URSI GASS 2021, Rome, Italy, 28 August - 4 September 2021

The INAF Radio Data Archive: towards a modern Science Gateway

A. Zanichelli*(1), C. Knapic(2), E. Londero(2), S. Zorba(2), F. Bedosti(1), S. Righini(1), M. Nanni(1), M. Stagni(1), A. Orlati(1), F. Tinarelli(1), M. Sponza(2), R. Smareglia(2), A. Fara(3), S. Poppi(3), R. Concu(3), M. Molinaro(2), A. Bignamini(2), A. Costa(4), E. Egron(3), V. Galluzzi(2), K.-H. Mack(1), A. Pellizzoni(3), F. Vitello(4)
(1) INAF - Istituto di Radioastronomia, via Gobetti 101, I40129 Bologna, Italy
(2) INAF - Osservatorio Astronomico di Trieste, via G.B. Tiepolo 11, I34143 Trieste, Italy
(3) INAF - Osservatorio Astronomico di Cagliari, via della Scienza 5, Selargius, I09047 Cagliari, Italy
(4) INAF - Osservatorio Astrofisico di Catania, via S.Sofia 78, 95123 Catania, Italy

Abstract

In the Big Data era, the amount and complexity of astronomical data more and more often prevents the scientist from locally store and process her/his data. As a consequence, the geographically distributed approach to data archiving and processing is rapidly becoming a requisite. To fulfill this need, we are realizing a prototype of Science Gateway (SG) for the Italian radio telescopes. The huge amount of significantly complex and resource-demanding datasets delivered by the Italian radio telescopes and the variety of use cases from the different observing modes represent an ideal test bed for the implementation and verification of a SG environment where the scientists can exploit, manage and analyse data. To this aim, we are exploiting our previous experience in the realization of a geographically-distributed radio data archive and processing tools as well as in the design of SG prototypes. Such a coordinated approach and harmonization of resources will maximize the return for the Italian observing facilities and, moreover, will match the requirements of the international community for a state-of-the-art, highly-performant environment in which to conduct successful science.

1 Introduction

The advent of new generation telescopes and instrumentation is radically changing the approach to the scientific exploitation of astronomical observations. The amount of data produced even in a single observation is going to prevent in most cases the possibility for the user to download her/his data locally. Moreover, the computing resources required to process such huge datasets are incompatible with even the most performant personal computer and in some cases also with small-to-medium scale local computing clusters. Thus, adequate processing solutions are mandatory, based on the concept of moving software to the data. On top of this, data curation and preservation as well as the application of the FAIR (Findable, Accessible, Interoperable, Reusable) principles are concepts nowadays largely accepted. Also, the modern multi-wavelength/multi-messenger approach demands for the possibility to interoperate between Data Centers that store different datasets, often composed by massive amounts of data and using different data models and data formats. FAIRness of data resources and services can be met by means of the technological standards and architecture provided by the International Virtual Observatory Alliance (IVOA) that allow data coming from different projects at different wavelengths to be retrieved, reduced and analysed using the same approach and possibly the same tools. This complex landscape is motivating the international community to design and implement state-of-the-art, VO-aware Science Gateways (SG) to guarantee the accessibility of data as well as advanced resources for their processing and interpretation. Examples of SGs have been already implemented or prototyped, for instance at the Canadian Astronomical Data Center[1], the Centre de Données astronomiques de Strasbourg[2], the European Southern Observatory[3] and, in Italy, the Space Science Data Center[4] and the INAF Italian Astronomical Archives[5] infrastructure. Most of the existing SG architectures are still in the process of increasing their capabilities in the re-processing and/or uploading of user data by means of dedicated User Spaces, a crucial aspect to maximize the effective scientific exploitation of archival resources. In particular, the combination of resources, user-friendly applications and specialized tools within a SG infrastructure is fundamental to ease the exploitation of highly specialised datasets like those coming from modern radio telescopes.

The change of paradigm in data handling, processing and analysis does not affect only international radio astronomical facilities like for instance the Square Kilometre Array (SKA) and its precursors and pathfinders. The most modern digital instrumentation on board of the Italian radio telescopes is already demanding for a similar new approach for the successful scientific exploitation of data. To this aim, we have developed a state-of-the-art VO-aware data archive for the Italian radio telescopes and are implementing a SG

[1] CADC, https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/
[2] CDS, http://cdsweb.u-strasbg.fr/
[3] ESO, http://archive.eso.org/scienceportal/home
[4] SSDC, http://www.ssdc.asi.it/
[5] IA2, http://ia2.inaf.it

prototype focused in the radio domain. Indeed, the variety of use cases represented by the complex and resource-demanding datasets produced by the Italian radio telescopes for a number of different observing modes represents an ideal test bed for the implementation and verification of a SG environment where the scientists can exploit, manage and analyse astronomical data.

In Section 2 we describe the INAF radio data archive main features and architecture, while the design and implementation of the radio SG are illustrated in Section 3.

2 The INAF radio data archive

INAF manages three fully-steerable radio telescopes: the 32-m Medicina and Noto antennas and the more recent 64-m Sardinia Radio Telescope (SRT). They operate as single-dish telescopes or in coordinated interferometric networks also with other international radio facilities. Data produced by the spectropolarimeters on board of these radio telescopes may easily reach the size of 0.5 TB for a single file and are saved in a web-based, publicly accessible data archive [1].

The radio data archive is capable to host and handle the different data formats produced by the available observing modes. Single-dish, interferometry and pulsar data are stored in different formats based on the FITS [2] standard. The single-dish and interferometric data types have been considered as the reference to build a general, MBFITS-based [3] database datamodel to be used as a baseline for the creation of the radio archive database. The generic structure of the MBFITS-based database makes it capable to handle radio data written in non-hierarchical FITS format as well, in order to serve a vaster range of users/instruments. With the start of the scientific operations at the SRT in 2016, pulsar observations become a commonly offered observing mode and data needed to be archived as well. The PSRFITS data format [4] used for SRT pulsar observations is easily handled by the MBFITS-based radio archive datamodel. Aiming at persistence and future scientific exploitation of data, ancillary information contained in the observing schedules and telescopes/correlator logs are archived for each dataset as well.

The flexibility in importing and handling scientific data coming from different instrumentation and telescopes is obtained thanks to a TANGO-based architecture and configurable software [5]. Software modularity also guarantees that new instruments can be easily integrated at runtime, provided that the output data can be represented with the MBFITS-based data model. The scalability of the management system for data archiving allows independent finalization and ingestion of data locally at the observing sites, distributing the storage in a number of data centres. This feature matches the geographically distributed nature of the three Italian antennas and is fundamental to overcome the difficulties implied by the transfer of huge data volumes to a centralized data storage.

In the Open Science era archive users include scientists (with different level of knowledge in the radio astronomy field) and telescope staff but also the general public. Different types of permissions on the data are thus to be applied. Following the INAF policy, during the proprietary period data can be accessed only by the PIs of the scientific projects and this is accomplished by means of a Single-Sign-On (SSO) authentication mechanism [6] that follows also the IVOA SSO profile for authentication protocols. Once data become public, they can be retrieved by any user without the need of registration.

The radio archive web interface has been designed keeping in mind the different user categories, to allow data exploitation also by astronomers who are not expert in that specific wavelength domain. A simple, generic search is possible by means of a subset of query parameters common to all the observing modes, like for instance the celestial object coordinates or the observed frequency. More specialised users can access dedicated web forms to execute queries on interferometric, single-dish or pulsar datasets through parameters that are specific to the observing mode, like the scan geometry, the front-end/back-end configuration or the employed subset of antennas. VO-compliant Services are integrated in order to increase both the scientific data accessibility and re-usage.

3 Towards an Italian radio astronomical Science Gateway

Building on the radio data archive described in the previous Section, we are developing a prototype of radio Science Gateway intended to provide an integrated infrastructure to let the user initialize, manage and operate her/his own workspace, where to access and retrieve proprietary data; (re-)process them through dedicated software and tools to produce advanced data products; visualize, analyse and share scientific results.

The software architecture of the radio astronomical SG is based on the science-driven requirements and the preliminary design delivered in the recommendations of the AENEAS project[6], funded by the European Union within the Horizon 2020 program. Figure 1 illustrates the schematic layout of the components of the SG design that we prepared for the European SKA Regional Centre within the AENEAS project and that will be the architectural foundation on which to build and develop the Italian SG prototype for the radio domain. The realization of such a cyber-infrastructure is a complex task that requires the development of software and hardware components and a considerable effort for their harmonization and orchestration.

[6] https://www.aeneas2020.eu/

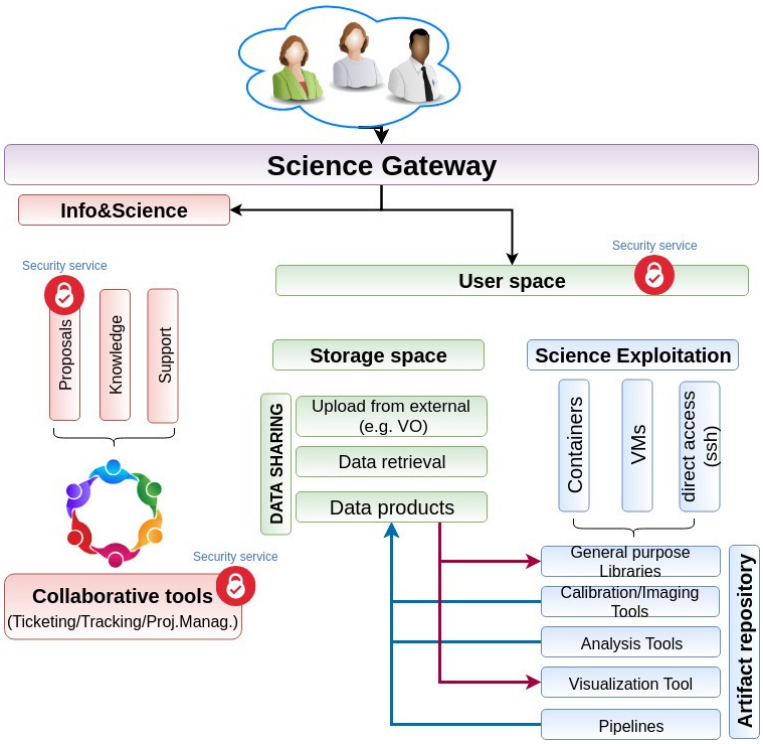

Figure 1. User's view of the Science Gateway. The diagram illustrates the composition of an Astronomical SG as depicted by the AENEAS H2020 project. The user can access several web-based or command-line services, some requiring Authentication & Authorization (security services). According to user privileges, different resources are available. For instance, the User Space is composed by storage and computation areas. Storage is devoted to proprietary data management (archived, uploaded or processed datasets), while the computation area offers several methods to run processes as well as analysis or visualization tools. Computational products are, in turn, made available in the User Space. Collaborative tools can be used to share the scientific results.

A number of building blocks required for the prototype implementation have already been developed by our group, like for instance the radio archive, and need to be interfaced and harmonized. Some other components, like the VOSpace and the data processing pipelines, are undergoing further development to be ported and integrated within the SG structure. Finally, some VO-oriented features that are desirable within a state-of-the-art SG are still to be designed and implemented.

A working implementation of Authentication and Authorization (A&A) is a key component to manage and grant permissions in the complex infrastructure of a SG. The working A&A implementation currently deployed for the radio archive is being adapted and improved to allow the interoperability of SG user attributes information and the delegation of user credentials. The User Space includes web clients and services for accessing resources, a storage space and a science exploitation area. The storage space must be compliant with VO standards to share and resolve the location of geographically distributed data in order to support not only data loading but also to ease discovery, access and retrieval of archival resources. An artifact storage area will be implemented in the User Space with appropriate links to the available libraries and tools. In this area the authenticated user will be able to search for the available resources for processing, analysis and visualization that are most suitable for her/his purposes and that can be loaded using software containers and/or virtual machines. The geographically distributed nature of the Italian radio data archive represents an ideal test bed to verify the efficiency of several functionalities like data transfer, search and handling. In particular, resource-demanding use cases like those posed by pulsar data, that can easily reach file sizes up to few TBs, are identified and used to test the computational performances of the SG prototype and their sustainability.

Data processing is possible through both dedicated pipelines that are integrated within the SG structure and custom software that can be loaded by the SG user in her/his User Space. Pipeline orchestration is controlled by the Yabi ([7]) workflow management system, an open-source Python application providing simplified web-based access to high-performance computing in an easy and powerful workflow creation environment. Initially the prototype will exploit two existing tools that are available at the Italian radio telescopes for the reduction of single-dish data in cross-scan and mapping mode. The Cross-scan Analysis Pipeline [8] is used to inspect, integrate and calibrate total-power on-the-fly cross-scan acquisitions on point-like sources for flux

measurements. The Single Dish Imager [9] produces total intensity and spectro-polarimetric maps from single- and multi-feed acquisitions. Also, we plan to deploy well consolidated analysis tools like the Common Astronomy Software Applications (CASA , [10]) and the PRESTO package [11] to allow the processing of interferometric and pulsar data respectively. Visualisation and analysis of intermediate and final results is allowed by the functionalities offered by the VisIVO [12] high-performance software, an open source collection of graphical applications which blend high performance multidimensional visualization, data exploration and visual analytics techniques.

Finally, a further development of this project will include the integration of VO-compliant Services and Clients in the SG, to maximize the effective scientific exploitation of archival resources by a wider community of users.

References

[1] Knapic, C., Zanichelli, A., Dovgan, E., Nanni, M., Stagni, M., Righini, S., Sponza, M., Bedosti, F., Orlati, A., Smareglia, R., 2016, "Radio data archiving system", *Proceedings of the SPIE* Volume 9913, id. 99132D.

[2] https://fits.gsfc.nasa.gov/fits_standard.html

[3] D. Muders, E. Polehampton and J. Hatchell, 2015, "Multi-Beam FITS Raw Data Format Revision 1.65" http://www3.mpifr-bonn.mpg.de/staff/dmuders/APEX/MBFITS/APEX-MPI-ICD-0002-R1_65.pdf.

[4] https://www.atnf.csiro.au/research/pulsar/psrfits_definition/Psrfits.html

[5] Dovgan, C. Knapic, C., Smareglia, R., 2016. "Radio Data Importer Report". OATS Technical Report n. 206. http://www.ira.inaf.it/Observing/download/Radio_Data_Importer_Report.pdf

[6] https://rap.inaf.it/Services/DEMO/

[7] Hunter, A.A., Macgregor, A.B., Szabo, T.O. et al. 2012, "Yabi: An online research environment for grid, high performance and cloud computing". *Source Code Biol Med* **7,** 1

[8] Giroletti, M., & Righini, S., 2020, "A flat-spectrum flare in S4 0444+63 revealed by a new implementation of multiwavelength single-dish observations". *MNRAS*, **492**, 2807

[9] Egron E., Pellizzoni A., Iacolina M.N. et al.2017, "Imaging of SNR IC443 and W44 with the Sardinia Radio Telescope at 1.5 and 7 GHz". *MNRAS*, **470**, 1329

[10] https://casa.nrao.edu/

[11] https://www.cv.nrao.edu/šransom/presto/

[12] Sciacca, E., Becciani, U., Costa, A. et al. 2015, "An integrated visualization environment for the virtual observatory: Current status and future directions". *Astronomy and Computing*, **11**, 146