Lessons learned and challenges in maintaining the ViaLactea Knowledge Base

Marco Molinaro[a], Robert Butora[a], Giuseppe Tudisco[b], Fabio Vitello[b], Milena Benedettini[c], and Sergio Molinari[c]

[a]INAF – Astronomical Observatory of Trieste, Trieste, Italy
[b]INAF – Astrophysical Observatory of Catania, Catania, Italy
[c]INAF – Institute of Space Astrophysics and Planetology, Rome, Italy

ABSTRACT

The ViaLactea Knowledge Base (VLKB) was designed and initially developed within the EU FP7 VIALACTEA project that included a Work Package dedicated to create infrastructure and tools to perform research in Milky-Way astrophysics. The infrastructure's goal was to set up an archive and services to enable that research. About 50 dataset collections (35k datasets of various sizes, in FITS format), 10 catalogues of compact sources (from thousands to a few million rows), a catalogue of morphological complex sources (few thousand sources), and a few other catalogues and simulated datasets were included in the archive, worth about 1TB of data. On top of those data, and their metadata descriptions, a set of services was deployed: a search and access (cutout and merge) service for the datasets, a general Table Access Protocol (TAP) service for all metadata and catalogues and some other dedicated attempts for serving specific user requirements. All the interfaces were developed in combination with the dedicated client, the ViaLactea Visual Analytics (VLVA) but were designed keeping in mind the discovery and access scenario that is continuously developed in the Virtual Observatory (VO) ecosystem. Indeed, interoperability was brought inside the VLKB afterwards, slowly (depending on the limited resources available after the end of the Vialactea project), mostly when the VLKB resources kept being used in galactic astrophysics projects or as a comprehensive resource of data and services in technical demonstrator projects. Those projects provided the continuity in funding basic maintenance of the VLKB and some updates (even if occasionally rather than continuously).

With the first release of the VLKB in 2016, the subsequent maintenance gap spanning from then until 2020, and the restart of development since then, the current adoption of standards in the VLKB includes: an ObsCore table to keep the metadata for the observational datasets' catalogue, the TAP service to expose the general metadata content for all its data resources (catalogues, images, radial velocity cubes and morphological complex objects, ...), a custom implementation of the SODA (Server-side Operation for Data Access) standard set up to replace the dataset cutouts (with UWS - Universal Worker Service - used to manage asynchronous cutout and merge requests). Furthermore, authentication and authorization infrastructure (AAI) solutions using OAuth/OIDC have been tested on top of the cutout service, and a multi-cutout solution has been presented at VO level as a feedback for the SODA and DataLink evolution. Other features (management of complex morphology, of simulated data, and registration of the VLKB resources into the VO Registry) are still undergoing or missing. In particular, enabling more standard interface for the VLKB and making VLVA aware and able to consume them, will let both the client be more general and easier to maintain and the server resources be consumed by non-dedicated client applications.

This contribution reports the challenges in maintaining and improving the VLKB, the actual status of the technologies and standards in use for its resources, and the present and future perspectives for the VLKB itself.

Keywords: data management, galactic astrophysics, interoperability, standardisation

Contact author: Marco Molinaro, e-mail: marco.molinaro@inaf.it

Software and Cyberinfrastructure for Astronomy VIII, edited by Jorge Ibsen, Gianluca Chiozzi, Proc. of SPIE Vol. 13101, 131011W · © 2024 SPIE
0277-786X · doi: 10.1117/12.3020176

1. INTRODUCTION

The history of the development of the ViaLactea Knowledge Base (VLKB) dates 2013, when the VIALACTEA* project started. The initial phase of the project was dedicated to collecting requirements from the project community, Milky Way astrophysics researchers, to provide the basis for the design and development of the VLKB. After that, a set of data collections and catalogues was put together and an archived built from them, and services were set up providing the first working implementation of the VLKB [1].

Unfortunately, the end of the VIALACTEA project meant also the end of direct funding to improve the VLKB archive and services, leaving many loose ends and putting the VLKB in basic maintenance, with limited curation of the metadata and services, and little resources to widen data collections and catalogues.

The gap in development of the VLKB was, initially, partially filled by the needs coming from other projects, (NEANIAS†, CIRASA‡) not necessarily focused on astrophysical research but seeking for technical use cases, and, later on, closed (for the most part) by the ECOGAL§ project, again driven by galaxy astrophysics.

This proceeding reports on the actual status of the VLKB archived data and service, with particular respect to their move towards IVOA standards driven compatibility. Sec. 2 provides an overview of the historical evolution of the VLKB; Sec. 3 describes the status of the data and services currently in place, while Sec. 4 describes the foreseen and potential evolution of those same resources. Finally, sections 5 and 6 draw some conclusions based on what has been learned in managing the VLKB as a project and the challenges faced while implementing and upgrading it.

2. VLKB HISTORY

Figure 1 shows, sketchily, the evolution of the VLKB over time and the projects that supported its development and upgrade. It stops at the time of writing even if the ECOGAL support will probably continue for a couple of years more, and the hinted SKA-RC connection is currently ongoing (see later for further details, Sec. 3.1). Only the *1st release* and *VO standards 1st integration* are pointed out in the diagram but more upgrades, experimentation, evolution steps, and integration happened during this time span. This section is meant to describe these latter further.

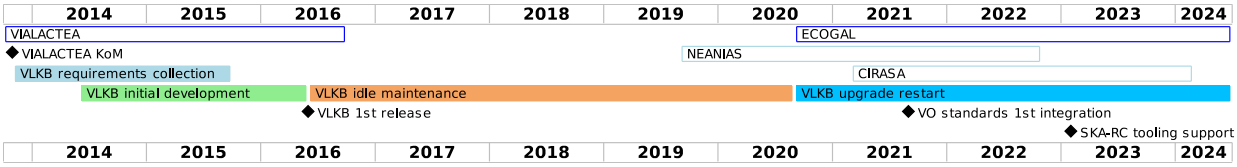

Figure 1. A sketchy overview of the ViaLactea Knowledge Base (VLKB) history.

2.1 Initial development

A great part of the development of the VLKB happened, obviously, during the lifetime of the VIALACTEA project, when the archive and services infrastructure requirements where initially defined and also a dedicated client application, the ViaLactea Visual Analytics (VLVA) was developed to specifically consume the VLKB contents (see [2] & [3] for the initial solutions with that respect.).

A good deal of time (about 2 years) was devoted to:

*VIALACTEA - The Milky Way as a Star Formation Engine - EU FP7 grant agreement n. 607380

†NEANIAS - Novel EOSC services for Emerging Atmosphere, Underwater and Space Challenges - EU H2020 grant agreement n. 863448. Website: https://www.neanias.eu/

‡CIRASA - Collaborative and Integrated enviRonment for Astronomical Source Analysis - PRIN-INAF TEC 2019. Website: https://www.oact.inaf.it/project/cirasa/

§ECOGAL - Understanding our Galactic ecosystem: From the disk of the Milky Way to the formation sites of stars and planets - H2020 ERC Synergy grant n. 855130. Website: http://www.ecogal.eu/

- gather community requirements;
- identify relevant data collections and catalogues to build the archive's database;
- identify potential existing solutions to match the requirements.

Since the beginning attempts were made at using IVOA standards[¶] for the modelling, discovery and access solutions of the VLKB. Given the requirements for:

- positional and energy band discovery and access;
- catalogue publishing and filtering;
- morphological sources storage and retrieval;

already from the first stages of the VLKB design, plans were put forward to use IVOA SIAv2, TAP, SODA (AccessData at the time), DataLink, MOC and connected standards (like UWS for asynchronous queries).

However the pressure for finalising the expected outcome of the project and the constraints coming from

- using galactic coordinates frames, instead of the mandatory ICRS one used in the IVOA Recommendations;
- dealing with velocity cubes, having a third axis not compatible with SODA wavelength cutout standard;

added to the lack of out-of-the-box tools to be integrated for the intended purposes, led to the decision of developing custom interfaces for the datasets search, cutout, and merge services. The only IVOA specification that made it since the beginning was a TAP server (still in place as of today) to provide catalogue data, dataset metadata and other table-like structures within the VLKB (it actually helped having in house a TAP implementation ready, even if a draft, not perfect one).

Not all the VO expertise available within the VIALACTEA project was anyway lost, because the decisions about how to design and implement the metadata archive, the architecture of the query, cutout and merge APIs and the programmatic behaviour of the interfaces were heavily influenced by knowledge available at IVOA community level. This is why, at the time of writing (see Sec. 3, but also Sec. 2.3) the IVOA Recommendations compliance has improved without re-building the VLKB architecture, but only improving metadata handling and interface alignment.

During these first 2 years a bulk of ∼8500 FITS files (images and cubes), subdivided in ∼10 collections (surveys and pointed observations) and ∼20 sub-surveys (sub-collections observed at different velocity bands), for a total of about 300GB of data, were collected and ingested in a prototype solution of the VLKB that already allowed discovery and access of catalogues and datasets.

Roughly at the same time, reports were given at Euro-VO related Forums and IVOA Interoperability Meetings, besides the internal VIALACTEA project meetings, about the advantages and disadvantages of the chosen solutions as well as feedback on VO standards. These reports, moreover, were showing features and aspects that continuously arose during further stages of the VLKB development, like:

- Authentication & Authorization (A&A) solutions for data policy requirements;
- tessellation (IVOA MOC) as a way to represent, store, retrieve and match source objects having different morphological complexity;
- IVOA DataLink solution for custom data access (especially for the datasets merge service).

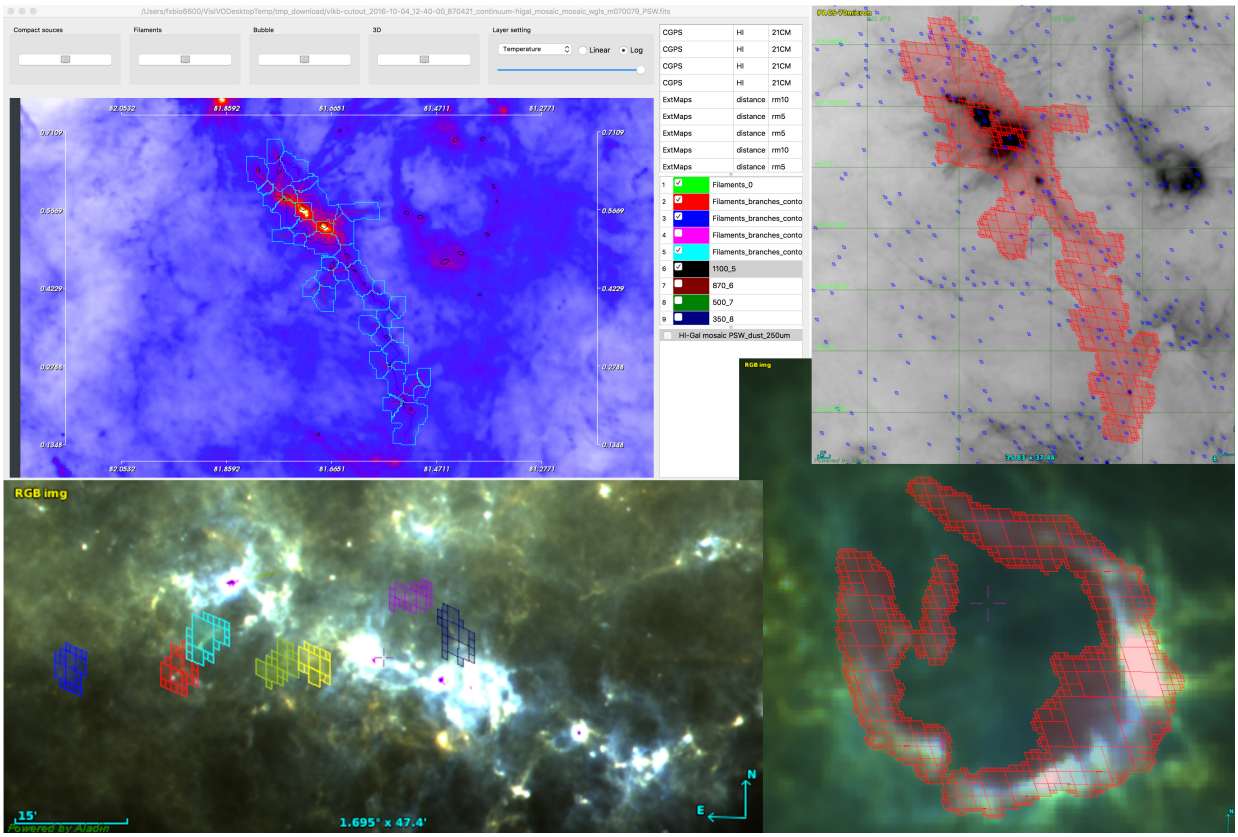

Figure 2. Exemplification of tests using IVOA MOC tessellation solutions to store and retrieve morphological complex objects into and from the VLKB. Top left image shows a first example of the ViaLactea Visual Analytics (VLVA) client accessing VLKB data. Top right image shows a tessel-rendered filament superimposed to compact sources from a catalogue. The bottom images show a set of *bubbles* (on the left) and a single *bubble* displayed as MOC tessels. The latter three images were generated using the CDS Aladin sky atlas tool.

Fig. 2 shows already, at the initial stages, storage and retrieval of MOC-based morphologies in the VLKB, while A&A solutions came only later on (see Sec. 2.3) and DataLink hasn't yet been put in practice (despite the merge interface being really similar in architecture).

After another year (from the internal VIALACTEA prototype release), the VLKB archive and services grew to a first release (fully described in [1]) that included (rough figures):

- 50 dataset collections;
- 30k individual datasets;
- 1TB storage content;
- 10 source catalogues;
- up to millions of rows;

and integrated a TAP server, and a set of APIs to perform discovery, access, cutout and merge on the dataset collections (see Fig. 3 for a quick overview of the interface parameters).

[¶] see https://ivoa.net/documents for details on the ones reported here

Other services were developed at the end of the VIALACTEA project, to serve specific needs to calculate distances of galactic sources and to let VLVA interpolate spectral energy distributions against numerical models. These were later, slowly upgraded but never reached a proper maturity level, apart from direct usability when consumed by dedicated clients.

	SEARCH	CUTOUT	MERGE	VALUES	DEFAULT
surveyname	√		√	<surveys table>	NONE
species	√		√	<surveys table>	NONE
transition	√		√	<surveys table>	NONE
pubdid		√		<search result provided>	NONE
skysystem	√	√	√	GALACTIC, ICRS	GALACTIC
l,b	√	√	√	0/360, -90/+90 [deg]	**MANDATORY**
r	√	√	√	0/2 [deg]	0 [deg]
dl,db	√	√	√	0/2, 0/2 [deg]	0, 0 [deg]
vl,vu	√	√	√	<dataset depending>	<full available range>
nullvals		√		flag key	not present

Figure 3. Description of the parameters used in the custom VLKB interfaces for the 1st release. The overlap of many of them when moving from the search to the cutout and merge interfaces was intentional and follows the SIA and SODA architecture of the IVOA Data Access Layer.

One, final, feature, that was proposed (but only tested as a prototype) during the initial implementation of the VLKB was a cross-match service for morphological complex structures (*filaments* using MOC tessellation and JSON objects directly stored in a relational database. This was abandoned both for lack of resources to complete it and also considering the (potentially oncoming) integration of MOC as a direct relational database type.

2.2 Idle maintenance

During the time span that goes from the end of the VIALACTEA project and the start of the ECOGAL support one only basic maintenance was possible for the VLKB. That means that services were kept running, critical bugs were fixed, and some attempt at improving the efficiency of some of the queries were made.

The actual improvements during this phase (most of them towards the end of it, thanks to the start of the NEANIAS project) relate to:

- software infrastructure changes to improve towards micro-servicing solutions;
- better insulation of services (that later eased containerisation) driven by mirror setups of the VLKB;
- tests for A&A solutions;
- increase of the amount of datasets and collections available through the VLKB (this is actually the only upgrade that was content related).

Besides that the VLKB content was used in demonstrations and also at VO level as feedback to source modelling when dealing with complex morphologies (again showing the potentials of the IVOA MIC standard).

2.3 Upgrade restarts

With the requirements coming from the ECOGAL project, in connection with the activities of the NEANIAS and CIRASA ones, the VLKB started again to be developed and improved. Roughly at the start of this stage, but overlapping the idle maintenance period, the VLKB contents nearly doubled (again, rough figures):

- 100 dataset collections;

- 40k individual datasets;
- 2TB storage content;
- 15 source catalogues;
- up to millions of rows;

That increase in size is not, however, what is most important to the VLKB development. In providing a partial mirror of the VLKB for the NAENIAS project, that required a set of technical improvements, the following features were put in place:

- the cutout system moved towards IVOA SODA compliance, including providing feedback to the IVOA community on the SODA standard itself;
- IVOA UWS became the asynchronous solution for VLKB data access (apart from direct download of bulk datasets);
- A&A solutions using token based authentication and authorization were set up against both the NEANIAS provided Authentication and Authorization Interface (AAI) and the INAF–IA2 (Italian center for Astronomical Archives[‖]) one

Other than these improvements related mostly to the NEANIAS project, the VLKB had resources also to move towards a better compliance with the IVOA standards architecture. Indeed VLKB now has:

- an ObsCore table that describes all the datasets ingested in the database;
- an (under testing phase) SIAv2 interface, alongside the custom search one, providing proper VOTable responses;
- a compliant SODA interface for the datasets cutout.

Moreover, this shift towards the IVOA scenario impacted also the development of the VLVA that integrated the PyVO package to connect properly to VO aware resources and is leading towards the simplification of catalogue access (using external service, where existing, and thus simplifying the maintenance of the VLKB data bundle).

The experience of the VLKB continues also to provide feedback to the VO community, like recently (May 2024) discussing the usage of *RESOURCE*s and *TABLE*s elements in the VOTable standard responses for the IVOA Data Access Layer.

3. CURRENT COMPONENTS

The possibility to restart improving the VLKB architecture and services was mainly used to improve some technical aspects and move the interfaces towards standardisation.

At present, the standards in use for the VLKB include:

- ObsCore table: metadata for the observational datasets catalogue;
- TAP service: general metadata content for all resources: catalogues, images, radial velocity cubes, morphological complex objects;
- SODA: dataset cutouts;
- UWS: asynchronous cutout and merge requests management;

[‖] https://ia2.inaf.it

- SIA: dataset discovery (currently under testing).

Besides the above, directly related to IVOA standards implementation, other standards and features have been investigated and put in practice:

- OAuth/OIDC AAI solutions have been tested on top of the cutout access service;
- a multi-cutout solution has been implemented (feedback to DataLink evolution);
- Resource content is under refactoring to take advantage of VLVA growing ability to query VO resources.

The evolution of the VLKB towards interoperable standards, as said, influenced also the VLVA client: moving both the VLKB and the VLVA towards IVOA (or other) standards will let both: the client-side be more general, and the server-side easier to maintain and enable VO-aware client applications to connect to its resources. Fig. 4 shows a few screenshots of the VLVA application, that continues to offer unique advantages to its users in the galaxy astrophysics research domain.

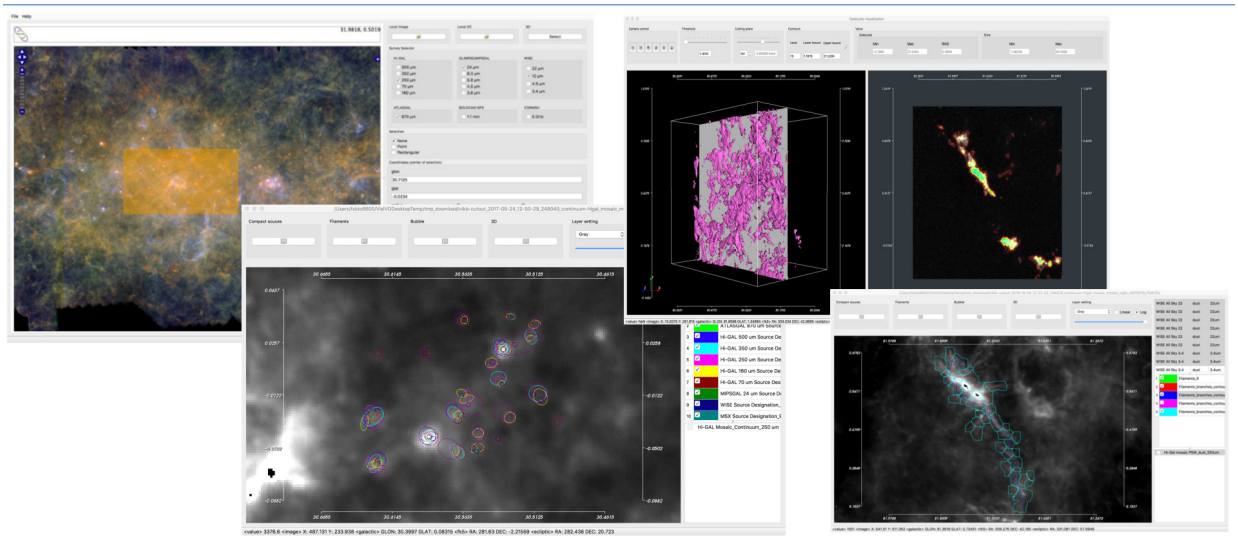

Figure 4. New requirements and support brought by the ECOGAL project allowed for upgrades of the interfaces, also in connection with the ViaLactea Visual Analytics (VLVA, see also [4]), the main client consuming the VLKB resources. Here above a few screenshots of the VLVA client application.

3.1 Involvement in other projects

The VLKB, besides serving the galaxy astrophysics community, has been used as a test solution for cloud solution porting and connection to federate authentication and authorization solutions (like within the NEANIAS project). Its SODA cutout component has also been used in SKA-RC Network prototyping activities and continues to evolve within that scenario.

The request for containerisation of the SODA development is leading towards a full containerisation of the VLKB services that would be helpful in re-deploying it, but also in re-using parts of its architecture in other contexts.

4. POSSIBLE EVOLUTION

Despite having improved recently, the VLKB still has quite a set of missing or incomplete features. Among them:

- management of complex morphology (probably using tessellation also to make it easy to provide cross-matches among them and with point-like source catalogue);

- management (and archiving) of simulated data. The latter need some further investigation in terms of requirements as well as available standards;
- registering the VLKB data collections and services as VO resources; The hope is that, having a stable set of resources and services, it will be easier to identify how to set up the connected set of potential resources;
- usage of DataLink as the solution for the merge service.

VLKB expertise and tools will anyway continue to be used in connection with SKA-RC activities and to contribute INAF IA2 data center as a discover and access API solution for general archive datasets and source catalogues. Finally, VLKB will continue to evolve following the changes and requirements in the VLVA.

5. LESSONS LEARNED

While designing and developing the VLKB architecture, components and working with the research community to provide good services for their data search and analysis, a few lessons were learned.

The most important one is possibly the need to spend a large amount of time in understanding each other's language, that is the only way to properly translate scientific research requirements in usable service components and features.

Language translation (between science and technology) is also time demanding when trying to provide a system that you want to be interoperable and based on community driven standards. This is a key feature nowadays, when FAIR principles play a big role in data and metadata management for research.

One other lesson is about allocating resources to reach sustainable solutions that can aim at long term preservation (here intended as being able to work when the hardware/software infrastructure inevitably change over time): already when facing the impossibility to directly use VO standards for the VLKB it was clear that this would have driven to a need to invest in standardisation later on to preserve the VLKB contents in the future. This is not a fault of the VIALACTEA project itself, it is only the common way to consider *ad hoc* solutions faster.

Finally, all possible smart interfacing solutions can easily break if not supported by proper metadata curation and content annotation. Following FAIR principles, even when not able to put them all at work, eases a lot these requirements.

6. CHALLENGES

In between lessons learned and challenges there's the need to disseminate the basics of metadata management. It is quite easy to underestimate the resources needed to do so, expecting that once the requirements are written down, the translation into proper management is easy. Data Management Plans are another nowadays common concept that fits into this challenge.

The obvious, ubiquitous challenge is in procuring suitable resources to build a system and foresee the ones needed for the maintenance of an infrastructure. The VLKB reported example is a quite common one, where resources are provided to build (part of) an infrastructure, through an initial competitive call, followed by uncertain resources for its sustainability on the long term, despite the interest of the community in keeping it alive.

Another foreseeable challenge is in trying to prioritise all the potential features that can come up in mind while working with data content. The VLKB has quite a number of loose ends that are still there due to a not optimal prioritisation on the server side.

There's also a challenge in foreseeing what software tools can be developed over time, leading to an ease or a complication in the development of one or another of the wanted features.

Finally, a challenge, again related to resources, is in keeping alive and transmit the knowledge needed to run and maintain the developed infrastructure, i.e. it should always be a collaborative effort.

ACKNOWLEDGMENTS

This work, and all the historical development reported, were possible thanks to the resources and support of the following projects:

- VIALACTEA - The Milky Way as a Star Formation Engine - EU FP7 grant agreement n. 607380 (it all started from here);
- ECOGAL - Understanding our Galactic ecosystem: From the disk of the Milky Way to the formation sites of stars and planets - H2020 ERC Synergy grant n. 855130 (allows for the infrastructure to be kept alive);
- CIRASA - Collaborative and Integrated enviRonment for Astronomical Source Analysis - PRIN-INAF TEC 2019 (helped in securing resources and try new developments);
- NEANIAS - Novel EOSC services for Emerging Atmosphere, Underwater and Space Challenges - EU H2020 grant agreement n. 863448 (helped in providing resources and technical solutions).

REFERENCES

[1] 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., and Vitello, F., "VIALACTEA knowledge base homogenizing access to Milky Way data," in [*Software and Cyberinfrastructure for Astronomy IV*], Chiozzi, G. and Guzman, J. C., eds., *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **9913**, 99130H (Aug. 2016).

[2] Molinari, S., Butora, R., Cavuoti, S., Molinaro, M., Riccio, G., Sciacca, E., Vitello, F., Becciani, U., Brescia, M., Costa, A., and Smareglia, R., "Integrated data access, visualization and analysis for Galactic Plane surveys: the VIALACTEA case," in [*Astroinformatics*], Brescia, M., Djorgovski, S. G., Feigelson, E. D., Longo, G., and Cavuoti, S., eds., **325**, 291–298 (June 2017).

[3] Vitello, F., Sciacca, E., Becciani, U., Costa, A., Bandieramonte, M., Benedettini, M., Brescia, M., Butora, R., Cavuoti, S., Di Giorgio, A. M., Elia, D., Liu, S. J., Molinari, S., Molinaro, M., Riccio, G., Schisano, E., and Smareglia, R., "Vialactea Visual Analytics Tool for Star Formation Studies of the Galactic Plane," **130**, 084503 (Aug. 2018).

[4] Tudisco, G., Vitello, F., Sciacca, E., Riggi, S., Molinari, S., Malikova, E., and Krokos, M., "ViaLactea: a distributed Visual Analytic system for exploring our Galactic ecosystem," in [*Astronomical Society of the Pacific Conference Series*], Hugo, B. V., Van Rooyen, R., and Smirnov, O. M., eds., *Astronomical Society of the Pacific Conference Series* **535**, 211 (May 2024).