A Public Data Archive for the Italian Radio Telescopes

Alessandra Zanichelli,[1] Cristina Knapic,[2] Righini Simona,[1] Nanni Mauro,[1] Matteo Stagni,[1] Francesco Bedosti,[1] Marco Bartolini,[1] Massimo Sponza,[2] Andrea Orlati,[1] and Riccardo Smareglia[2]

[1]*INAF-IRA, Bologna, BO, Italy;* a.zanichelli@ira.inaf.it

[2]*INAF-OATS, Trieste, TS, Italy*

Abstract. The amount of data delivered by modern instrumentation and observing techniques is bringing radio astronomy in the era of Big Data, and the nowadays widely adopted Open Data policies allow free and open access to data from many radio astronomy facilities. A fundamental ingredient to enable Open Science in the radio astronomical community and to engage also public participation (the so called Citizen Science) is thus the availability of public archives in which data can be accessed and searched with modern software tools. A web-based, VO-compliant public archive has been built to host data from the Italian radio telescopes managed by the National Institute for Astrophysics (INAF). The archive main features consist in the capability to handle the various types of data coming from the different observing instrumentation at the telescopes; the adoption of a policy to guarantee the data proprietary period; the accessibility of data through a web interface and the adoption of VO standards to allow for successful scientific exploitation of the archive itself in the data mining era. We present the progress status of the public Data Archive for the Italian radio telescopes being developed to provide the international community with a state-of-the-art archive for radio astronomical data.

1. Introduction

INAF manages three fully steerable reflector antennas for radio astronomy: the 32m dishes at Medicina and Noto and the 64m Sardinia Radio Telescope (SRT). These observing facilities can be used separately as single-dish instruments (SD) or in a coordinated manner within the international European VLBI Network (EVN) for interferometric observations. While data from VLBI experiments performed within the EVN are usually stored in the EVN Database, single dish observers were in charge of individually saving data coming from their own projects. In recent years two factors motivated the development of a data archive for systematic data storage. On one hand, the availability of a modern observing system resulted in a vastly increasing interest in the use of the Italian radio telescopes as single dish instruments. On the other hand, with the advent of the new SRT it is now possible to realize a fully Italian VLBI array (hereafter VLBI-IT) by using the three antennas in a coordinated manner, with our without other EVN facilities, and to pre-process VLBI-IT data in-house thanks to the local expertise on the DifX software correlator.

The main requirement for the radio data archive has thus been identified in the capability to host and handle both kinds of data (those coming from VLBI-IT obser-

vations and those coming from SD ones) coming in a variety of formats that will be described in the following section.

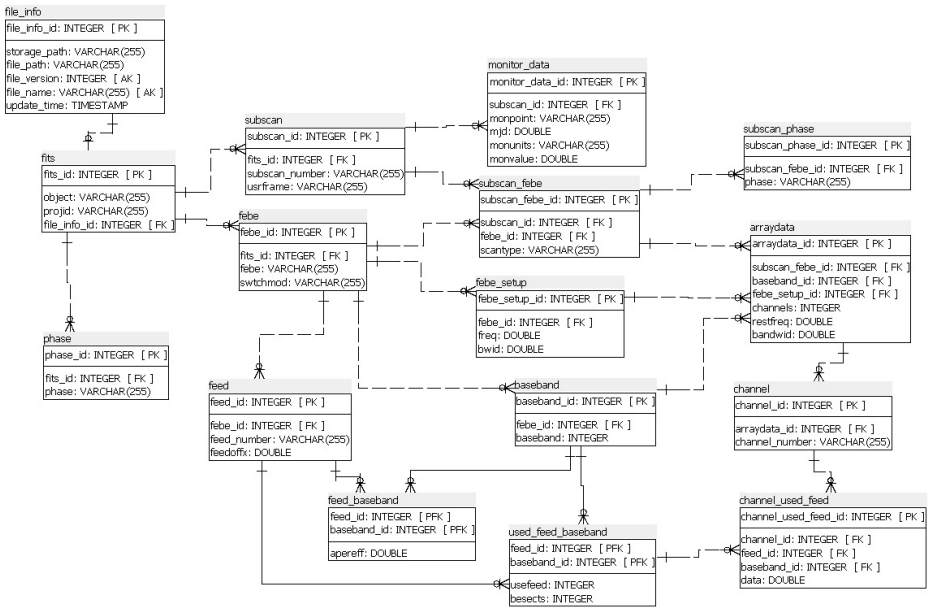

Figure 1. Radio data model.

2. Data formats and the Radio Data Model

The VLBI-IT and SD observing modes are characterized by different output (meta)data formats.// The correlation of VLBI-IT observations produces a so-called Visibility Data file: a monolithic FITS file some Gby in size. The Visibility File contains all the data from an observing session. It may contain, for instance, data for a number of target sources plus data from calibrator sources. The metadata set stored in the Visibility File is by no means complete in terms of the most commonly used query parameters. For archival purposes, each FITS Visibility Data file is thus associated with an XML Summary file grouping all the metadata information relevant for archival search, organized in blocks of keynames/keyvalues.

Two data formats are foreseen for SD observations (Zanichelli et al. (2016)). Currently, SD data from the Italian radio telescopes are written in a single-dish FITS format, a kind of table-based FITS which can be easily handled by the most common data analysis tools. Each observation may produce a number of FITS files hierarchically grouped in a directory tree. Metadata relevant for archival and query purposes are summed up in an accompanying FITS Summary file. In the future it is planned the adoption of the MBFITS standard format ((Muders et al. 2015)) developed for the Atacama Pathfinder Experiment (APEX) and in use at many radio astronomical facilities. MBFITS is particularly suited for the storage of multi-beam data when many data streams need to be written at the same time, and strongly rely on the FITS hierarchical grouping standard ((Jennings et al. 1997)). The hierarchical structure of MBFITS and

the presence in the headers of all the information relevant for archive queries makes the Summary file unnecessary.

Starting from the data formats described above, a general MBFITS database (1) has been built and is used as a baseline for the creation of the radio archive database. The radio archive database is a subset of the general MBFITS database suitable to store all the needed metadata from the currently adopted formats. However, the generic structure of the MBFITS 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.

Generally observers or archive users will not be interested in the type of information contained in the telescope observing schedule and log files. However, there are cases in which such information may prove to be fundamental to understand the data quality or to allow correct data processing. For instance, offsets in the position of a source may be caused by problems in the antenna tracking, which are recorded in the log file. Aiming at persistence and future scientific exploitation of data, ancillary information contained in the observing schedules and telescopes/correlator logs needs to be archived for each dataset as well. Given the different data formats in use and in particular the hierarchical structure of SD datasets, a dedicated software (the Finalizer) has been developed for data preparation before ingestion in the archive.

3. The Radio Data Archive structure

In order to contemporarily handle all the three types of radio data formats coming into the archiving system, a specific architecture and a configurable software were foreseen. The main characteristics of the archiving system software are the storage of data models into a dedicated database and the configuration of the software behavior using specific information for each instrument (or telescope) saved in dedicated tables of the data model. The Radio Archive is based on TANGO Distributed Control System and its overall schema is represented in Figure2.

The Radio Data Importer (RDI, (Dovgan et al. 2016)) is a TANGO server designed to handle and import the FITS, MBFITS and XML files in the Radio Archive containing both the observational metadata (data descriptors) and the data themselves. RDI configuration is stored in the datamodel database, which determines for each observing mode the set and the structure of metadata that has to be read from the input files, as well as the storage directory where the files are preserved. A general MBFITS database is used as a baseline for the structure of the Radio Archive database.

Two approaches are applied to speed-up the queries on the Radio Archive database. Firstly, all the data that might be retrieved while querying are calculated in advance and stored in the database itself. Consequently, these data are easily indexable and there is no need to calculate them at each query. Secondly, indexes are used on various columns and combinations of columns to speed-up the most frequent queries.

The modular and flexible design of the radio archive permits easy integration of new and different instruments, provided their data can be described by means of the general MBFITS database.

The Radio Archive architecture is scalable in the number of RDI devices so that data ingestion can be locally distributed at the telescopes. The Finalizer and the data storage as well may be geographically distributed and run independently at the various telescopes.

Database replication is based on a number of programs, namely data exporter, data importer, metadata exporter and metadata importer. The metadata exporter and importer are used to copy the database data from the master database (where the metadata exporter is running) to the replication database (where the metadata importer is running). In addition, the data exporter and importer are used to copy the files - whose links are stored in the master database - from the master server (where the data exporter is running) to the replication server (where the data importer is running).

Such an architecture is characterized by a flexible revision of policy and versions. It is robust with respect to logging and error handling, and allows for the development of services in the major Object Oriented programming languages (C++, Java, Python).

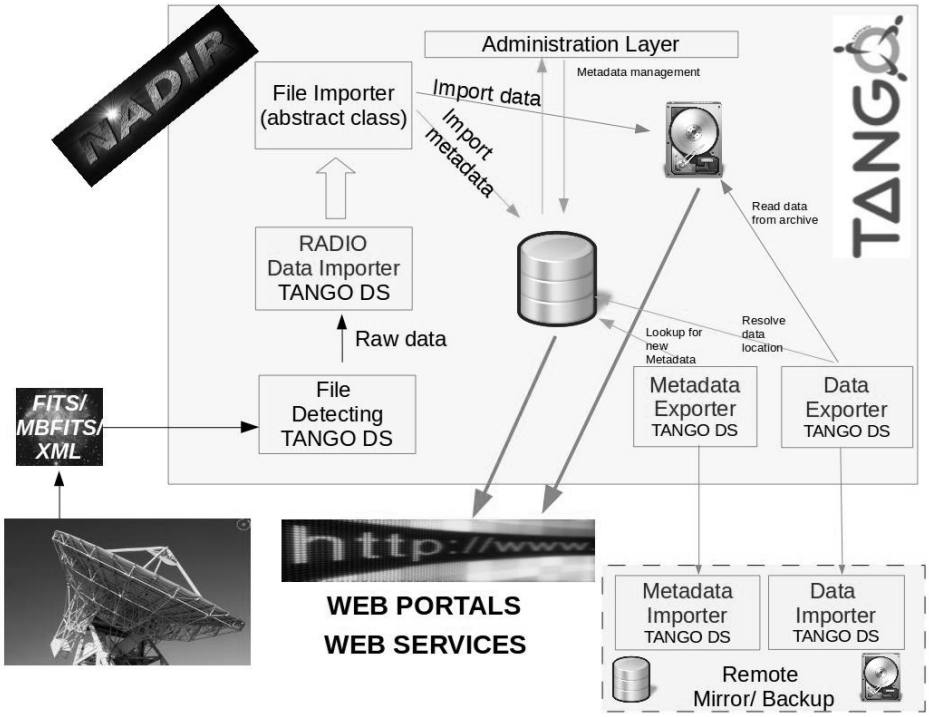

Figure 2. The Radio Data Archive.

4. The Radio Archive on the Web

Access to the Radio Archive is possible through a dedicated web interface. Archive users in the Open Science era include scientists and telescope staff but, in principle, also the general public. To guarantee the data proprietary period, according to INAF rules, to the PIs of the scientific projects a Single-Sign-On login authentication mechanism is foreseen. Public data follows the INAF policy and are available for download without user registration. In both public or private cases, users perform dedicated queries in the Radio Archive by means of web forms depending on the instrument (telescope) features or modes.

A simple, generic search is possible by means of a subset of query parameters common to SD and VLBI-IT modes, like for instance the celestial object coordinates or the observed frequency. Dedicated forms for SD and VLBI-IT data search are available, and contain query parameters specific to the observing mode, like the scan geometry, the front-end/back-end configuration or the employed subset of antennas. As an example, in Figure 3 the radio archive query form conceived to search for VLBI-IT data is shown.

To improve the query performance and increase the speed of results delivery, an indexing process of the most commonly used columns and some SQL functions were implemented. To limit real-time calculation, all the parameter computations that can be performed during the query execution are carried out in advance and stored in the database following OAIS (Open Archival Information System) standard rules. Also, indexes are created on various columns and combinations of columns. These processes are tailored to the necessities of the Radio Archive query form, taking into account the effective curiosity and needs of (radio)astronomers.

Typically the output page contains a summary of query results in tabular form, to allow users to download of checkbox-selected datasets. Optionally, also ancillary information in terms of schedule and log files associated to the recorded data can be retrieved as well.



Figure 3. The Radio Data Archive query form for VLBI-IT data search.

5. The Radio Archive and the Virtual Observatory

The Virtual Observatory offers a unique opportunity to guarantee the accessibility of radio data to the astronomical community and to the public, thus maximizing the scientific impact of the observing facilities.

In order to increase both the open data re-usage and the accessibility of scientific data to the astronomical community and the general public, VO-compatible standards like the Table Access Protocol (TAP) are preferable. To this aim, a TAP-based service to export the Radio Archive Database to VO-compliant clients is going to be published.

The use of Observation Data Model Core Components and its appropriate application to the radio astronomical raw data from the Italian radio telescopes is also under study. The goal is to investigate how and to what extent the VO already addresses the use cases for the Italian Radio Archive.

A more detailed discussion of the Radio Data Archive in connection with VO requirements, services and data models is presented in Knapic et al. (2018).

References

Dovgan, C., Knapic, C., & Smareglia, R. 2016, Radio Data Importer Report

Jennings, D. G., et al. 1997, A Hierarchical Grouping Convention for FITS, Rev. 8. URL http://fits.gsfc.nasa.gov/registry/grouping.html

Knapic, C., et al. 2018, in ADASS XXVII, edited by TBD (San Francisco: ASP), vol. TBD of ASP Conf. Ser., TBD

Muders, D., Polehampton, E., & Hatchell, J. 2015, Multi-Beam FITS Raw Data Format. URL http://www3.mpifr-bonn.mpg.de/staff/dmuders/APEX/MBFITS/APEX-MPI-ICD-0002-R1_65.pdf

Zanichelli, A., et al. 2016, Data formats for the Medicina and Noto radio telescopes in view of a common Archive. URL http://www.ira.inaf.it/Library/rapp-int/488-15.pdf