EALFA Memo 040804B:
Data Processing Stream (Mode b) for ALFA Drift Surveys
rg & mh – 12 August 2004

Definitions for data units were given in a previous memo ("Bookkeeping for E-ALFA Drift Surveys"). Here we briefly summarize a proposal for data processing stream within the IDL environment in the mode (b) described in the above mentioned memo, to be used as a starting point for discussions.

- 0. A **Drift**, containing seven **Strips** each associated with one of the ALFA beams plus an eighth one possibly used for RFI monitoring, is obtained at the telescope and delivered as a single FITS file. This file gets stored in a (perhaps two?) raw data directory at NAIC and transferred on a regular basis to data processing sites (mode tbd).
- 1. The FITS file is decoded and converted into a sequence of **d** structures, one for each Drift Scan in the file. Those are immediately stored as named files, e.g. dR210000D021000, with the starting RA and Dec (epoch 2000) of the group, for the central feed of the array. In order to avoid clogging the IDL memory with many **d** structures and encountering the need to use different names, each Drift Scan is converted into a **d** structure, then immediately saved to disk with the assigned filename. The structure with the next Drift Scan will occupy the same location in memory as the previous one. If the noise calibration relies on cals being fired for, say, 1 sec On and 1 sec Off, before each Drift Scan, the Cal records can be stored in the form of **c** structures. They are similarly saved on disk as named files (e.g. cR205959D021000) immediately after being created. The **d** structure contains data for 7 **Strip Segments** contiguous in R.A., each taken, simultaneously, with a different ALFA beam. All strip segments contain $\sim$ 600 records of 1 sec. The **d** structure has the same format as PP's **m** structures in the AO/IDL package: 4 dimensions (2 pols, $n_{rec} \simeq 600$ records or R.A. samples, $n_s = 7$ strip segments, $n_{chn} = 4096$ spectral channels. The **c** structures are similar to the **d** structures, except that they contain a single record each.
- 2. **Noise calibration** and **Bandpass subtraction** are applied next, to one **d** structure at a time and sequentially for each Drift Scan in the strip.
  - First, all the **c** structures for a full Drift are restored to IDL, $(Cal_{on} - Cal_{off})/Cal_{off}$ ratios are computed for the observing session sequence, 'bad' cals (e.g. those that were obtaining while drifting over a continuum source) are flagged and a linear fit for $(Cal_{on} - Cal_{off})/Cal_{off}$ is obtained for the remainder. This will be used to obtain calibration values for each of the Drift Scans in the session, e.g. in the form of an array with calibration solutions indexed along the Drift.
  - The **d** structures are then restored one by one to IDL, and they are noise–calibrated and bandpass–subtracted, possibly using a 1D bandpass derived from the data within each Strip Segment in the Scan.
  - Immediately after, the bandpass–subtracted structure is interactively **Baselined**.

  The resulting structure we shall refer to as a $\mathbf{d}_1$ structure and will constitute the unit of Level 1 data products and will be stored away in a named file, e.g. d1R210000D021000 (same as the storage name of the input **d** structure, but with different prefix. Important Q: do we need to keep around both "d" and "d1" structures?). The noise calibration and bandpass subtraction is automatic and blind. The baselining is interactive, and will provide the first visual inspection of the data and its quality validation. Preliminary spectral signal extraction may be carried out at this stage, either by means of an automated matched–filter algorithm that operates in the spectral–RA space or by means of visual inspection. These processes produce a set of ancillary files:

1. a set of **cals** structures, which retain memory of the noise calibration;
2. a set of **cont** structures, which contain the fluxes of continuum sources encountered along the strip (separately for polarization and for several spectral bands, to yield some indication of spectral index);
3. a set of **bl** structures, which contain coefficients of polynomial baselines for each spectrum;
4. the **bp** bandpasses applied to each strip segment,
5. a **bad** structure, which flags pixels in position–frequency maps excluded from the estimate of the bandpasses.

- 3. Once all the data within a tile have been acquired, all strip segments with data within that tile are combined to produce an **m** structure of the tile, possibly by stitching together four **quadrants** of the tile, which may be processed separately. Continuum maps of the tile are produced and comparison of fluxes extracted from the continuum maps with those in a flux catalog are used to (i) re–calibrate the flux scale of the tile and, for a multiple pass tile, (ii) identify variable continuum sources. Regridded and compressed versions of the tile are produced — Level 2 data products — for public access.

- 4. Automatic signal extraction of spectral sources is carried out on the tile **m** structure, via a matched filter approach that can separate spurious features with unphysical characteristics (for a cosmic signal), producing a first–look catalog of sources in the tile, with an automatic quality index assigned. Candidate detections are visually inspected and assigned a quality index by an observer. Cross–referencing with other extragalactic catalogs and other tools within the NVO environment are used.

- 5. In a multiple pass survey, steps 1–4 are repeated for each pass. Catalogs of candidate detections are cross–referenced, expected shifts in frequency in the geocentric rest frame are tested. Tiles are regridded and signal extraction is carried out again, to take advantage of higher sensitivity. Coadded tiles form Level 3 data products. Final catalogs of detections constitute Level 4 data products.