Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921317000163

Integrated data access, visualization and analysis for Galactic Plane surveys: the VIALACTEA case

Sergio Molinari[1], Robert Butora[2], Stefano Cavuoti[3], Marco Molinaro[2], Giuseppe Riccio[3], Eva Sciacca[4], Fabio Vitello[4], Ugo Becciani[2], Massimo Brescia[4], Alessandro Costa[2] and Riccardo Smareglia[3]

[1]INAF-Istituto di Astrofisica e Planetologia Spaziale, Via Fosso del Cavaliere 100, 00133 Roma, Italia
email: `molinari@iaps.inaf.it`

[2]INAF-Osservatorio di Trieste, Via G. Tiepolo 11, 34131 Trieste, Italia

[3]INAF-Osservatorio di Capodimonte, Salita Moiarello 16, 80131 Napoli, Italia

[4]INAF-Osservatorio Astrofisico di Catania, Via Santa Sofia 78, 90125 Catania, Italia

Abstract. The VIALACTEA project brings to a common forum the major new-generation surveys of the Milky Way Galactic Plane from 1μm to the radio, both in thermal continuum and in atomic and molecular lines, to attack in a systematic way the characterization of the Milky Way as a star formation engine. Images, catalogues, spectroscopic datacubes and radiative transfer models of the Spectral Energy Distributions (SEDs) of sites of star formation have been incorporated and indexed in the VIALACTEA Knowledge Base (VLKB). The VLKB consists of a combination of a relational database where the VIALACTEA data and metadata are homogenised and stored, and a filesystem-based stored information. This infrastructure allowed, among others, the generation of extensive catalogue for compact sources and extended structures in the Galactic Plane, the implementation of data-mining algorithms for the band-merging of multiwavelength data and expert systems for the automated analysis of molecular line surveys to extract critical kinematical information and derive distances using Galaxy rotation curves and new 3D extinction maps. A new VIALACTEA 3D Visual Analytics interface has been developed that provides integrated access and analysis of continuum and spectroscopic images together with catalogue data directly interfacing with the VLKB.

Keywords. stars: formation, methods: data analysis, astronomical data bases: miscellaneous

1. Introduction

The Milky Way Galaxy, our home, is a complex ecosystem where a cyclical transformation process brings diffuse barionic matter into dense unstable condensations to form stars, that produce radiant energy for billions of years before releasing chemically enriched material back into the InterStellar Medium (ISM) in their final stages of evolution. Star formation is the trigger of this process, eventually driving the evolution of ordinary matter in the Universe from its primordial composition to the present-day chemical diversity necessary for the birth of life.

Although considerable progress has been made in the last two decades in understanding the evolution of isolated dense molecular clumps toward the onset of gravitational collapse and the formation of stars and planetary systems, a lot remains still hidden. We do not know the relative importance of gravity, turbulence or the perturbation from spiral arms in assembling the diffuse and mostly atomic Galactic ISM into molecular dense

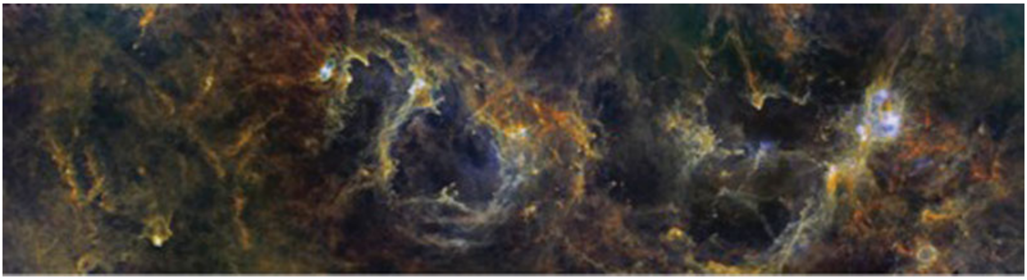

Figure 1. The Galactic Plane at 106° $< l <$ 116° as seen in the Herschel/Hi-GAL survey (R: 350μm, Y:160μm, B:70μm). It is apparent the unprecedented ability of new-generation infrared surveys to trace cold dust in the diffuse ISM, cold filamentary structures and bright and active star-forming regions.

filamentary structures and compact clumps. We do not know the relative importance of gravity or external triggering in the onset of the gravitational collapse leading to star formation, nor we know the role of magnetic field in the process. We do not know how the role played by these different agents changes from extreme environments like the Galactic Center to the quiet neighborhoods of the Galaxy beyond the solar circle. We do not know how to quantitatively relate the different physical agents at work, to the rate and the efficiency with which they are able to turn gas and dust into stars. This lack of a “fundamental theory” or, rather, of a galaxy-scale predictive model for star formation, is the key issue to be addressed by VIALACTEA.

Today for the first time, it has been possible to engage this ambitious challenge thanks to a new suite of cutting-edge Milky Way surveys that provide a homogenous coverage of the entire Galactic Plane and that have already started to transform the view of our Galaxy as a global star formation engine. New instruments in space and on the ground have delivered information of unprecedented depth and spatial detail, covering wide swathes of the plane, and spanning the electromagnetic spectrum. The combination of near-Infrared (IR) ground surveys data, mid-IR and far-IR dust continuum obtained by ESA’s HERSCHEL (Pilbratt *et al.* 2010) and NASA’s SPITZER and WISE (Wide-Field Infrared Survey Explorer) satellites, with radio free-free continuum and gas-tracing atomic and molecular spectroscopy from ground-based observatories, has allowed to compile the first complete census of sites of ongoing and potential star formation in the Milky Way. Figure 1 illustrates the complexity of the datasets at hand, with compact sites of star formation interspersed with intricated network of extended bubbles and filamentary structures.

The volume and complexity of these new survey datasets calls for a radical re-evaluation of the current science and data analysis techniques. A number of data visualization packages exist, and yet none of them integrates data access and analysis of 2D images, catalogue source properties and 3D spectral datacubes into a single client application. The Virtual Observatory (VO) concept of distributed data is not optimised for deep scientific exploitation of the VIALACTEA data unless an intermediate layer of astronomical metadata information is provided to homogenise the description of the data. The normal discovery services for data archives are limited to point-like objects, or compact objects that can be described with position-size information. Extended structures like clouds or complex filaments that are critical actors in the star formation process are not currently catalogued with descriptors that enable there query & access via standard Structured Query Language (SQL) queries to databases. Current building of Spectral Energy Distributions (SEDs) by naive source catalogue band-merging purely based on positional

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press

matching do not provide additional metrics to evaluate the quality of the various associations, leading to complex SEDs that ramify more and more as the spatial resolution increases as we proceed from sub-millimeter and far-infrared wavelengths to the mid-IR and near-IR.

2. The VIALACTEA Knowledge-Base

Implemented in a SQL-accessible relational DataBase Management System (DBMS), the VLKB (Molinaro *et al.* 2016) content is deployed through a web service that follows the International Virtual Observatory Alliance (IVOA) Table Access Protocol (TAP) (Dowler *et al.* 2010) recommendation, and requires user authentication to preserve data policy. The VLKB provides access to images and catalogues of the Galactic Plane, including: i) dust filamentary structures catalogues (Schisano *et al.* 2014), ii) dust bubble structures catalogue, iii) compact source single-band as well as band-merged photometric catalogues, as well as iv) survey images from Hi-GAL (Molinari *et al.* 2016), MIPSGAL (Carey *et al.* 2009), WISE (Wright *et al.* 2010), CORNISH (Purcell *et al.* 2013), and 3D extinction maps (Arab & Cambresy, in prep.) and v) full set of atomic hydrogen and molecular lines (mainly CO andisotopologues, but also more complex molecules) data cubes survey metadata from Galactic Plane surveys.

Besides the surveys' metadata description available through the TAP service, the spectroscopic datacubes can be searched to investigate the spatial coverage of the dataset with respect to a desired line of sight and circular region around it and, given the availability response, each interested cube can be cut along the positional and velocity (spectral) dimensions to allow for a more efficient and less band consuming data transport over the network as well as a lighter data volume to perform subsequent scientific processing. The VLKB is specifically designed as a resource that can be queried and consumed from data analysis codes and the VIALACTEA Visual Analytics application (see §4) via standard SQL queries based on position, search radius, and data type/subtype. The same search & cutout service can be accessed from a web interface. The VLKB TAP and search & cutout services † are designed keeping in mind the goal for an IVOA Observational Core data model compliant service for the full set of radio surveys and a unique TAP resource to expose all of the VLKB content. A dedicate service was implemented on the server side to the merge adjacent datacubes. The "Montage" library ‡ was used for this task, incorporating the velocity axis regridding to homogenize the velocity scale; additional parameters useful for the analysis are also reported, including the amount of undefined pixels in the datacube, overlap quality information when calculating an overlap with the region covered by the datacube, added CUNIT3 to cutout files when missing from original, vertex points to search results, and ensure that search operates only on primary Header Data Unit (HDU).

Filaments & Bubbles tableset holds all the information related to the diffuse objects identified from Hi-GAL continuum tiles (for filaments) and Hi-GAL and CORNISH tiles (for bubbles). Filamentary structures are described using 3 tables, identifying filaments as the primary complex object, branches as their components in terms of "linear" areas but also their spines (nearly 1-dimensional), and nodes as the connection points of the various branch segments that compose a filament. Bubbles are described as a unique catalogue of diffuse objects, because their roundish shape doesn't require further relationships among

† The search and cutout services of 2D and 3D datasets incorporate AST library (Berry *et al.* 2016) for FITS header interpretation and coordinate transformations.

‡ http://montage.ipac.caltech.edu/

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press

their components. Both filaments and bubbles tables are completed with positional and global details plus morphological information, contour and area representation. Contours are represented as ordered sequences of sky positions (i.e. a polygon outlining the diffuse structure), while area, i.e. celestial sphere coverage is represented in MOC format (Multi-Order Coverage Map, Fernique *et al.* 2014 an IVOA Recommendation), that is an HEALPix (Hierarchical Equal Area isoLatitude Pixelization) tessellation of the diffuse object's area to be used for easier cross-match with other positional features in the VLKB or other databases. MOCs are stored as string JSON objects in the database. Compact sources tableset contains all the single band catalogues used by the project to build up a band-merged catalogue of compact sources.

The VLKB's holdings for compact source photometric catalogues include: Hi-GAL (Molinari *et al.* 2016), ATLASGAL (Schuller *et al.* 2009) , BGPS (Aguirre *et al.* 2011), MIPSGAL (Carey *et al.* 2009), MSX in 4 bands (Egan *et al.* 1999) and WISE (all 4 bands plus the additional 2MASS fluxes already contained in the catalogue) catalogues for the Galactic Plane. The bandmerged catalogue is produced from these ones using data mining techniques included in the *Q-FULLTREE* tool developed in the VIALACTEA project itself (see §3). All of the catalogue records contain a MOC derived tessel index for quick match against diffuse objects.

The VLKB also includes other tables needed for the scientific analysis of the band-merged SEDs from compact sources. One table is devoted to the grid of synthetic protocluster SED evolutionary models, consisting of 20 million SED records to synthesize energy distributions in the observed bands. Another one is meant to keep track of the velocity information processed using the radio cubes FITS and database used in the distance estimation process of the various sources on the galactic plane. Velocity records including distances was also used to input distance estimates and cartesian positions of the band-merged catalogue sources. All the tables, for diffuse and compact sources, include positions in galactic as well as equatorial coordinate systems. Also, as already said, HEALPix derived information has been used for quick indexing reference in matching positions among the two types of sources. The full set of tables and indexing used within the database server sums up to about 50GB of database space, with table going from a few KB up to a couple of tables spanning up to 10 GB. These catalogues are accessible through a standard TAP interface, its TAP_SCHEMA is part of the database where the VLKB sits, nearby the above described table-sets. Additionally a "XMATCH" web service was developed specifically for direct cross-match on positions between diffuse objects (filaments, bubbles) and compact sources. It is based on HEALPix tesselation and reachable by web-browser or tool/library which can do HTTP requests. The service allows to search among compact sources, filaments or bubbles by specifying the identifier of an object. Alternatively specifying a region on the sky in form of a polygon. To speed up the databases queries a separate virtual machine was set up running the database server. Apart from indexing, to speed up metadata search and match an "in-memory" solution has been adopted to reduce the amount of time needed for searches on large catalogue tables.

3. Complex source band-merging

The proper bandmerging of photometric source catalogue, in which we included all the bands of MSX and all the bands of WISE (pushing the coverage down to 3.4 μm), the 5 Hi-GAL bands augmented with ATLASGAL and BGPS coverage, require that all the multiple counterpart associations found as we go toward shorter and shorter wavelengths are captured.

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press

The tool Q-FULLTREE (Quick Full Tree on Ellipse), has been implemented to trace and register all partial and full band-merged sequences based on topological cross-matching of sources and developed as a multi-thread Python open-source application. The tool is designed to handle an arbitrary number of additional bands, coming from other survey projects external to Hi-GAL.

The source matching (see Riccio *et al.* 2016) is defined based on the following function: $Ell(i,j) = [(x_{bi} - x_{bj})^2]/a^2 + [(y_{bi} - y_{bj})^2]/b^2$, where a and b are the two semi-axes of the ellipse (calculated upon the two given values of the Full Width at Half-Maximum (FWHM) of the source, centre of the ellipse), x and y are the coordinates of the higher resolution counterpart (opportunely corrected by the position angle variation). If $Ell(i,j) \leqslant 1$ then there is a positive match between the two counterparts.

In case there is more than one higher resolution source included in the ellipse of a more extended counterpart, the value of the ellipse formula provides a by-product reliable and fast distance estimation of the object from the ellipse centre that is used to assign a different score to all the candidate counterparts included in the same ellipse. Furthermore this scoring system takes implicitly into account the relationship among candidate sources, their bands and dimensions of ellipses. This scoring system is useful to generate an ordered reliability index for each candidate match (CM) between two bands, also helpful to navigate a posteriori into the catalogue and to make easy several kinds of correlated information extraction. The quality scoring and flagging is based on a simple criterion, by combining the multiple values of the ellipse function calculations along a sequence of multiple band matching.

Let be a sequence of $\mathrm{b}_1, \ldots, \mathrm{b}_N$ bands ordered in decreasing wavelengths. We define CM = (x_{bi}, x_{bj}) a generic couple of sources in two different bands where x_{bj} is included in the ellipse centred at x_{bi}. The couple is indeed a CM between two bands. For each candidate match CM we define the Confidence Level (CL) of the candidate match as: $CL(b_i, b_j) = 1 - Ell(i,j)$. By definition of CM, all the occurrences of source couples whose ellipse function value > 1 are not considered (simply because in this case there is not any match between the two sources). Therefore CL is always less or equal to 1.

CMs are then linked band by band to form Candidate SED Sequences (CCS). In other words a CSS is composed by a number of matches which we indicate as the number of elliptical matches NE between all the possible two-elements bands combinations.

For a candidate SED sequence among M bands, we may have a theoretical maximum number of elliptical matches

$$TNE = M!/[2!(M-2)!] \tag{3.1}$$

CL terms. These CL terms are formally independent, although implicitly correlated by the fact to belong to a candidate SED sequence, hence candidate to represent the same compact source in all bands at different spatial resolutions. In this sense the CL terms take into account the cross-correlation among various sizes of ellipses and relative positions of the candidate sources all over the chained sequence of matches. This information can hence be used to assign a different degree of likelihood in case of multiple matches (candidate SED sequences with common sources in some bands). Starting from such CL series, it is possible to define a Merit Score (MS) defined as follows:

$$MS(CSS) = NE/TNE \sum CL_i \tag{3.2}$$

The weight term NE/TNE can be treated, in statistical terms, as the well-known Dice coefficient, also known as Sorensen index, a statistical figure used to compare the

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press

similarity among samples. When taken as a sequence similarity measure, the coefficient may be calculated for two entries, x and y as follows:

$$DICE(x, y) = 2n_{x,y}/(n_x + n_y) \tag{3.3}$$

where $n_{x,y}$ is the number of common elements present in both x and y (i.e. their intersection), while n_x and n_y are respectively the elements in x and y (in practice their union in set-theoretical jargon). In our case the function in eq. 3.3 has a high correspondence with the weight term of the equation 3.2, because NE is just the number of common entries in the sequence, while TNE is the total theoretical number of matching sequence elements.

An additional tool, named FT-Recap (Full Tree-Recap) has been designed and developed to perform a post-processing re-organization of Q-FULLTREE output, in order to fulfil the SED visualization system expectations. In particular, all partial and complete band-merged sequences with common sources, found by Q-FULLTREE, are grouped by assigning them to a unique merging tree identifier, making possible to filter the visualization of SEDs (Spectral Energy Distributions) with higher cardinality (i.e. number of bands in the merging sequence).

The resulting SEDs are no longer single SEDs, but are rather SED Trees with branch departures at each band in which multiplicities are found. Quality figures are computed and assigned based on the relative distance of counterparts at the different wavelengths as well as to the number of bands in the merging sequence, so that each branch in the SED is assigned a Merit Score that can be used to filter out less reliable branches in the science analysis.

4. 3D Visual Analytics

A new Visual Analytics client application has been designed and implemented closely integrated to the VLKB services (see §2) to handle in an integrated framework the different types of datasets in an astrophysically sensible fashion, via the design and implementation of specific science use-cases. The tool is a cross-platform application implemented in C++ for the core functionalities using Qt† for the user interface and the adopted rendering engine is the Visualization Toolkit‡ (VTK). The access to VO-compatible databases and archives is performed using the Table Access Protocol (TAP) service interfaces. The interoperability with the other VO-based tools (e.g. Aladin, Topcat) is guaranteed thanks to the implementation of the SAMP (Simple Application Messaging Protocol) protocol.

A Main Window (hereinafter named "visual query") has been implemented to allow the users to easily query and retrieve data from the VLKB search & cutout services. The visual query allows the user to navigate into a full view of galactic plane (with common zoom, scroll and pan functionalities) in order to better identify the region of interest from which to start the analytic operations. The selection of the region of interest can be carried out choosing a point on the map and specifying the radius of the selection or drawing a rectangular region inside the map. When queried, the service returns the selected 2D fits image containing the user selected region. The 2D map visualization allows to add other surveys images as layers on top of the visualised one. The new layers are aligned (position, scaling pixel size, rotation) to the "base image" according to the information contained into their header. The user can interact with each layer

† Qt web page: `https://www.qt.io/`
‡ VTK web page: `http://www.vtk.org/`

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press

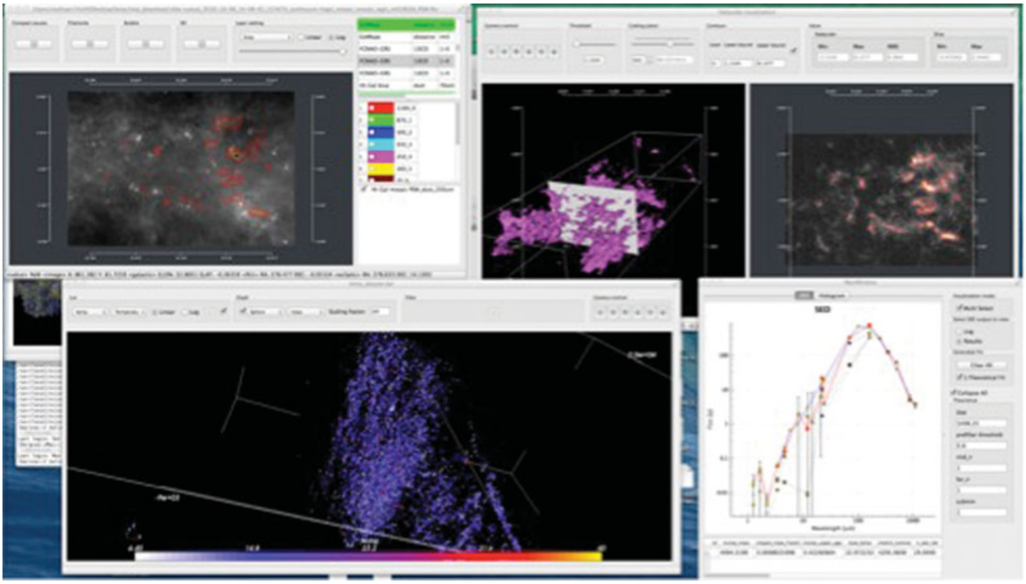

Figure 2. Screenshots of the various windows of the VIALACTEA 3D Visual Analytics application to analyse the multi-facet data contained in the VLKB in an integrated framework.

activating or deactivating the visualization, changing the opacity or changing the order in the visualization stack. It is also possible to add filaments and bubbles overlayed to the visualised 2D image. Bubbles and filaments morphological information are stored into the VLKB.

A new 3D user interface has been implemented to visualize spectral datacubes. The rendering window are splitted in two panel. On the left one a 3D visualization of the data (using isosurface algorithm) is shown, while the right panel shows a single slice of the velocity datacube. The possibility to have isocontours displayed on top of the selected slice has been implemented. The contours are also reported on the 2D map image. The user can obtain a 3D visualization of compact sources on the galactic plane selecting a region on the FITS image or specifying the coordinate in order to query the VLKB. The 3D rendering is interactive, and the user can change the colourmap selecting one of the scalar field which is in the database. The glyphs visualization can be activated to change the shape and size of the visualised points according to the selected field in the database.

Finally, SED visualization has been implemented supporting the view of tree-SEDs capturing multiple counterparts matching in the band-merged photometric catalogues (see §3) contained in the VLKB. It is possible to have different fits plotted as layers that the user can activate or deactivate and the possibility to show histogram plots of the fit results. Figure 2 shows a snapshot of a typical analysis session using the Visual analytics tool.

References

Aguirre, J. E., Ginsburg, A. G., Dunham, M. K., *et al.* 2011, *ApJS* 192, 4

Berry, D. S., Warren-Smith, R. F., & Jennes, T. 2016, *Astronomy and Computing* 15, 33

Carey, S. J., Noriega-Crespo, A., Mizuno, D. R., *et al.* 2009, *PASP* 121, 76

Dowler, P., Rixon, G., & Tody, D. 2010, *IVOA* Recommendation, 27 March 2010

Egan, M. P., Price, S. D., Shipman, R. F., Gugliotti, G. M., Tedesco, E. F., & Moshir, M. 1999, *ASP Conference Series*, 177, 404

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press

Fernique, P., Boch, T., Donaldson, T., Durand, D., O'Mullane, W., Reinecke, M., & Taylor, M. 2014, *IVOA* Recommendation, 2 June 2014

Molinari, S., Schisano, E., Elia, D., Pestalozzi, M., Traficante, A., Pezzuto, S., Swinyard, B., Noriega-Crespo, A., *et al.* 2016, *A&A* 591, 149

Molinaro, M., Butora, R., Bandieramonte, M., Becciani, U., Brescia, M., Cavuoti, S., Costa, A., Di Giorgio, A. M., Elia, D., Hajnal, A., Gabor, H., Kacsuk, P., Liu, S. J., Molinari, S., Riccio, G., Schisano, E., Sciacca, E., Smareglia, R., & Vitello, F. 2016, *SPIE Proceedings*, 9913

Pilbratt, G., Riedinger, J. R., Passvogel, T., *et al.* 2010, *A&A*, 518, L1

Purcell, C. R., Hoare, M. G., Cotton, W. D., Lumsden, S. L., Urquhart, J. S., Chandler, C., Churchwell, E., *et al.* 2013, *ApJS* 205, 1

Riccio, G., Brescia, M., Cavuoti, S., Mercurio, A., di Giorgio, A. M., & Molinari, S. 2016, *PASP, in press*

Schisano, E., Rygl, K. L. J., Molinari, S., *et al.* 2014, *ApJ*, 791, 27

Schuller, F., Menten, K. M., Contreras, Y., *et al.* 2009, *A&A* 504, 415

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., *et al.* 2010, *AJ* 140, 1868

https://doi.org/10.1017/S1743921317000163 Published online by Cambridge University Press