPS, 8 bit I + 8 bit Q, 2 polarizations), at maximum, or the LWA single baseline interferometer. We note that at LWA-SV the maximum data rate, 301.8 Mbps, is half of the maximum rate at LWA1 due to the maximum sample rate of LWA-SV being 9.8 MSPS.

We also describe some options related to the LWA User Computing Facility (hereafter UCF) which is located at the site. Current wisdom suggests that it is generally more efficient to move the computing facilities close to the data rather than moving the data to the computing facilities. Also, not every LWA user may have facilities capable of handling the 20+ TB of data produced by a typical observing program.

We furthermore assume that data can be physically collected from the sites at most only once every one or two months by the UNM project office. This may change but is likely to be the case indefinitely.

2 Spectrometer Mode

Most users employing beams in single station experiments should be satisfied with the spectrometer observing mode which provides either linear XX and YY or Stokes I, Q, U, and V for 1024 channels at a time resolution of 40 milliseconds with the widest bandwidth mode (filter 7, 19.6 MSPS; 80 ms at LWA-SV with filter 6, 9.8 MSPS). Some trade-off between integration time and number of channels is possible up to a maximum of 2048 channels at 160 millisec and down to 256 channels at 10 millisec (see Table 1 for LWA1 and Table 2 for LWA-SV options). The spectral resolution and integration time can be determined from:

$$\text{channel width} = \frac{\text{sample rate}}{\text{transform length}}$$

and

$$\text{integration time} = \frac{\text{transform length} \times \text{integration count}}{\text{sample rate}},$$

where the "sample rate" is in samples/second, the "channel width" is in Hz, the "transform length" is the number of channels, and "integration count" is an integer. By way of example consider the widest DRX bandwidth sample rate at LWA1 (19.6 Msamples/sec; filter code 7), an "integration count" of 768, and 1024 channels gives a channel width of 19.140 KHz and an integration time of

Table 1. Integration Times in Seconds for DRX Filter #7

Channels	Integration Count				
	384	768	1536	3072	6144
64	0.001	0.003	0.005	0.010	0.020
128	0.003	0.005	0.010	0.020	0.040
256	0.005	0.010	0.020	0.040	0.080
512	0.010	0.020	0.040	0.080	0.160
1024	0.020	0.040	0.080	0.160	0.321
2048	0.040	0.080	0.160	0.321	No
4096	0.080	0.160	0.321	No	No

Table 2. Integration Times in Seconds for DRX Filter #6

Channels	Integration Count				
	384	768	1536	3072	6144
64	0.003	0.005	0.010	0.020	0.040
128	0.005	0.010	0.020	0.040	0.080
256	0.010	0.020	0.040	0.080	0.160
512	0.020	0.040	0.080	0.160	0.321
1024	0.040	0.080	0.160	0.321	0.642
2048	0.080	0.160	0.321	0.642	1.284
4096	0.160	0.321	0.642	No	No

40 milliseconds. An interactive table of all available spectrometer modes can be found at `http://www.phys.unm.edu/~lwa/astro/scheds/spec.html`. The file size for this typical spectrometer observation (1024 channels, 40 millisecond integrations) for 1 hour is 2.74 GB at a data rate of 6.2 Mbps. All such observations will be converted to HDF5 format, archived, and available for download from `https://lda10g.alliance.unm.edu/ldadb/`.

In the case of spectrometer data, no data management plan is necessary in the proposal.

3 The Raw Data Option

For all TBW/TBN/TBF observers and for observers who require no averaging of the DRX beam output in time or frequency (e.g., pulsar observations, certain solar observations) it will be possible to request the raw data. Justification for this choice will be required in the observing proposal. Furthermore all observations planned using the raw data option will require a data management plan that specifically describes how the user will cope with the large data volumes generated. The specific issues that should be addressed are:

1. an estimate of the total data volume,
2. where the user plans to reduce the data, i.e., UCF, UNM, or home institute, and
3. details of how the data will be reduced. Specifically:
   (a) For the UCF/UNM: how much time they need to crunch the data and retrieve the results; or
   (b) If at their home institute: how they plan to get the data back to their home institution, e.g. how many drives they will need.

By way of example there are a couple possibilities including:

(A) Reduce the data at the User Computing Facility. The UCF (see [2]) is a cluster of 6 hexacore nodes each equipped with two to three CUDA compatible GPUs. The UCF is located in the old VLA correlator room within the VLA control building[1]. The cluster is connected to LWA1 by a dedicated fiber so that it is possible for the LWA operators to copy raw data soon after the observations to the cluster. Data may also be copied automatically using the "auto-copy" feature if specified in the session definition files (SDFs) using the `sessionGUI.py` script. The cluster currenlty has a common storage area of 138 TB, corresponding to roughly 380 beam-hours, although users are generally limited to 10 TB through the file system's quotas. Raw data will be deposited to the observer's directory in `/data/local/recent_data` by the operator. Users should process the raw data from this location and schedule the files for deletion with the `rmrd` utility available on each cluster node. Each node has roughly 6 TB of temporary local storage so observer's should manage their reduced data accordingly (see [2] for the quotas and file age limits currently enforced on the UCF). **In general the UCF time will be granted immediately following the observations.** If there is a time-window during which the user cannot make use of the cluster in a timely fashion then scheduling constraints for the observations should be provided. No raw data will be archived by the LWA.

(B) Ask for data to be processed at UNM. A limited amount of data can be brought back from the site and processed at UNM. This will require a UNM collaborator and proper coordination to be described in the proposal.

(C) Ask for data to be shipped to the user. Valid options for storage media may be either (1) 3 TB or larger bare drives formatted as ext4 or (2) by transferring raw data over the internet (requires a very good, Gigabit or better, connection). For CFP7 and the duration of the MSIP there are limited funds set aside for the purchase of external drives so users need not supply these.

4 Interferometric Data

CFP7 introduces a new mode for LWA observers to use: a single, ∼70 km baseline interferometer formed using the LWA1 and LWA-SV stations. This new interferometric mode allows observers to increase the sensitivity of observations by combining the collecting area of both stations while also reducing the confusion level with the interferometer's fan beam. Although the final data rate for this mode is relatively low at ≈175 MB per hour per beam for 1 sec integration, 256 channels, and two 9.8 MHz tunings, observers should be aware that there is a hidden data cost since all of the observations have to be recorded in raw data mode at both stations and then correlated off-line. Using the observing procedure defined in [1] this amounts about 530 GB of raw data per wall clock hour which needs to be correlated. It is recommended that observers use this mode for blocks of no more than 18 wall clock hours at a time and then wait for the data to be correlated by the LWA staff before scheduling any additional observations. This will help ensure that there is sufficient space available on the cluster for correlation and give observers the opportunity to inspect their data and assess their strategy before continuing observations.

5 MCS Metadata

MCS-generated metadata for all projects will be deposited at a location from which they can readily retrieved. This can either be done by accessing the LWAdb at `https://lwalab.phys.unm.edu/lwadb/` or directly through the LWA archive at `http://lda10g.alliance.unm.edu/metadata/`.

[1]The availability of the UCF is thanks to the generosity of NRAO for hosting it, and JPL, VT and UNM for supplying the nodes

6 Proprietary Period

All LWA users will have a one-year proprietary period, after which data in the LWA archive will be made available to anyone requesting it. Data obtained from the archive after the proprietary period has passed will not be subject to the LWA publication policy (see `http://www.phys.unm.edu/~lwa/publ.html`) except for the requirement that they include the following statement in their acknowledgements:

> Construction of the LWA has been supported by the Office of Naval Research under Contract N00014-07-C-0147 and by the AFOSR. Support for operations and continuing development of the LWA1 is provided by the Air Force Research Laboratory and the National Science Foundation under grants AST-1835400 and AGS1708855.

The current archive distribution scheme automatically enforces the proprietary period of files. Observers wanting to access their files via the archive will need to request a project access key from the archive administrator, Jayce Dowell, via e-mail.

7 Summary

The recent availability of a spectrometer mode with good spectral resolution should dramatically reduce the data volume for single station LWA observations. For those still wanting access to the raw data then this will be possible but users will need to describe how they intend to cope with the large data volume in a data management plan . Some possibilities include, but are not limited to, use of the LWA Users Computing Facility, and data shipment. We have described methods by which we will manage this activity that will require some action from LWA users. These policies may need adjustment as we learn exactly what data rates can be supported by the stations, the UCF, and the UNM network.

This plan describes LWA data management activities while the LWA is supported under the Mid-Scale Innovations Program which is expected to extend through August 31, 2021. Depending on the availability of funding, these services may change.

8 Document History

- Version 4 (Jan 3, 2019): Updated for LWA-SV and the LWA single baseline interferometer. Updated for recent changes to the archive and publication policy.
- Version 3 (Dec 3, 2012): Revision.
- Version 2 (Feb 13, 2011): Revision.
- Version 1 (Feb 11, 2011): First version.

References

[1] Davis, I. & Taylor, G. 2018, "LWA Single Baseline Interferometry Tutorial", `http://www.phys.unm.edu/~lwa/singleB_tutorial.pdf`

[2] Dowell, J. 2018, LWA Memo #193, Ver. 3.0

[3] Taylor, G. B. 2007, LWA Memo #110

[4] Taylor, G. B., & Ray, P. 2008, LWA Memo #131

[5] Wolfe, C., Ellingson, S., & Patterson, C. 2009 LWA MCS memo #0019, `http://www.ece.vt.edu/swe/lwavt/doc/MCS0019.pdf`