

EALFA Memo 040729:
Bookkeeping for ALFA Drift Surveys
rg & mph – 29 July 2004 – revised 11 Aug 2004

Data Unit Definitions:

Survey data for E-ALFA drift scan programs will be collected with the telescope at rest at fixed azimuth as the sky drifts by, each beam collecting data along constant declination strips. This data unit, which includes spectra taken simultaneously for all seven beams over the course of several hours in a given observing session, is referred to as a single **Drift**. The sequence of spectra from a single beam (and thus at a single declination) make up a single **Strip** and 7 strips acquired simultaneously at the telescope make a drift.

For ease of reference, data processing and distribution, we subdivide the Arecibo sky into **Tiles** of 20^{min} in R.A., by 4° in Dec. For a declination coverage of 36° ($+0^\circ < \text{Dec.} < +36^\circ$), the survey will consist of 648 tiles, each of approximately $20 \cos \delta \text{ deg}^{-2}$, where δ is the declination of the tile. They will be assigned to 9 **Bands** in Dec., centered at $+2^\circ, +6^\circ, +10^\circ$, etc. Each tile will be “filled” by a succession of drifts. For ease of data handling, each tile will be subdivided into four **Quadrants** of $t_{\text{Quadrant}} = 10^{min}$, or about 2.5° in R.A., and covering a 2° span in Dec. The tile itself will however be the basic unit of the final regridded public data product.

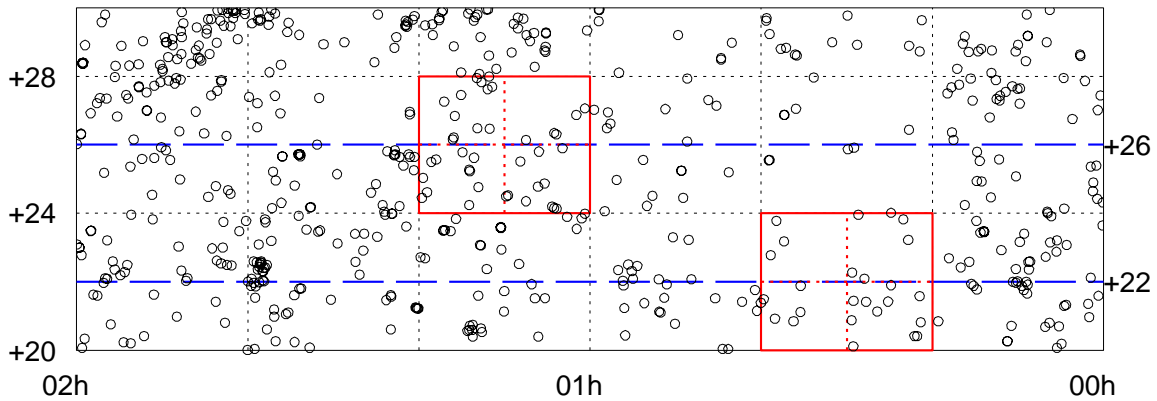


Figure 1: Data configuration for a $2^h \times 30^\circ$ region centered on R.A. = 1^h , Dec. = $+25^\circ$. The dashed lines indicate the central **Drifts** of the $+22^\circ$ and $+26^\circ$ tile **Bands**. The solid red lines outline the **Tiles** centered at $(00^h 30^m, +22^\circ)$ and at $(01^h 10^m, +26^\circ)$, while the dotted red lines delineate the four **Quadrants** that make up those tiles. Circles mark the locations of galaxies with known redshifts $cz < 20000 \text{ km s}^{-1}$.

When ALFA is rotated optimally, the seven beams will sample equally spaced regions on the sky, offset by $126''$ ($2.1'$). Drifts will be spaced in two basic configurations to provide both *areal coverage* and *optimal sampling*:

- Wide step: Successive drifts will space the central beam offset by $14.6'$ to cover separate regions.
- Fine step: Successive passes of the same region will offset drifts by a submultiple of $2.1'$ to achieve denser sampling in the angular domain. As discussed in EALFA memo 040702, a two-pass strategy, with the second drift offset by $1.05'$, would provide better rfi identification, increased signal reliability at low SNR, positional accuracy and characteristic source size information.

For convenience, data will be recorded in raw data units, currently referred to as **Groups** in new AO parlance (what used to be called a 'scan'). Each group consists of N_{rec} data dumps or **Records**. A drift

will thus consist of a sequence of groups, or **Drift Groups**. A Drift Group is composed of 7 **Segments** each corresponding to a different Declination strip swept by an ALFA beam. We may thus refer to such segments also as **Strip Segments**.

The length of a Strip Segment (same as for a Drift Group), $t_{segment}$, is partly dictated by the process of bandpass subtraction. Based on previous experience with single pixel data sets taken in drift mode, it appears that bandpass subtraction will be most efficient on timescales of $\sim 600\text{--}900$ sec. It is also safer to break the data-taking process into relatively short units, in order to minimize the impact of system malfunctions and to fit better into scheduled time blocks. Hence, we propose to record data in continuous ~ 600 sec groups. A record duration of 1 sec is well-matched for dense RA sampling and rfi identification. So a group will consist of 600 records. Cal firing may take place between data groups. If a cal is fired between data groups, taking 2 or 3 sec, the actual duration of the group may be 2 or 3 sec shorter than 600 (Synchronization of group starts may become an issue if the data taking system does not allow starts at sec ticks).

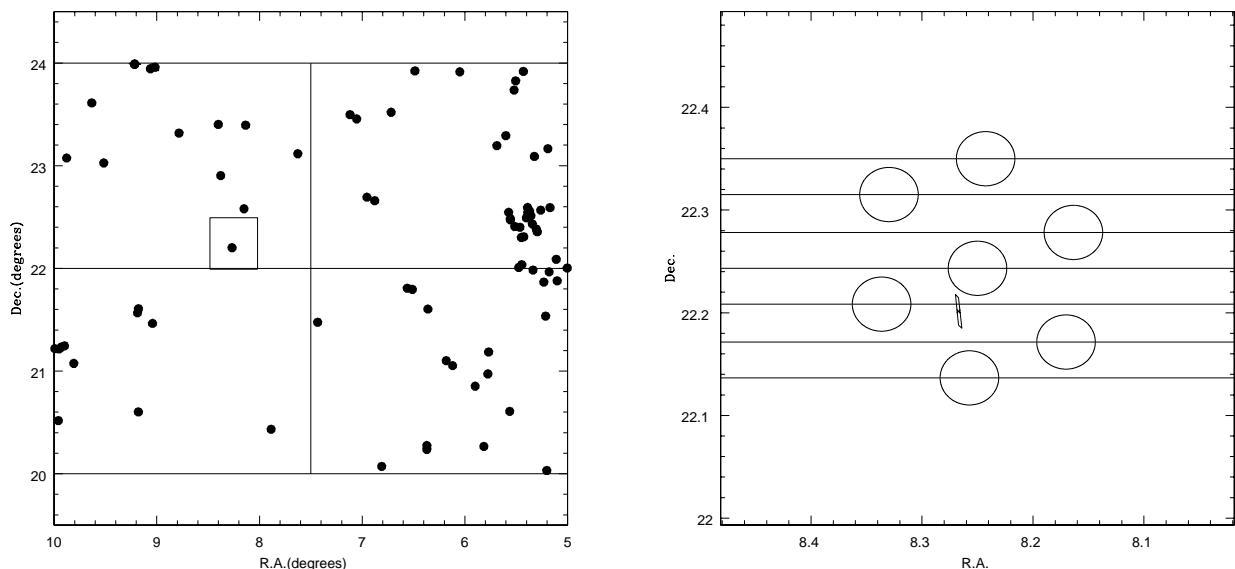


Figure 2: Left: The **Tile** centered at $(00^h30^m, +22^\circ)$ and its constituent **Quadrants** also shown in Figure 1. Symbols indicate the locations of catalogued galaxies within this tile. Right: Configuration of the 7 ALFA beams and their individual equally-spaced **Strips** during a single **Drift** stepped $14.6'$ north of the center of the $+22^\circ$ tile. The box, shown also in the left panel, is $30'$ on a side and is centered at $(00^h33^m00^s, +22^\circ14'36'')$. The polygon indicates the location within this field of UGC 325.

Data Rate:

The basic unit of a raw data file will correspond to the sequence of spectra for the 7 beams and 2 polarizations that correspond to a **Drift Group** of 7 segments, each 600 records long. For a WAPP total bandwidth of 100 MHz at 3 level (1.6 bit) sampling, each spectrum will have 4096 channels. A drift group will thus be $4096 \times 2 \times 600 \times 7 \times 4 = 137$ Mb plus headers, where the factor 4 refers to 4-byte (real) spectral values. The data rate will thus be about 0.9 Gb/hour. Each individual 10-min, single beam **Strip Segment** requires 25 Mb. The total size of a single **Quadrant** data set for a single pass survey (single beam strips separated by $\sim 2.1'$ in Dec) will thus be ~ 1.4 GB (~ 2.8 Gb for a double pass strategy). Regridding, polarization averaging, will reduce the data bulk by a factor of a few. For a number of processing purposes, it is convenient for a map to be fully loadable into memory, so this data unit size appears reasonable for current top of the line, inexpensive workstations.

Processing Sequence and Record Keeping:

As mentioned above, data will be taken as a sequence of drifts containing successive groups composed of 7 strip segments at different Decs., corresponding to the 7 beams. Note the important fact that each strip segment within a group will cover a slightly different range of R.A. from the others because of the offset geometry of the ALFA beams. In order for data quality to be validated shortly after the data are taken, bandpass subtraction is desirable right away. Flux calibration using noise diodes, bandpass calibration and baselining will thus be carried out in one of two modes:

- (a) one beam at a time, i.e. acting on sequences of single beam strip segments contained in adjacent groups, *adjacent to each other in R.A., rather than in Dec.*, or
- (b) one drift group at a time, i.e. acting on the 7 drift segments of a group. ¹

Only after the solid angle covered by a quadrant has been fully sampled by drifts at different Decs. will we proceed to producing a spectral map cube, a continuum map, possibly a re-calibration of the data by using a flux catalog of the sources within the continuum map and extracting spectral line sources.

It is foreseen here that noise-diode calibrated, bandpass subtracted single beam strip segments will be the **Level 1** data products of the ALFA Drift Survey.

Re-calibrated, baselined spectral data cubes and continuum maps will be **Level 2** data products. Gridded, smoothed spectral data cubes can also be produced at the same **Level 2**.

If multiple passes are available for a given quadrant or tile, coadded spectral data cubes and information on continuum source variability will be **Level 3** data products of the survey.

Catalogs of extracted signals are the final **Level 4** data products.

Record keeping will be a challenging task. Here the following strawman proposal is made. Note that the names here are examples and subject to the change to conform to adopted AO conventions.

- It is expected that NAIC will assign a unique observation number to each data Record (of 8 boards, 7 of which with ALFA beam data, 2 polarizations and n_{chn} spectral channels). Records will be clustered in Drift Groups of n_{rec} records. We will assign to such a unit a unique source name, identified by the starting RA and Dec. (epoch 2000) of the segment of the central pixel of the array, e.g. R021000D201500, where R and D stand for RA and Dec. If *exactly* the same Drift Group is run a second, third time, a letter 'B', 'C' is added at the end of the name string. A log of each of these data units will be produced *on the fly* automatically at the telescope. There will be a "Day Log", a cumulative "Month Log" and a cumulative "Survey Log". The Day logs will be named **tel.day.yymmdd.n.log**, where the letter **n** stands for the version number of each log, which is useful if there is more than one observing session per day. The Month logs will be named **tel.month.yymm.log**, where **yymm** identifies year and month. The cumulative Survey log is named **tel.survey.log**. The Month and Survey Logs are automatically updated at the end of each observing session. Each has an "updates section" at the beginning, listing time, date and Day log for each update. Each log will also have a "comments" section at the beginning, for 'manual' comments, each timestamped and authored. Ideally, there will be a "log line" per each Strip Segment of a Drift Group: the observation number would have an appended single digit integer to identify the beam number. For each Strip Segment line, one would list a source name, start and stop AST times and date, start and stop RA and Dec for each beam strip, AZ and ZA of the telescope, observation id number and other parameters of interest (tbd). On-line ability to comment each line of the log would be desirable, e.g. in case of board, receiver malfunction or severe RFI, etc. By protocol, all comments should be authored. It would be useful to be able to

¹In IDL data processing with PP's AO tools, a data cube is an "m" structure, of n_s (e.g. Declination) strips, n_r records (e.g. R.A. samples), n_{chn} spectral channels and 2 polarizations. A structure of the 'm' type can be obtained for either of the processing schemes discussed. In scheme (a), the n_s dimension of the structure described in the text will be replaced by the number of strip segments, sequential in R.A.; in scheme (b), $n_s = 7$, for the seven strip segments swept by the 7 ALFA beams. In either case, similar tools to those designed for 'm' structures can be applied.

flag poor data, interrupted segments, etc., “on the fly”, so that downhill processing would receive alert message. A graphic display of the sky coverage by the survey can be produced from the log files, coding differently those that have been flagged as adequate, bad, poor, etc. A software facility should be produced that will allow immediate recovery of the coordinates of any given beam, say, for strip segment xxx and record number yyy, by access to the log file. This should be available on–line at data taking, as well as off–line.

- A **Quadrant Log** is kept for each of the sky quadrants within the targetted tiles. It has all the fields in the Day Log (but it lists only the Drifts of Strip Segments that cross the quadrant solid angle). The Quadrant Log name is of the type **tel.QR0205D21.log**, where the RA and Dec are the center of the quadrant region. After each data taking session, the Day Log is mapped onto the relevant Quadrant Logs. A graphic display of the sky coverage of the tile region by the survey can be produced from the log file.
- The **Tile Log** is a running summary of the status of the 648 sky tiles. It is intended for public informational purposes and contains less information than the Quadrant Log.
- Soon (days? weeks? certainly less than one month) after data taking, the raw data will be converted to IDL structures and processed to obtain noise–diode calibrated, bandpass subtracted structures.
 - In the processing scheme (a) described earlier, Strip structures (we’ll refer to them as **s structures**) will contain data of a complete observing session *for a given Declination, i.e. for a single beam*. The **s** structures will be stored in files with names similar to those of the Drift they pertain to, but the starting RA and Dec will be specifically those of the beam (rather than those of the central beam 0) and a beam identifier will be appended to the name, e.g. R021023D201606B3.
 - In the processing scheme (b) described earlier, Drift Group structures (we’ll refer to them as **d structures**) will contain all data from a single Drift Group (i.e. all beams). The **d** structures will be stored in files with names as described earlier for Drift Groups, e.g. R021000D201500.
- At the time of creation of the **s** or **d** structures, validation of the data quality will also take place (placement in the public domain will take place only after full coverage of a tile, however). A data processing log will be produced for each processing session, named **proc.1.yymmdd.n.log**. This will list date, each processed Strip Segment name, start and stop RA and Dec, AST start and stop, AZ and ZA, beam nr, names of ancillary files produced in the reduction, cal values used, T_{sys} measured at zero continuum level, name of processing person, etc. [here, we’ll have to decide on a protocol for making the processing ‘approved’]. Once approved, the session log can be appended to a cumulative, level 1, processing log, **proc.1.all.log**. A graphic display of the reduction stage of the survey can be produced from this log. Tile Logs of data processed to Level 1 (named **proc.1.QR0205D21.log**) are also produced, if the tile is affected, each time that **proc.1.all.log** is updated. All processing log files have a section listing updates.
- When all the data for a given tile quadrant have been taken, a 3D spectral map cube and a continuum 2D map can be produced. The Quadrant Log can be used to produce a list that contains the record range of each **s** structure, that fits within the tile quadrant. Thus a map **m structure** (following PP’s terminology) can be built.
- Further processing (baselining, more accurate flux calibration, continuum map generation, spectral index and variability determination, spectral signal extraction, gridding, etc.) can now take place by acting directly on the ‘m’ structure. Although the format of the **m**, **d** and **s** structures is identical, we will reserve the **m** nomenclature for the map of a full tile.
- The four quadrants of each tile will finally be combined to produce a smoothed data cube covering the whole tile. This will be the final product made available through the NVO as a data cube.

Glossary

- **Tile:** One of the 648 basic sky subdivisions of 20 minutes ($\sim 5^\circ$) in R.A. by 4° in Dec. The tile name, e.g. TR0210D22, contains the center position of the tile. A **tile** is the basic unit of higher level data storage.
- **Quadrant:** One quarter of a **tile**, 10 minutes ($\sim 2.5^\circ$) in R.A. by 2° in Dec. The quadrant name, e.g. QR0205D21, contains the center position of the quadrant. The **quadrant** is the basic unit of Level 1 data storage.
- **Drift:** An ALFA dataset containing all spectra for all 7 beams and both polarizations at a constant declination, a sequence of 600 second segments recorded in a single observing session. File names follow the AO convention for WAPP SDFITS.
- **Drift Group:** A unit of a drift spanning a 600 seconds containing 7 **strip segments** each pertaining to a different beam and collected simultaneously by the ALFA array. Each Drift Group has a unique name, e.g. R021000D201500, corresponding to the starting RA and Dec of the central beam (beam 0) of the array. The data in a Drift Group make up an IDL **d** structure.
- **Strip:** A subset of a drift corresponding to a single beam; a single declination sequence of segments.
- **Strip segment:** a 10-min, single beam, spectral line data time series in drift mode (constant Dec.). Each **strip segment** has a unique name, e.g. R021023D201606B3, relating the starting RA and Dec of the strip, as well as the array beam number.
- **Record:** a 1 second integration dump of spectral line data, the smallest data unit.
- **d structure:** an IDL structure of 'm' type, with Level 1 processed data of a Drift Group. It has fixed size: $4096 \times 2 \times 600 \times 7$ spectral values.
- **s structure:** an IDL structure with Level 1 processed data of a Strip, i.e. a sequence of Strip Segments: a $2 \times 600 \times n_s \times 4096$ structure, where n_s is the number of consecutive Segments in the Strip. A unit similar to the **m** structure in format, except that the strips are adjacent in R.A. rather than in Dec. It is the basic processing unit to produce Level 1 output.
- **m structure:** an IDL structure with Level 1 (or higher) processed data of a full quadrant: a $2 \times 600 \times (\text{number of strip segments}) \times 4096$ structure.
- **Day Log:** a log file of all the data taken during an observing session. Name type: tel.day.yymmdd.n.log, where yymmdd stands for the day of data taking, bf n is a numeral used for multiple datat taking sessions within the same day.
- **Month Log:** a log file of all the data taken for the survey during a month. Name type: tel.month.yymm.log, where yymm stands for the month during which data was taken. The file is automatically updated at the end of each observing session.
- **Survey Log:** a log file of all the data taken for the survey. Name type: tel.survey.log. The file is automatically updated at the end of each observing session.
- **Quadrant Log:** a log file of all block array Strip Segments which cross a given quadrant. Name type: tel.QR0205D21.log where the coordinates correspond to the center of the quadrant. The file is automatically updated at the end of each observing session, if observations within the quadrant were made.
- **Tile Log:** a log file providing public information about the survey status and data availability.
- **Level 1 Processing Session Log:** a log file reporting summary information of a calibration/bandpass processing session (Level 1). Name type: proc.1.yymmmdd.n.log
- **Level 1 Cumulative Processing Log:** a log file reporting summary information of all calibration/bandpass processing sessions (Level 1). Name type: proc.1.all.log. It is updated after validation of each processing session.

- **Level 1 Quadrant Processing Log:** a log file containing a record of all Level 1 processing sessions of data relevant to a given quadrant. Name type: proc.1.QR0205D21.log, where the coordinates are those fo the center of the quadrant. It is updated after validation of each processing session.