Radio interferometry data in the VO

A. M. S. Richards, UK ARC, JBCA, University of Manchester

February 2, 2010

Examples of selection of radio interferometry data for science cases and the associated metadata requirements and VO models. The use cases are based on data in the range of approximately 1 GHz to 1 THz (cm to sub-mm wavelengths), from instruments such as *e*-MERLIN, the EVLA, VLBI and ALMA. Many of the considerations also apply to lower frequency interferomety e.g. LOFAR or CMB observations, to IR/optical interferometry or to single dish radio observations.

The general properties of interferometric data are described first in order to avoid repetition in the use cases.

1 Character of radio interferometric data

This section describes the quantities likely to be needed for data selection and assessment by the end user (as distinct from any on-demand processing by the data provider whom, it is assumed, has access to any necessary information). In all cases, it is very desirable also to provide links to the data processing history, in particular references and uncertainties for the flux density of the source(s) used to set the amplitude scale and the position of the source(s) used as astrometric references.

1.1 Visibility data

The most basic data from a radio interferometer which are usually accessible off-line are the correlated visibilities. These consist of complex visibilities (amplitude, phase, weight), recorded per integration time (typically 0.1 – 10 sec), per baseline. The visibility function is a 2-D (or, strictly, 3-D) Fourier transform of the sky brightness distribution which can be inverted to produce an image of the sky. The locus of a baseline in the visibility plane is the projected length as seen from a source at the interferometer pointing centre, and so it changes with time due to the rotation of the Earth (Fig. 1, *right*). The visibility axes are u, v and w. Astronomers refer to this as the uv plane or the visibility plane, and to the length of a vector from the centre to a particular visibility as the uv distance. The w term is significant for some data processing but not, in general, for selection and other VO purposes. The proposed utype for visibility-plane data is is "spatial Fourier".

An interferometry observation may last a few minutes or many days or longer (not necessarily continuous), and usually consists of duty cycles switching between targets and phase-reference sources, with occasional observations of other calibrators. This is equivalent to *data quality level 1* in the VO standard. The resulting data set contains from hundreds to many millions of visibilities. Radio telescope feeds normally observe in two orthogonal polarizations, e.g. Left and Right Circular Polarization (see IVOA Note on Polarization) and the cross products can also be made, giving four polarization products. Some or all of the calibration (and flagging of bad data) can be carried out using pipelines and/or by instrument staff.

The first stage relevant to the VO is probably access to calibrated visibility data for each separate source (target or calibrator), *data quality level 2*. Such data can directly be used to make images, data cubes etc., which are the most commonly-used science products. The very large fields of view with respect to resolution elements (10^9 pixels might be required to cover the primary beam in a single image) and the need to allow for sky curvature and differential calibration, means that multiple images or facets are usually made, often initially only at the positions of known bright or interesting sources. There is no unique image-plane data product. The nature of interferometry means that a single data set can be weighted to increase resolution at the expense of sensitivity, or averaged in different fashions in time or spectral channel. Data sets from different arrays can be combined to improve the coverage of spatial

scales. Non-image products include spectral and variability curves. Many of these products involve going back to the visibility data. Moreover, *uv* plane properties are useful in selection of calibration sources and this involves information which must be shared between observatories.

This means that the VO (and radio astronomer data providers) have to present descriptions of data in two forms. Firstly, properties of potential images, spectra etc., (Section 1.2) are familiar to all astronomers but generally have to be expressed as ranges due to the flexibility in their generation, which can lead to ambiguity (since not all combinations of resolution and semnsitivity etc. are possible) or to excessive complexity. Secondly, *uv* plane properties are commonly used for data selection by people more familiar with radio astronomy, in order to, for example,

1. Download the *uv* data for local processing;
2. Extract data products such as images/spectra/time series via processing by the data provider;
3. Assess the range or scales of information present in a ready-made data product or catalogue of source properties.

Visibility data descriptions did not fit into SIAP, SSA etc. and in the past it seemed unecessary to invent a special standard. Now, the flexibility of Characterization (allowing an extensible number of axes) and the multi-dimensionality permitted by SIAv2, make it more feasible for the VO to describe visibility data. The need is also greater given the increase in the number of astronomers using radio interferometry as the EVLA, e-MERLIN, ALMA etc. are comissioned.

A visibility data set has many axes (depending on origin and contents) but these quantities are required for data selection:

- The pointing position or phase centre of the observed data set (default of ICRS in equatorial coordinates). The data may be in the form of a mosaic with multiple centres; for practical data retrieval it would be useful to give a central coordinate and to include the component image centres in metadata if advanced information is requested.
- The field of view. This is not hard-edged.
  - The field of view of a single pointing can be expressed as a radius to 50% of sensitivity/smearing, to be used as the default region of regard. In some cases it may be possible for users to override this by searching for data within a larger radius of the requested position, or requesting images (etc.) from regions further from the observational centre.
  - The field of view for a conventional single pointing is a function of the frequency (channel) resolution, the integration time and the sizes of the primary beams of the individual telescopes in the array (see further points below).
  - The field of view of a mosaic is more restrictive and might be best expressed as a complex polygon (STC).
  - Alternatively, a weight or sensitivity map may be provided, see Section 1.2. This might be impractical where the field of view is too many million resolution elements.
- Component telescope properties, such as location and diameter, can be used to deduce other properties if these are not provided explicitly. Should these be included in this model or in Provenance/Observation (and/or a link to the array web page in the absence of the latter models)?
- Diagnostics of the *uv* plane coverage.
  - Most conventional radio archives provide links to plots (such as Fig. 1) of visibility amplitude as a function of *uv* distance This is in dimensionless units of (projected baseline length)/(observing wavelength), or SI multiples, written as e.g. Mλ (mega-lambda).
  - Table 1.1 provides a machine-readable quantification of Fig. 1.
  - The range of the maximum and minimum *uv* distances present in a data set is the simplest fairly accurate quantifier. The example shown in Table 1.1 gives 114.93–3553.52 kλ.

Table 1: Parameterization of binned visibility amplitude as a function of *uv* distance, corresponding to crosses in Fig. 1.

uv distance (Mλ) in bin centre	0.19	0.56	0.93	1.30	1.67	2.04	2.41	2.78	3.15
Amplitude (Jy)	0.3898	0.3682	0.2622	0.2305	0.1714	0.1497	0.1117	0.1264	0.1248
Std. deviation (Jy)	0.0013	0.0014	0.0005	0.0009	0.0010	0.0005	0.0007	0.00094	0.00066

  - The range of intermediate spacings available as well as the limits of *uv* coverage affect the data quality. There is no commonly-used universal quantifier for this although the coverage could be described in more detail using the finer levels of Characterization. Astronomers usually inspect plots such as Fig. 1 or the dirty beam (see Section 1.2 and Fig. 2). The time axis also provides some information, see below.
  - A crude estimate of coverage can be obtained from the number of participating telescopes, the maximum and minimum baselines on the ground in m or km and the duration of the observation.

- The limiting sensitivity of the data (assuming processing for optimum sensitivity). This would usually be expressed for visibility data as the $\sigma_{\rm rms}$ noise in Jy. Strictly speaking this varies with the field of view (see above).

- The frequency range and resolution of the data.

  - Radio data are almost invariably sampled regularly in frequency (channels) over sub-bands, but a single data set from the newer instruments can contain many sub-bands of different widths, divided into different-width channels. At the coarsest level, the data can be described by minima and maxima of wavelength and resolving power, but detailed Characterisation of is only useful in frequency (or velocity, see next point) units, e.g. GHz, kHz.
  - Data may be observed at fixed velocity, in which case the convention, reference frame and rest frequency are needed (as defined by STC) for detailed Characterization, in velocity units.

- Time span and resolution, etc. – all could be expressed in ISO-8601. The conversion to MJD or other decimal units is linear and unique so it is easy to perform if the user wants to plot a time series requiring this.

  - The start and stop times of the observation.
  - The integration time (finest time resolution).
  - The on-source time, scan length etc. can be described by the detailed coverage provided for in Characterization or provided by a link to scan listings. In Fig. 1 *right* each continuous short strip corresponds to a scan of about 7.5 min, containing 56 8-sec integrations. The total duration of the observation was 13.9 hr of which 9.87 hr was spent on the source shown.

- The polarization products (also known as correlations) present, as described in the Note on Polarization which includes suggested utypes.

- Accuracy affected by calibration

  - Astrometry: Properties of phase reference source and effects due to calibration and array; ideally each should be given as a contribution in arcsec (or other angular units) to the total, so that if better measurements become available in future the accuracy can be improved.
  - Flux scale: Properties of the flux standard and estimate of uncertainty in Jy or similar (this may be broken down as for astrometry).
  - Polarization purity and polarization angle accuracy, including properties of standard sources.

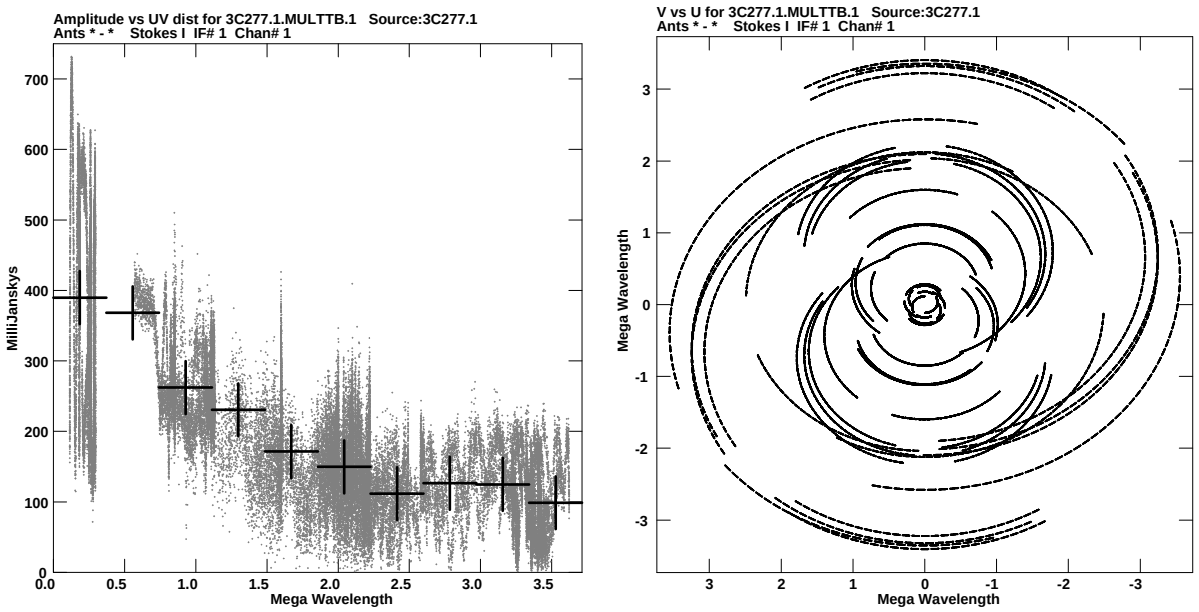

Figure 1: *left* Visibility amplitude as a function of *uv* distance for the source 3C277.1. The crosses shown the binned amplitudes as in Table 1.1. *right* The *uv* tracks for this observation.

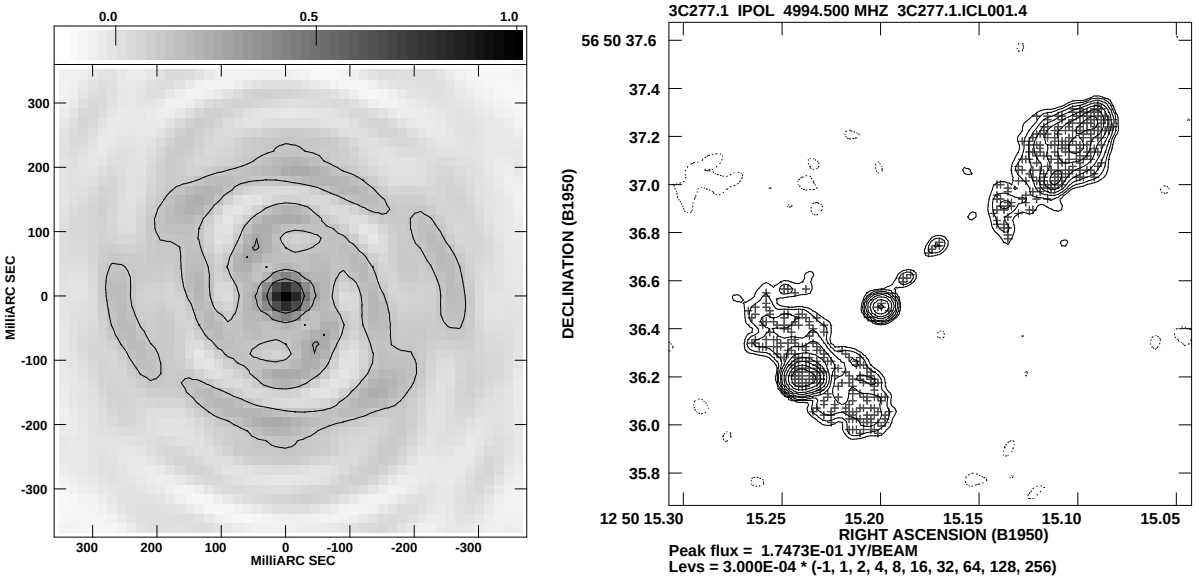

Figure 2: *left* The Fourier transform of Fig. 1 (*right*), known as the Dirty Beam. Contours are at 12.5, 25, 50% of the peak. *right* Contours of the cleaned (deconvolved) image or 3C277.1, overlaid with the Clean Component positions

1.2 Images, spectra, cubes, time series

Radio interferometry science-ready data products can be described by the general models such as SIAv2. Some of the metadata are common to the visibility data description above (e.g. accuracy limitations involving calibration); other considerations are listed here.

- A single visibility data set can be used to produce images etc. in a range of sizes and resolutions (on all axes) with a trade-off with sensitivity. Therefore, the characterisation of virtual data products will be ranges on most or all axes, and as these are often non-linear in the units mandated for the coarsest level of VO data discovery, limits (maxima and minima) are most useful.
  - This requires a means of indicating whether the visibility data are available and if so, what services (if any) are offered by the data provider.
  - Image (etc.) Quality is determined by the amount and distribution of the visibility data actually used. Images and other products offered (ready-made or virtual) will normally be made from fully edited and calibrated data as these processed have to be completed in the visibility domain. Non-specialists might find the dirty beam (Fig. 2) most inutitive, which could be quantified as the FWHM axes (see resolution description above) and the absolute value of the worst sidelobe as a percentage of the peak (12.6% in this example). Anyone who understood *uv* coverage. would use that, as described in Section 1.1.
- FITS and other radio interferometry images commonly have 4 axes e.g. R.A., Dec., Polarization, Frequency. Axes may contain one or many 'pixels', e.g. the cubes provided by NVSS have a single frequency (100 MHz centred on 1.4 GHz) but 3 polarizations; a spectral datacube might have only one polarization but 512 or more frequency pixels (channels). It is also possible to have 4 – or more – multi-valued axes.
- The fundamental spectral unit is usually frequency, see Section 1.1.
- Data products can have anything from one axis (plus observable), in the case of simple spectra, to four or more (two position, frequency, polarization ....) (see Section 2).. This could be covered using Characterization to support SIAv2. Note that spectra may be presented as 1-D 'images' with just a spectral axis, or as multi-dimensional with a single pixel on the spatial axes (but possibly with a polarization axis). There is no need to embed them as per SSA since they are usually supplied as valid FITS files, although VOTables might be used by the single dish community.
- The Resolution of a radio image is usually expressed in terms of the synthesised beam size (major and minor axes, usually in arcsec, and position angle in degrees). This is a function of the weighting used in data processing as well as the constraint of the longest *uv* distance present in the data (or the longest baseline if this is not known, see section on visibility data). There is a largest angular resolution (determined by the shortest baseline) as well as a smallest (determined by the longest baseline). It can be described using Characterization (Resolution). Virtual images would require a range of values for the beam major and minor axes and angle. The beam shown in Fig. 2 has a FWHM (the highest contour) of 58×54 milliarcsec at a position angle of $89°$. The spatial pixel size must (with current techniques) be chosen to give > 3 pixels across the beam (as determined from the longest baseline) during processing and it is inadvisable to use a larger or much smaller size. If a user wanted a coarser resolution, the image would be resampled after imaging. Resampling after imaging can always be done to any scale (as with any image, exceeding the natural resolution limits, here determined by the baseline lengths, can be misleading, but that is up to the user).
- Flux density is usually in units of (μ etc.) Jy beam^{-1}, where the beam is the synthesized beam used in making that particular image. It is not clear to me how the beam size (even a single pair of major and minor axes for a specific image) can be tied to the flux unit, nor whether making the coarsest level of data discovery use fixed units, e.g. Jy arcsec-2, would confuse more users than it helped.
- The sensitivity of a radio image is the $\sigma_{\rm rms}$ noise, which is not quantised (unlike images made using photon-counting detectors). It may vary across the field of view (see Section 1.1).

  - Certain polarization and spectral products (Stokes V and HI absorption, for example) can legitimately be negative, so the selection of valid Observable values can involve two ranges, one negative and one positive. Noise pixels can also be negative.
  - The noise can be different in different parts of the field of view or in different polarization (or other) channels, but extrema would do for initial characterization.
  - Weight or noise maps could be provided. The simplest form of these is a cut-off for the shape and extent of the primary beam, e.g. a mask to blank a rectangular image outside a certain radius.

- Images are usually 'clean'ed by iteratively identifying the brightest map pixels and subtracting a scaled version of the beam sidelobes. The measurement of the cumulative flux per pixel is stored in a list of Clean Components (Fig. 2), one entry per pixel with signal above a noise-based threshold. The Clean image consists of these components convolved with the restoring beam and added to the residual image. The Clean Components list comprises e.g. arcsec offsets from the reference position, flux density at position, eventually possibly spectral index or other information.

- Image-plane models are required for calibration, both to provide standards and to allow confusing sources to be cleaned and/or subtracted to avoid artefacts. These may be:
  - Similar to observed images but with arbitrary origins for some coordinates.
  - Lists of Clean Components (possibly transportable as VOTables).
  - Parameters to describe Solar System objects (e.g. disc axes, albedo, limb darkening).
  - Catalogues or other all-sky models of known sources within wide or sensitive fields.

- Positional data and beyond
  - Galactic plane surveys such as CORNISH (`http://www.ast.leeds.ac.uk/Cornish/index.html`) and MMB (Green et al, 2009MNRAS.392..783G) (also in other regimes such as the IR *Spitzer* Glimpse) are most conveniently accessed in Galactic coordinates. It is important to support detailed data acquisition in Galactic coordinates.
  - Solar system objects are of importance to ALMA and sometimes observed by other arrays. Thus, support for ecliptic coordinates and ephemerides could be important (covered in STC) or at least a tag to note that these are needed for such data.
  - Non "x, y" data: The main example already in common use is the HEALPIX coordinate system used for CMB (Cosmic Microwave Background) data. LOFAR, SKA (pathfinders) and some observations by other arrays such as EVLA and e-MERLIN, will have a wide enough, sensitive enough field of view to detect many, relatively bright sources. These must be included in models used for calibration. Conventional listing of sources and models are described in Section 1.2, but arrays like LOFAR and the SKA will observe the entire half-sky simultaneously and are likely to use wavelets, shapelets or other systems.

1.3 More advanced data products

A range of *data quality level 3* products (involving more than just cutouts and changing resolution) can be produced by feeding a few parameters to data provider pipelines; this is desirable either to avoid downloading large visibility data sets or because the software or at any rate the heuristics are best maintained at the data centre. Hence, the VO does not need to know how to extract these products but it does need to have a vocabulary for their labels, and to be able to parse the units. Examples are given in Table 2:

1.4 Data formats and packages

There are three main families of radio interferometry data storage; FITS, Measurement Sets (MS) (`http://aips2.nrao.edu/docs/notes/229/229.html`) or similar and Science Data Models (SDM). SDM are not really intended for public data export but are able to carry more metadata than FITS or even MS and can be used to generate information for the VO, inter alia (Viallefond 2006, 2006ASPC..351..627V;

Table 2: These examples are not exhaustive and products may be characterised slightly differently but these have known use cases. Examples of advanced radio interferometry data products. Units may be SI multiples. The Dimonsionality excludes the observable.

Product	Dimensionality	Observable units	Origin
Rotation measure image	2D space	rad m^{-2}	Multi-ν polarization angle maps
Spectral index image	2D space	Dimensionless	Images at 2 frequencies
Spectral curvature image	2D space	Dimensionless	Images at > 2 frequencies
Variability curve	1D time	Jy	Multi-epoch visibility data
0^{th} moment[1]	2D space	Jy km s^{-1}	Spectral cube
1^{st} moment[1]	2D space	km s^{-1}	Spectral cube
2^{nd} moment[1]	2D space	km s^{-1}	Spectral cube
Optical depth image	2D space	Dimensionless	Spectral cube

[1]The 0^{th}, 1^{st} and 2^{nd} moments correspond to weighted sums of intensity, velocity and velocity dispersion, usually with smoothing/blanking. They may also use frequency units.

links under ASDM from `http://almasw.hq.eso.org/almasw/bin/view/OFFLINE/WebHome`). A common SDM is being developed for ALMA and the EVLA (with local modifications) and is likely to be adopted for other interferometers. Some observatories generate other proprietary formats but, like SDM data, these are usually converted to FITS or MS before releasing to astronomers, or at least before science-ready data are offered to the VO.

The AIPS package is most widely used at present; it reads FITS images, UVFITS and FITS-IDI (Interferometry Data Interchange `http://fits.gsfc.nasa.gov/registry/fitsidi/FITSIDI.pdf`) and some proprietary formats. It exports FITS (UVFITS and images) and can be used to extract metadata and some parameters etc. and print out ascii files. AIPS is likely to be required for its wide functionality, such as specialised tasks for calibrating VLBI data, for some time to come. CASA is being developed for ALMA and the EVLA and will be widely used for next-generation interferometers, at least for imaging and the later stages of calibration. CASA is a python interface to the AIPS++ toolkit and imports, stores and process MS and a related image format. CASA can import some SDM and proprietary formats and both import and export UVFITS and FITS images. However, many extension tables and some other metadata are lost during FITS $\Leftrightarrow$ MS/related image conversion unless these can be captured and stored separately; VO standards may have a rôle here.

Both packages have some single-dish functionality and can handle (near-) science-ready images from many domains.

Most generic image display and measurement packages e.g. GAIA, Aladin, SExtractor, can handle science-ready FITS radio image cubes and spectra well enough for qualitative work but commonly run into one or more problems. Firstly, units of Jy $beam^{-1}$ are not recognised or the beam size is not easily accessible. Secondly, negative pixels are legitimate. Thirdly, sub-milli-arcsec and sub-km s^{-1} precision may be required. Fourthly, many images have a polarisation axis in addition to 2 spatial axes and often a spectral axis. On the plus side, radio images are almost always in physical units with orthogonal spatial axes aligned with a recognised coordinate system (although exceptions may become increasingly common as wide-field instruments are developed).

2 Use Cases

2.1 Astrophysical calibration sources

This Use Case is derived from the needs of observatory staff to draw up lists of potential calibration sources meeting a range of criteria, and of users who need to supplement these lists. The description of a potential calibration source should ideally include most of the following, depending on its rôle. All sources are assumed to be continuum unless otherwise stated.

- Position and position accuracy, possibly at the milli-arcsec level. This may require:
  - Proper motion, ephemeris or orbital information and its temporal reference.
- Model of flux distribution. This may be:

  - A simple point;
  - Parameters describing a fitted model such as a Gaussian component, a disc, etc.;
  - A table of Clean Components;
  - A FITS- or MS-format image.

  and should also include information about scope, such as:

  - Date of measurement and/or time variability curve parameters (including origin);
  - Frequency of measurement and any information about spectral index etc.;
  - *uv* ranges used to obtain measurement and/or parameterization or plot of amplitude v. *uv* distance, or description of array configuration used;
  - Limiting sensitivity of a complex model.

- Model of (usually linear) polarization properties.
  - Polarization angle;
  - Distribution of polarized intensity
  - Further details as for flux distribution, as applicable.

- For wide fields and field-based calibration, the model needs to include other sources in the field of view around a calibration source.

- Spectral lines/interference. Some sources in some frequency ranges and configurations may be affected by spectral features and these should be avoided if possible or modelled if not.

Where possible, all information should be accountable, with references and uncertainties. In practice, partial information will be available from different sources and at different frequencies from that required, etc., necessitating interpolation/extrapolation by observatory or user tools. Hence, it would be helpful to have some means of searching around a 'position' (on any axis), and returning the n closest matches. This is akin to the Top[1] facility suggested for SIAv2. It would be possible but less convenient to search within ranges, in the absence of prior knowledge of how far the user might have to go (e.g. up or down in frequency) to find useful sources.

Sources of information include:

- Published data e.g. via online journal links, NED, CDS (e.g. SPECFIND, Vollmer et al. 2009).

- (Other) observatory lists. These may be flat tables of parameters, or contain links to plots or FITS or other data.

- Data extracted from previous observations. Ideally this will include metadata extracted from an SDM or otherwise to supplement the limitations of FITS/MS related formats.

2.2 Extracting variability and rotation measure curves

This is based on a project devised and implemented by Tony Rushton for data mining the MERLIN archive of calibrated visibility data. X-ray binaries contain a compact core with variable flux density and can also have variable, usually fainter, extended emission.

2.2.1 Time variability curves

The first aim of this project was to obtain a radio 'light' curve for the core by summing the visibility data over, e.g., 1 min intervals. It is not possible to make a reliable interferometry map in a very short time, but the flux density of a point source can be measured. This requires shifting the phase centre of the calibrated visibility data to the position of the core. The following steps form a suggested series of SIAv2 requests, with sample values.

1. Data discovery: query (probably specific) radio archives to investigate the availability of suitable data:

[1] we should choose a new name to avoid confusion with SQL command TOP.

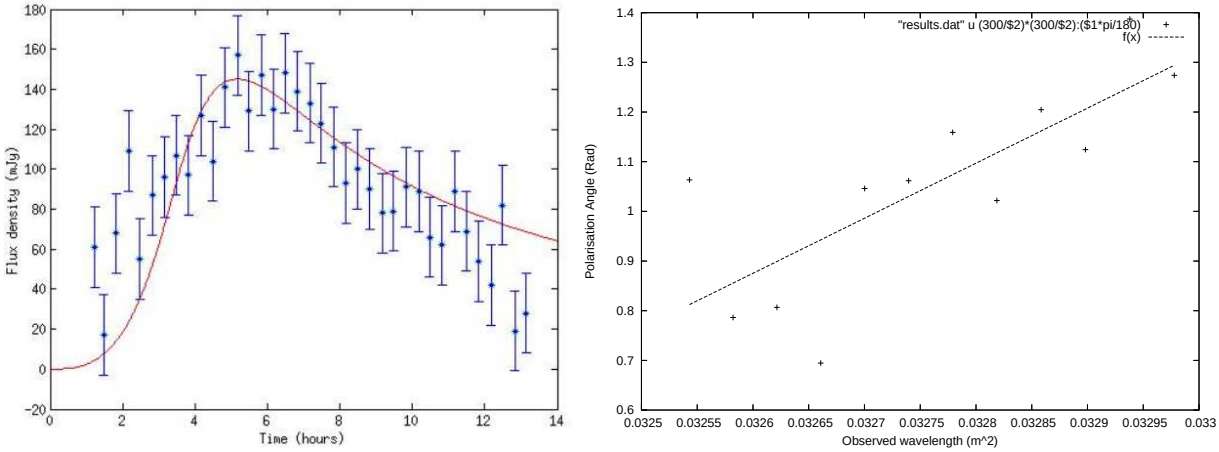

Figure 3: X-ray binary GRS 1915 (courtesy Tony Rushton). *left* Visibility amplitude as a function of time distance. *right* Polarization angle as a function of observing wavelength[2].

- Calibrated visibility data
- Covering the position of GRS 1915 provided by SIMBAD
- Within the frequency range 1 to 30 GHz
- Resolution $\lesssim$0.3 arcsec

2. Data selection: Desirable information, including properties relating to the above list (would queries on all fields have to be included in order to get the details?). If not all is available, the user could probably use a link to the observatory web page to deduce enough.
   - Frequency bounds
   - Potential image angular resolution or some kind of *uv* coverage information
   - Total time bounds of observation and time on GRS 1915.
   - Sensitivity (e.g. predicted limiting $\sigma_{\rm rms}$ of map made from all data, or information allowing this to be calculated by the user).
   - Observational pointing position and field of view (e.g. to 50% smearing).
   - Origin or accuracy of flux scale
   - Services offered e.g. variability curve extraction or user-customisable pipeline.
   - Is a map already available?
3. Data extraction of variability curve. For each epoch and frequency:
   - Make or obtain map and measure position of GRS 1915 core
   - Rotate phase centre of total intensity visibility data to this position
   - Sum visibility data over all baselines (or within a specified *uv* range)
   - Bin resulting data into 1 min time intervals
   - Measure flux density and scatter in each time bin
   - Return table and binned visibility data to user.

The output should enable the user to make a plot like Fig. 3 (*left*).

2.2.2 Rotation measures

A similar excercise is possible to calculate the rotation measure. At step 1., add a query as to what polarizations are present. At step 2., the frequency (channel) resolution is also required. The total time is not needed if the sensitivity is available. Step 3. then becomes:

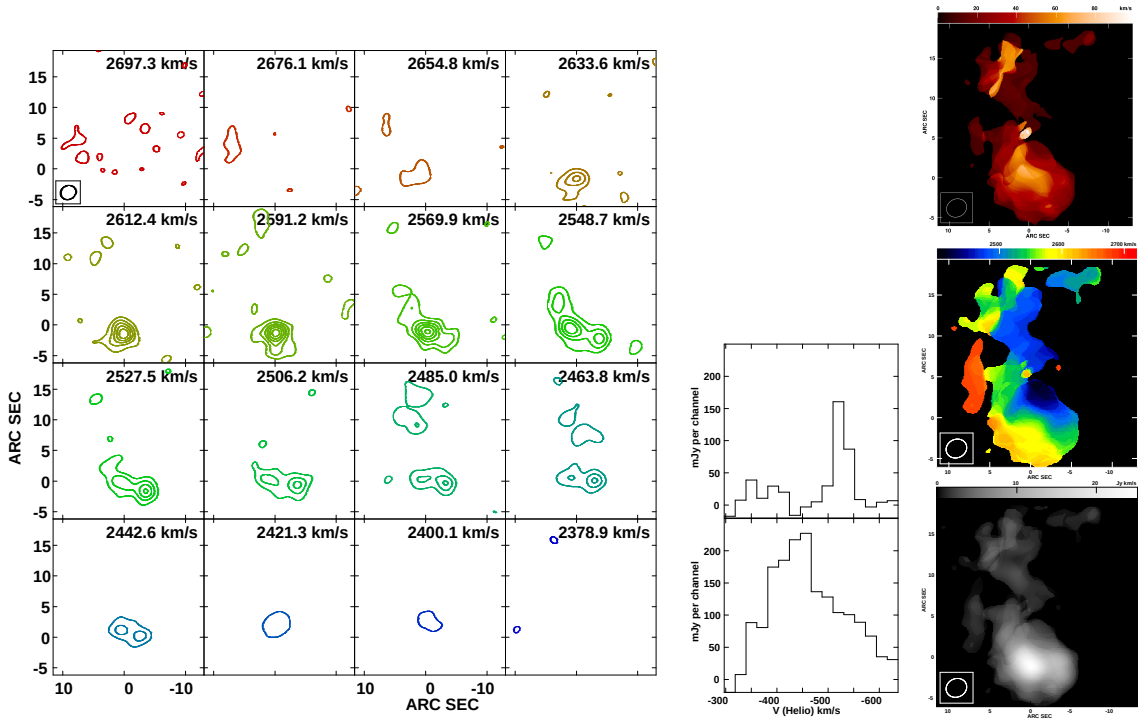

Figure 4: The Medusa merging galaxies in CO J=1-0, observed with OVRO (Aalto & Hüttemeister 2000 2000A&A...362...42A). Representation of spectral data cube and products (see Table 2). From left to right: spectral data cube; spectra extracted from the N and S components; 0^{th}, 1^{st} and 2^{nd} moments (total intensity in greyscale, velocity in rainbow, velocity dispersion in flame).

3. Data extraction of rotation measure plot. For each epoch and frequency band:
   - Make or obtain total intensity map and measure position of GRS 1915 core
   - Make or obtain polarization angle and polarized intensity map(s) within specified frequency intervals
   - Measure polarization angle from each map at position of core
   - Return table of polarization angle v. frequency and maps to user.

The output should enable the user to make a plot like Fig. 3 (*right*), the slope of which gives the Rotation Measure.

3 Moments

Figure 4 shows a contour plot of the planes of a 3D image cube together with data products extracted using specialised software (to make it easy to apply blanking and handling of units). The user would specify spatial and spectral coordinate ranges to generate the spectra. The spectra could returned as ascii files or VOTables incorporating the originating image details and window applied (used to generate these plots), or as a FITS file (by a different processing route). Moment images are sufficiently common to merit VO terms (e.g. MOMO, MOM1, MOM2) which would be recognised by the interface to a pipeline capable of deducing the appropriate blanking (in an advanced case the noise-based blanking and smoothing could be user-specified). 2D FITS images would be returned.