EMU Data Requirements v.2.0

Ray Norris
30 January 2011

1. Introduction

This document is an updated version of the EMU data requirements submitted to ASKAP WG5 in June 2010. It takes for granted items covered in the ASDAF plan and the EMU Requirements document, and distils the wiki and face-to-face discussions since June 2010. It does not attempt to include the data requirements of POSSUM or other projects, but obviously overlaps with them.

2 Essential Requirements

2.1 Provision of Data Services

EMU data will be processed in the ASKAPSOFT pipeline processor, and the resulting data products placed on a VO server in the ASDAF where they can be accessed by EMU team members prior to quality control , and can be accessed by anyone after being validated by the quality control process. The data products will include uv data, images, and catalogs, as detailed below. Standard VO protocols will be used to enable use of Aladin, TopCat, etc. Search tools will be provided to select a subset of data by searching on any metadata range (e.g. RA, Dec, observation date, etc). Sufficient bandwidth and computational power will be provided in the server to allow automated searches that might return millions of objects.

2.2. UV Data

- All EMU uv data, image data, and source catalogues produced by the ASKAP pipeline processor will be stored and made available to users
- UV data will be available in all four Stokes parameters, and in all 300 frequency channels.

2.3 Images

The server will provide calibrated images in IQUV, including:

- An intensity image,
- a coverage map (showing which parts of sky are included in the image),
- A residual image containing all uncleaned data, in which no components have been added in (this is necessary for stacking),
- The first three Taylor terms representing spectral index and curvature (i.e. the intensity, spectral index, and derivative) in each of those stokes parameters,
- A collection of point-spread functions sampled finely enough across the image that a point-spread function can be accurately interpolated for any position.

Metadata associated with each image will give

- Date(s) of observation, details of antennas used, configuration, processing algorithms, etc
- Any known faults or error conditions on the array during the observation
- Quality control flags, including a passed/not passed flag, and details of any problems detected during quality control, and steps (if any) taken to rectify them
- A history of changes to the image entry

Individual images from different days will be spliced in the image catalog so that a request for an image centred on a position spanning two different images will return a composite image seamlessly spliced together from the two images.

2.4 Source Catalogue

The server will provide source catalogue consisting of radio components extracted from the data. Each component should be listed with the following attributes, each of which is accompanied by an estimate of its standard error.

- An unique IAU-registered identification number (e.g. ASKAPEMU1 hhmmss.s-ddmmss),
- Ra, Dec (J2000) of the peak,
- A peak and integrated (obtained through floodfill or similar algorithm) flux density,
- A peak and integrated spectral index,
- The parameters of a Gaussian (Ra, Dec, peak, bmax, bmin, position angle) of a Gaussian fitted to the component,
- In the case where a Gaussian fit fails, an estimate of the extent of the source,
- A postage-stamp image (or a pointer which extracts a postage stamp image from the image server),
- A history of changes to the catalogue entry.

2.5 Quality Control

Simple statistical tests will be done on each image as part of the quality control process, and the results written into the image metadata. These will include

- A check that the rms noise is ok,
- A check that the differential source counts distribution conforms to the expected distribution.

Tools will be provided (e.g. an image viewer with the functionality of kvis etc.) so that EMU staff can visually inspect the images prior to quality control, and perform tests on them. EMU staff will be able to write quality control flags into the data set.

2.6 VO Tools

VO tools will be made available to

- Extract images or parts of images, or parts of the source catalog, selected by searching on any metadata range (e.g. RA, Dec, observation date, etc), including automated searches which may return millions of objects.
- Extract parts of the source catalog selected by keywords (e.g. position, flux density)

3 Desirable Requirements

3.1 Provision of Data Services

- After quality control, the data catalogue will routinely be provided to NED
- ASDAF will host a VO server to serve added-value data written by the EMU team. This will include (a) a table grouping the radio components into sources (b) a table of cross-identifications to other wavelengths, and redshifts, generated by the EMU cross-ID pipeline
- All catalogue entries will be linked to publications that use these data, possibly using a DOI or similar to track the use of data.

3.2 Facilities for Data-Intensive Research

We need access to facilities with high bandwidth to the ASDAF for the following data-intensive research tasks

- Automated cross-identifications with optical/IR/other surveys (Perhaps followed by searches for clusters, superclusters),
- Cross correlation with CMB data for ISW effect,
- Cross correlation of different redshift ranges to measure cosmic lensing,
- Cross-correlation of low-surface-brightness radio data with optical galaxies,

- 2- and multi-point auto- and cross-correlation functions (or other equivalents) with different data cuts (e.g. size, crude spectrum, flux),
- Stacking of EMU image data at millions of positions chosen from other catalogues (e.g. Herschel, SkyMapper, PanStarrs, LSST),
- Extraction of low surface brightness emission cross-correlated with millions of galaxies selected from other surveys (e.g. to measure the WHIM continuum emission).

4 Points for Discussion

- We do not yet have the optimum source extraction algorithms. Work is continuing on this in the EMU team.
- For data-intensive research, EMU members need to list the tasks that we will need, develop use cases, and estimate the computational load.
- It is possible that we will need two parallel pipelines with different tapers so we get maximum sensitivity for both compact and extended sources. Alternatively, a combination of an appropriate robustness parameter plus multi-scale clean may get us close to the optimum in both cases. We need to run simulations to see if two pipelines are necessary.
- What else does the ASKAP Computing team need from the EMU team?