Astrophysics Visual Analytics services on the European Open Science Cloud

Eva Sciacca[a], Fabio Vitello[b], Simone Riggi[a], Ugo Becciani[a], Cristobal Bordiu[a], Filomena Bufano[a], Robert Butora[c], Alessandro Costa[a], Sergio Molinari[d], Marco Molinaro[c], Carmelo Pino[a], and Eugenio Schisano[d]

[a]INAF, Osservatorio Astrofisico di Catania, Via S Sofia 78, I-95123 Catania, ITALY
[b]INAF, Istituto di Radioastronomia, Via Gobetti 101, I-40129 Bologna, ITALY
[c]INAF, Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, I-34143 Trieste, ITALY
[d]INAF, Istituto di Astrofisica e Planetologia Spaziali di Roma, Via del Fosso del Cavaliere, 100, 00133 Roma, ITALY

ABSTRACT

The European Open Science Cloud (EOSC) aims to create a federated environment for hosting and processing research data, supporting science in all disciplines without geographical boundaries, so that data, software, methods and publications can be shared seamlessly as part of an Open Science community. This work presents the ongoing activities related to the implementation and integration into EOSC of Visual Analytics services for astrophysics, specifically addressing challenges related to data management, mapping and structure detection. These services provide visualisation capabilities to manage the data life cycle processes under FAIR principles, integrating data processing for imaging and multidimensional map creation and mosaicking and data analysis supported with machine learning techniques, for detection of structures in large scale multidimensional maps.

Keywords: Cloud Computing, European Open Science Cloud, Visual Analytics, Deep Learning

1. INTRODUCTION

The European Open Science Cloud[1] (EOSC) is materialising the Open Science paradigm in Europe, providing researchers with a federated platform including fit-for-purpose services, developed and operated by European research institutions. In this context, the H2020 NEANIAS project[2] is driving the co-design, delivery, and integration into EOSC of innovative thematic services, derived from state-of-the-art research assets and standard practices in three major sectors: underwater, atmospheric and space research.

New approaches to data processing, archiving, analysis and visualisation are nowadays mandatory to deal with the data deluge expected in next-generation facilities for astronomy, such as the Square Kilometer Array[3] (SKA). The unprecedented data volumes produced by these facilities will demand not only enhanced visualisation capabilities but also efficient extraction of meaningful knowledge, allowing for breakthrough discoveries. To address such needs, Visual Analytics (VA) has emerged as a novel approach to support interactive data exploration and analysis [4].

This paper presents the first release of the NEANIAS Visual Analytics services for (1) management and visualisation of astrophysics data adopting FAIR (Findable, Accessible, Interoperable, Re-usable) principles, (2) assembling those images into custom mosaics (mosaicking), and, finally, (3) to perform pattern and structure detection in astronomical surveys.

Section 2 presents the NEANIAS Astrophysics user communities and the user-requirements collected to guide the software development. Section 3 describes the NEANIAS Visual Analytics services and their delivery processes. The NEANIAS release procedures are reported in Section 4. Finally, Section 5 outlines the conclusions and future developments.

Further author information: (Send correspondence to E. Sciacca)
E. Sciacca: E-mail: eva.sciacca@inaf.it, Telephone: +39 095 7332321

End-User	User-requirement	Reason
Astrophysicist	To analyse maps at different wavelengths	To account for different emitting components in the same astrophysical source from the same UI or service
Astrophysicist	To reduce images to the same technical features	To compare images from different surveys
Astrophysicist	To enhance current source finding algorithms and software	To identify, classify and characterize compact and extended sources with a minor impact of artefacts and spurious detection
Software engineer	To improve the software's portability	To easily share software and code within the community not being limited to available local computing resources
Astrophysicist	To improve the algorithms' reproducibility	To share the acquired knowledge and allow upgraded experiments
Astrophysicist	To access all observations and catalogues available for a certain sky region and epoch	To have a dataset to perform data analysis or train and validate analysis algorithms
Software engineer	To integrate existing data access and processing components	To create a more complex integrated system capable of producing new scientific information

Table 1. NEANIAS User Requirements for Space Research.

2. NEANIAS ASTROPHYSICS COMMUNITY AND USER REQUIREMENTS

Space end-user communities within NEANIAS mainly include astrophysics and planetary scientists, as well as computer scientists and software engineers interested in computer vision and machine learning. Astrophysics scientists are primarily focused on Radio astronomy studying data from large radio interferometers (including SKA) and on Infrared astronomy studying the processes of star formation in our Galaxy.

An initial set of user requirements for the development of the services was collected by a group of experts within the project, using the paradigm of 'User Stories' (see Table 1) in order to capture the description of software features from end-user perspectives. A User Story describes, in a concise way, the type of user, what they need to achieve and why. This high-level approach helps to create a simplified description of a requirement that can be subsequently transcribed into development specifications. Later, these user stories were validated by means of a targeted survey about the needs and trends within the astrophysics community, involving more than 300 professionals from different research institutions across Europe.[5]

The final user requirements, summarised in table 1, can be translated into the following specific goals:

Increase multi-wavelength studies. Multi-wavelength astronomy is the study of the emission of a particular object throughout the entire electromagnetic spectrum, i.e. covering a wavelength range as wide as possible to better disentangle and analyse its different components and physical conditions. This kind of studies is gaining relevance in recent years, with more than 60% of the surveyed professionals perceiving it as crucial for future discoveries. Effectively supporting multi-wavelength astronomy implies two tasks: i) access different public surveys and data archives, subject to their specific query formats; and ii) manage files with different kinds of technical features (e.g. dimensions, resolution). Thus, a service offering direct data access plus multidimensional image creation capabilities is necessary, both to speed up research and to offer the scientific community a complete knowledge database.

Seamlessly compare different surveys. Dealing with different survey images poses a major technical problem: how to compare images and maps with different dimensions, spatial resolution, pixel size, or flux units? A tool able to manipulate images in order to make them directly comparable represents a service of crucial importance for the community.

Improve source finding. Most of the current source finding algorithms and tools have non-negligible false-positive rates, especially when processing maps with complex backgrounds, which could be the case for the Galactic Plane. Ideally, all these artefacts and spurious sources should be removed from the final source catalogues. As of today, this curation process relies on manual procedures (e.g. visual inspection) in almost all surveys. This approach is error-prone, not reproducible, time-consuming and unfeasible at the scale of SKA. Likewise, existing source finders do not provide the astronomical identity of the extracted sources and, when working in the low signal-to-noise regime, compact source detection performances are abysmal, e.g. parameter estimates (e.g. source flux and position) are biased due to inaccurate background subtraction, deblending and fitting. In this regard, Machine Learning may help in: i) discriminating genuine sources from false/spurious detections in radio catalogues automatically generated by existing source finders; ii) identifying astronomical object classes of the extracted compact sources; and iii) outperforming detection and characterisation capabilities of existing finders, thus improving catalogues' completeness and reliability.

Improve portability, distribution and performance. Many software packages and tools employed to study astrophysical phenomena are mostly executed in local laptops and PCs, and only shared among few researchers, often within the same research group. The overwhelming data volumes expected in next-generation telescopes will render local processing unfeasible. It is then clear that computing infrastructures should approach data archival facilities to minimise costs and facilitate data transfer and result replication.

Improve reproducibility. In every science field, the reproducibility of the results obtained is fundamental to prove their own reliability. Astronomy is producing data at an unmatched rate. The installation of new telescopes, combined with marked improvements in pattern-finding algorithms, has led astronomers to turn to sophisticated software to do the data-crunching they cannot do manually. With more complex analyses, there is less transparency concerning how they have been performed. Fully reproducible research would require the publication of detailed procedures carefully addressing every step done in data processing and analysis, making all the intermediate and final products available for peer-review. Result reproducibility is indeed a major concern for the astrophysics community, identified as a significant barrier for Open Science by 59% of the survey participants. Once the scientific community can freely glean not only to the same data archive but also to all published software and algorithms, then we will reach the optimal degree of reproducibility, giving everybody the tools to test any experiment and validate its results, in a sort of worldwide laboratory.

Improve data access. In Astrophysics, datasets are described using different metadata, usually with different formats and stored in diverse locations. This situation complicates datasets exploitation, making it complex and tedious. There is the need to facilitate the data access through a common approach based on a common dataset description, standards for data formats and efficient and distributed data stores. While there have been significant steps in this direction with the spread of Virtual Observatory tools, there is still a long way to go. The community perceives data accessibility as one of the principal challenges for astronomy in the next decade.

Improve integrability. Apart from being used as standalone space services, the existing applications within the NEANIAS landscape can also be efficiently combined to create new services of extended functionality, integrating single space services as elements of a higher level pipeline. For this to be accomplished, the space service architecture has to be designed with this requirement in mind, relying on solid communication standards (e.g. REST) to ensure the interoperability of all its constituent pieces.

For further technical information about the effective implementation of these requirements, we encourage the reader to consult [6].

3. NEANIAS VISUAL ANALYTICS SERVICES

NEANIAS is developing a wide service portfolio addressing the current needs in the astronomical data lifecycle and eventually laying the foundations to overcome the scientific and technological challenges expected in the next decade. The services providing Visual Analytics capabilities are:

SPACE-VIS The *FAIR Data Management and visualisation services* deliver an advanced operational solution for data management and visualisation for space FAIR data. They provide tools that enable an efficient and scalable visual discovery, exposed through advanced interaction paradigms also exploiting virtual reality.

SPACE-MOS The *Map Making and Mosaicking of Multidimensional Space Images services* deliver a user-friendly cloud-based version of the already existing workflow for map-making and mosaicking of multidimensional map images. They create multidimensional space maps through novel mosaicking techniques to a variety of prospective users/customers (e.g., mining and robotic engineers, mobile telecommunications companies, space scientists).

SPACE-ML The *Structure Detection on Large Scale Maps with Machine Learning services* provide cutting-edge solutions for detection of compact (e.g. stars, galaxies) and extended sources (e.g. supernovae remnants,

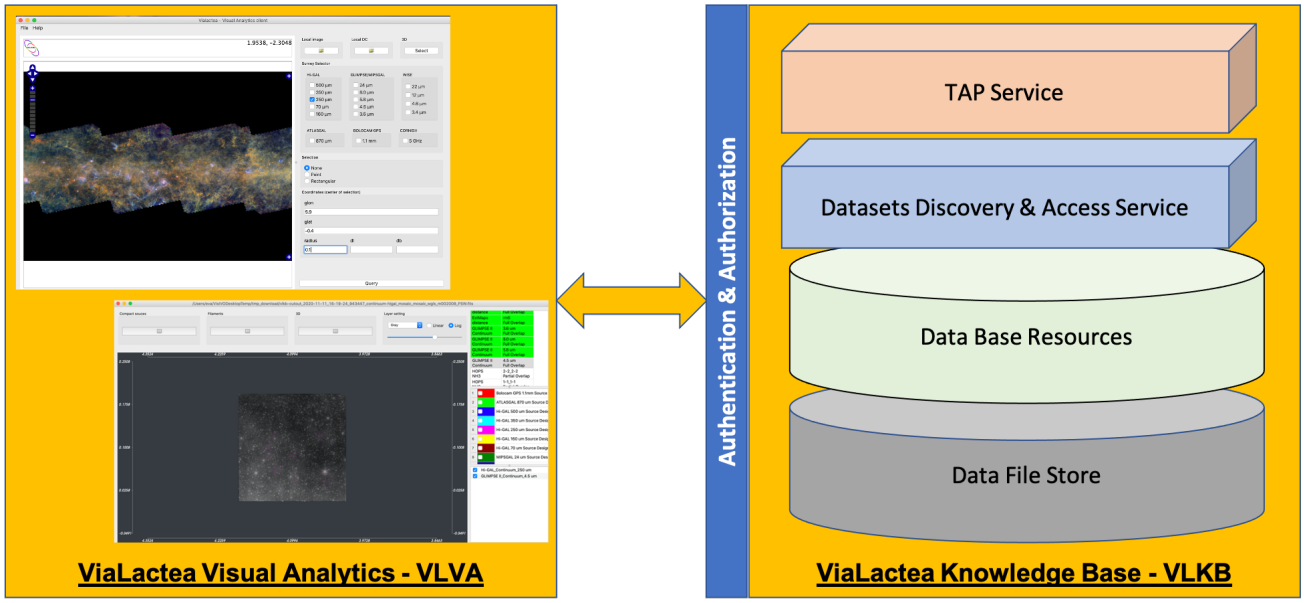

Figure 1. ViaLactea Visual Analytics (VLVA) application interacting with the ViaLactea Knowledge Base (VLKB) exposing the contained file store and metadata trough the TAP service and the datasets discovery & access service

galactic filaments) in infrared and radio maps, exploiting state-of-the-art machine learning algorithms. These services are meant to support astronomers in the extraction, classification and characterisation of astronomical sources in future all-sky surveys.

3.1 SPACE-VIS: ViaLactea Service

The ViaLactea Service (see Figure 1) is aimed at exploiting astrophysical surveys of the Galactic Plane focused on studying the star formation processes of the Milky Way. The ViaLactea Visual Analytic (VLVA) tool [7] combines different types of visualisation to perform the analysis exploring the correlation between the different data managed in the ViaLactea Knowledge Base (VLKB) services and resources [8] including 2D and 3D (velocity cubes) surveys, numerical model outputs, point-like and extended object catalogues.

The ViaLactea Visual Analytic tool allows for an integrated analysis of compact sources and their spectral energy distributions (SEDs), studies on diffuse bubble-like structures, and analysis of filamentary structures exploiting the new-generation multi-wavelength infrared and radio surveys of the Galactic Plane. It employs a novel data and science analysis paradigm based on 2D/3D visual analytics and data mining frameworks. VLVA is implemented as a cross-platform application, seamlessly interacting with VLKB, as shown in Figure 1. The core functionalities of VLVA are implemented in C++, extended with Qt[9] for graphical user interface and the Visualization Toolkit[10] (VTK) for rendering.

The ViaLactea Knowledge Base is a file and metadata storage utility currently providing access to: (1) 2D continuum maps; (2) 3D molecular line cubes; and (3) 3D extinction maps of the Galaxy. Apart from the data collections, a relational database completes the resource content in terms of knowledge derived from the analysis of the map, with information about the sources present in the archives and their parameters. All the data and metadata provided by VLKB can be readily accessed by Virtual Observatory tools and services by means of the standard Table Access Protocol (TAP).

3.2 SPACE-MOS: Astro MapMerging Service

The Astro MapMerging service is built upon Montage.[11] It is currently integrated into the ViaLactea Knowledge Base (VLKB), being part of the dataset discovery service and providing capabilities to merge contiguous datasets into a single image/cube, based on positional and velocity constraints. Besides, Scutout, a tool to extract scientifically comparable cutouts of astronomical sources from multi-wavelength FITS images is currently under development.

3.3 SPACE-ML: Source Detection Service

The Source Detection Service is based on CAESAR [12, 13]. It can extract and parameterise both compact and extended sources from astronomical radio interferometric maps. The processing pipeline consists of a series of independent stages for both compact and extended source detection: (i) Compact sources are extracted with flood-fill and blob finder algorithms, processed and fitted using a 2D gaussian mixture model; (ii) for extended sources, the procedure involves image pre-filtering, denoising, compact source identification and removal, enhancement of diffuse emission and final segmentation. These stages can run in a distributed, parallelised environment, employing multiple cores and processors. The outputs of the service include source fitted parameters, regions and contours. The workflow is being enriched with a separate Machine/Deep Learning stage that improves source identification, classification and characterisation in large-scale radio surveys.

4. SERVICES DELIVERY ON EOSC

In NEANIAS, services are built on top of TRL6 software solutions (i.e. software which is fully functional in its original domain) with the aim of reaching TRL8 (fully operational in a cloud environment). NEANIAS thematic services are integrated with a series of core services aimed at filling current gaps to offer operational services into EOSC. Core services can exploit other services that exist in the EOSC ecosystem already to form the basis of additional value chains in interdisciplinary research and business. Core services are grouped in distinct clusters addressing: a) research lifecycle empowering services, providing essential tools for registering, locating, inspecting, sharing data and services, b) integration services facilitating the exploitation of EOSC and research infrastructures offerings in a unified manner, including access to storage and computation and mechanisms for authentication and authorisation, c) Artificial Intelligence services, building capacity for Machine Learning and, finally, d) visualisation services delivering state-of-the-art visualisation capabilities.

The first release aims at deploying the software in the context of the NEANIAS infrastructure, using cloud infrastructure and services provided by GARR,[14] to be used as a test run for the development of NEANIAS services.

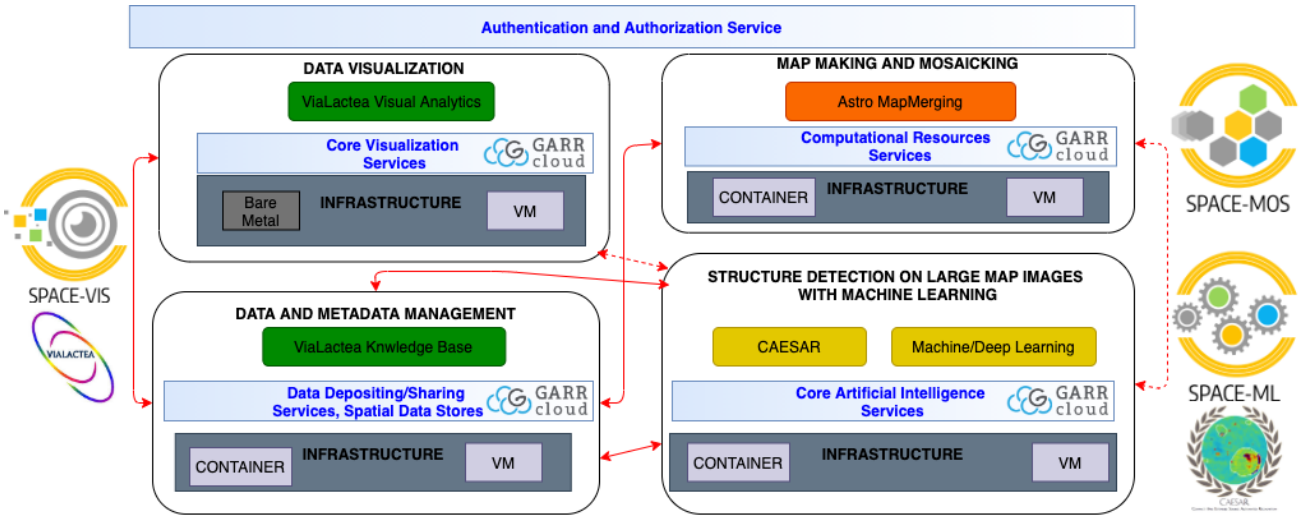

Figure 2. NEANIAS Visual Analytics Services delivery and integration with core services.

Figure 2 shows the Visual Analytics Services integration with NEANIAS core services: authentication and authorisation, computational and data resources access, visualisation and artificial intelligence.

So far, the ViaLactea Service has been released as a distributed system including the VLVA desktop application and the VLKB data service. VLVA has been updated with latest UI and visualisation libraries and extended to work on Unix-based Operative Systems. VLKB has been deployed on the GARR Cloud Platform Service and has been integrated with the NEANIAS AAI Service to allow authorised access to specific surveys based on their original policy. Similarly, the AstroMapMerging service is provided through the VLKB.

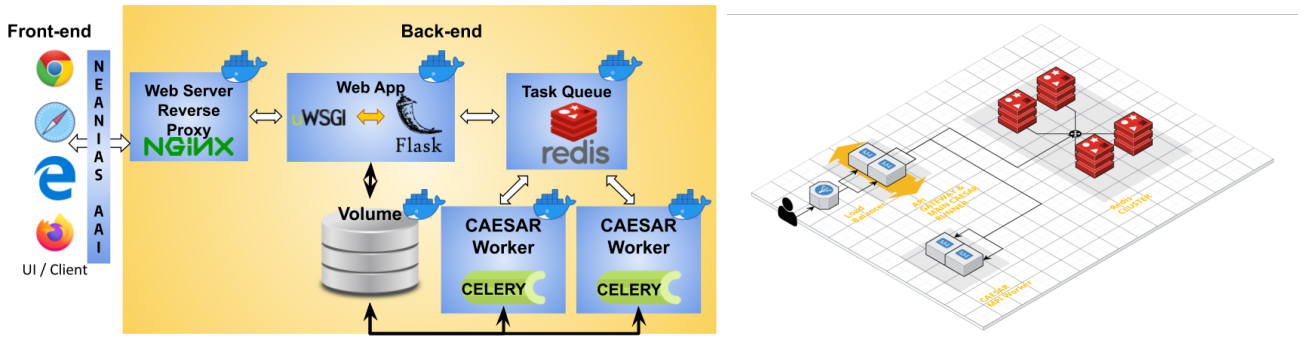

Figure 3. NEANIAS CAESAR Service delivery with REST-API and connection to the NEANIAS AAI (left figure) and deployment on the Kubernetes cluster (right figure)

The CAESAR Service has been deployed on the GARR Cloud interfaced with REST-APIs based on the Flask python framework[15] as shown in Figure 3. Celery[16] task queue is used to execute the CAESAR application jobs asynchronously, and it has been configured to use a RabbitMQ[17] broker for message exchange and Redis[18] as task result store. In the production environment CAESAR service run behind Nginx[19] and uwsgi[20] server connected to the NEANIAS AAI for authentication and authorisation processes. The service has been tested on the GARR Cloud Container Platform,[21] based on Kubernetes[22] to deploy the CAESAR tool in a modern cloud computing environment while maintaining support for the HPC requirements of the software. The deployment requires a Redis cluster, also used as a global store of currently accessible running containers, and is conducted using Helm Charts. Since MPI is a technology that is not natively supported by the Kubernetes cluster architecture paradigm, an open-source project called Kube-OpenMPI[23] was employed as the foundation for the Kubernetes CAESAR cluster.

The Machine/Deep Learning Service has been integrated within the NEANIAS Core AI Services providing facilities offering features of a typical Machine Learning (ML) workflow lifecycle, and that is also intended for the composition of higher-level services. In particular, the AI Science Gateway service[24] has been deployed for the development of ML/DL models using JupyterHub,[25] allowing for code implementation and interpretation using an IPython web interface.

The NEANIAS space thematic services are reachable through the NEANIAS space portal[26] designed for optimal community engagement with project outcomes for both specialist (e.g. astronomers and planetary scientists) and non-specialist target audiences. The webpages have been developed with HTML5, CSS and JavaScript and include the RSS feed from the NEANIAS Portal to collect news, blog articles and events related to the Space research sector. All documentation related to the Space services has been delivered on the NEANIAS documentation repository[27] and the relevant source code is handled by the NEANIAS GitLab repository[28] eventually mirroring some repositories from the GitHub.

5. CONCLUSION AND FUTURE WORKS

The NEANIAS project is prototyping innovative thematic Space services, driving the co-design, delivery, and integration into EOSC. In this paper, we have briefly introduced the first release of visual analytics services for the astrophysics community, enumerating the user requirements and needs that have driven their development, and outlining the methodology for their delivery. Future works involve further refinement of user requirements based on the results of [5], as well as two additional software releases, addressing a complete materialisation of the EOSC ecosystem.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission Horizon 2020 research and innovation programme under the grant agreement No. 863448 (NEANIAS).

REFERENCES

[1] "European Open Science Cloud (EOSC)." https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/eosc_en (2020). Accessed: 2020-11-30.

[2] "NEANIAS web portal." https://www.neanias.eu/ (2020). Accessed: 2020-11-30.

[3] "SKA web portal." https://www.skatelescope.org/ (2020). Accessed: 2020-11-30.

[4] Yi, J. S., ah Kang, Y., and Stasko, J., "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics* **13**(6), 1224–1231 (2007).

[5] Bordiu, C. et al., "Astronomical research in the next decade: trends barriers and needs in data access management visualization and analysis," in [*ADASS XXX*], Ruiz, J.-E. and Pierfederici, F., eds., *ASP Conf. Ser.*, ASP, San Francisco (2021).

[6] Sciacca, E., Topa, E., et al., "NEANIAS Deliverable D4.1 Space Research Services Report on Requirements, Specifications & Software Development Plan.," tech. rep., H2020 NEANIAS Project (06 2020).

[7] Vitello, F., Sciacca, E., Becciani, U., Costa, A., Bandieramonte, M., Benedettini, M., Brescia, M., Butora, R., Cavuoti, S., Di Giorgio, A., et al., "Vialactea visual analytics tool for star formation studies of the galactic plane," *Publications of the Astronomical Society of the Pacific* **130**(990), 084503 (2018).

[8] Molinaro, M., Butora, R., Bandieramonte, M., Becciani, U., Brescia, M., Cavuoti, S., Costa, A., Di Giorgio, A. M., Elia, D., Hajnal, A., et al., "Vialactea knowledge base homogenizing access to milky way data," in [*Software and Cyberinfrastructure for Astronomy IV*], **9913**, 99130H, International Society for Optics and Photonics (2016).

[9] "Qt Framework." https://www.qt.io/ (2020). Accessed: 2020-11-30.

[10] "Visualization Toolkit (VTK)." https://vtk.org/ (2020). Accessed: 2020-11-30.

[11] "Montage: An Astronomical Image Mosaic Engine." http://montage.ipac.caltech.edu/ (2020). Accessed: 2020-11-30.

[12] Riggi, S., Ingallinera, A., Leto, P., Cavallaro, F., Bufano, F., Schillirò, F., Trigilio, C., Umana, G., Buemi, C. S., and Norris, R. P., "Automated detection of extended sources in radio maps: progress from the scorpio survey," *Monthly Notices of the Royal Astronomical Society* **460**(2), 1486–1499 (2016).

[13] Riggi, S., Vitello, F., Becciani, U., Buemi, C., Bufano, F., Calanducci, A., Cavallaro, F., Costa, A., Ingallinera, A., Leto, P., et al., "C aesar source finder: Recent developments and testing," *Publications of the Astronomical Society of Australia* **36** (2019).

[14] "GARR Cloud Platform." https://cloud.garr.it/ (2020). Accessed: 2020-11-30.

[15] "Flask lightweight web application framework." https://palletsprojects.com/p/flask/ (2020). Accessed: 2020-11-30.

[16] "Celery - Distributed Task Queue." https://docs.celeryproject.org/en/stable/index.html (2020). Accessed: 2020-11-30.

[17] "RabbitMQ message broker." https://www.rabbitmq.com/ (2020). Accessed: 2020-11-30.

[18] "Redis data structure store." https://redis.io/ (2020). Accessed: 2020-11-30.

[19] "Nginx web server." https://www.nginx.com/ (2020). Accessed: 2020-11-30.

[20] "uWSGI project." https://uwsgi-docs.readthedocs.io/en/latest/ (2020). Accessed: 2020-11-30.

[21] "GARR Cloud Container Platform." https://cloud.garr.it/support/kb/kubernetes/ (2020). Accessed: 2020-11-30.

[22] "Kubernetes." https://kubernetes.io/ (2020). Accessed: 2020-11-30.

[23] "kube-openmpi: Open MPI jobs on Kubernetes." https://github.com/everpeace/kube-openmpi (2020). Accessed: 2020-11-30.

[24] "NEANIAS AI Science Gateway." https://docs.neanias.eu/projects/c3-1-ai-gateway/en/latest/ (2020). Accessed: 2020-11-30.

[25] "JupyterHub." https://jupyter.org/hub (2020). Accessed: 2020-11-30.

[26] "NEANIAS Space portal." https://thematic.dev.neanias.eu/SPACE/ (2020). Accessed: 2020-11-30.

[27] "NEANIAS documentation repository." https://docs.neanias.eu/en/latest/#space-services (2020). Accessed: 2020-11-30.

[28] "NEANIAS GitLab repository." https://gitlab.neanias.eu/ (2020). Accessed: 2020-11-30.