PROCEEDINGS OF SCIENCE

The LOFAR Transients Pipeline

John Swinbank[*†]
University of Amsterdam
E-mail: swinbank@transientskp.org

The LOFAR Transients Key Science Project will exploit LOFAR's unique wide-field, high-sensitivity view of the low frequency radio sky to carry out a large scale transients monitoring program: the LOFAR Radio Sky Monitor. Exploiting this new and unprecedented capability has required the development of a number of new techniques and technologies. Here, we highlight some of the approaches which have been adopted and describe how they are being applied both within the Radio Sky Monitor and across the LOFAR project.

ISKAF2010 Science Meeting - ISKAF2010
June 10-14, 2010
Assen, the Netherlands

[*]Speaker.
[†]On behalf of the LOFAR Transients Key Science Project.

PoS(ISKAF2010)082

1. Introduction: the Radio Sky Monitor

The Radio Sky Monitor, or RSM, [6] is a key component of the Transients Key Science Project's (TKP) [5] goal to exploit LOFAR's unique capabilities to explore the low-frequency transient radio sky. As illustrated in Figure 1, LOFAR's multi-beaming capability makes it possible for multiple beams from the core to tile out a large area on the sky (varying from 65.8 square degrees per beam at 30 MHz to 4.0 square degrees per beam at 240 MHz), while simultaneously using individual beams to monitor noteworthy objects. Images will be made on a logarithmic range of integration times between one and 10^4 seconds, with transients and variable sources being identified by a combination of image differencing and statistical analysis of lightcurves. While the precise survey strategy will be determined in the light of practical experience as LOFAR comes online, and will likely concentrate on the zenith and galactic plane, it will be possible to use this mode to survey the majority of the visible sky to a reasonable depth (tens of mJy, depending on frequency) in a 24 hour observing period.

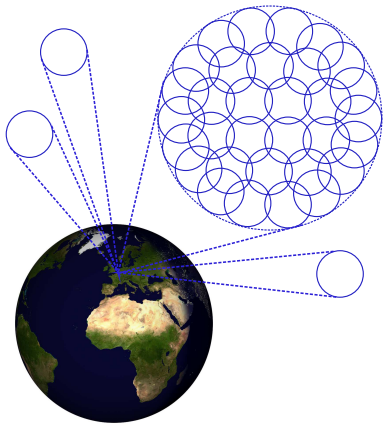

Figure 1: The Radio Sky Monitor concept: multiple beams from the LOFAR core tile out a large area on the sky.

Observing in this mode will stretch the limits of existing techniques and technologies. The TKP has therefore been developing a sophisticated pipeline system to process and respond to the large amounts of data generated. A schematic overview is shown in figure 2. In brief, the transients pipeline (TP) is tightly coupled with an optimised version of LOFAR's Standard Imaging Pipeline (SIP). This provides data flagging, compression, calibration and imaging, eventually delivering an "image cube" [2] (a group of simultaneous images of the same area of sky at different frequencies) to the TP. The images are then searched for sources, and the results fed into a database which will automatically associate them with known objects and generate lightcurves. Interesting lightcurves are extracted from the database and fed to a source classification and response system, which can then arrange for appropriate follow-up actions to be taken. The pipeline will also listen for notification of potentially interesting events from other observatories using the *VOEvent* format (see Section 6), and process them in much the same way.

2. Pipeline Framework

The various tools which are used as part of the TP do not all present a uniform interface : some are developed in-house; some are adopted from external sources. Some are compiled executables; some are shared libraries; others are Python modules. However, they must all interface with each other, and run under the control of the MAC, the LOFAR control system [10].

A pipeline framework system has therefore been developed to provide a uniform interface to all the various pipeline components. Each component is wrapped in a so-called 'recipe', which exports its functionality to other pipeline components (and, indeed, the whole LOFAR system) through a consistent and convenient interface. The framework provides a number of useful services

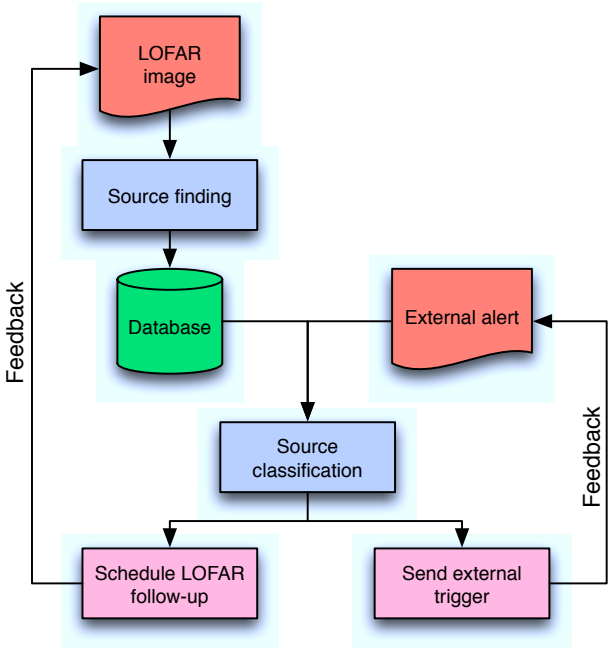

Figure 2: A simplified outline of data flow through the TKP pipeline.

which may be exploited within the recipe, such as distribution of jobs across the LOFAR processing cluster (based on the IPython[1] system).

Of course, the configuration of a pipeline component must often be adjusted depending on the task in hand. For instance, depending on the scientific requirements, one may wish to adjust the detection thresholds for the source finding system (see Section 3). A recipe can be combined with a set of configuration parameters to form a 'task', which operates on some input data and provides results to the next stage in the pipeline.

The pipeline developer can connect the tasks via a straightforward Python script, making use of arbitrarily-complex logic: it is not adequate to simply chain them sequentially, as pipelines can be responsible for decision making, looping, and so on.

Once a pipeline has been constructed in this way, it can be automatically started when appropriate data is available by MAC, and will ensure that results and logging data are fed back in a consistent way to the observer.

The development of this framework has built upon previous work performed for the Westerbork Synthesis Radio Telescope[2]. Although designed initially specifically to meet the requirements of the TP, it is now being used for several LOFAR science pipelines, including the Standard Imaging Pipeline [7] and the Known Pulsar Pipeline [1].

[1] `http://ipython.scipy.org/`
[2] `http://www.astron.nl/~renting/pipeline_frame.html`

3. Source Finding

Fast and accurate source identification and measurement is critical to the TP. This is used both for the direct identification of new transient sources (by identifying them in difference images) and for the addition of new measurements to the lightcurve database (Section 4). After evaluating available packages, the TKP has implemented a new source finding system as a Python module [13].

Two source finding systems are available: one based on simple thresholding (identifying islands of pixels above a certain multiple of the local RMS noise), the other based on a false detection rate algorithm [8]. The latter provides a much more convenient way of controlling the statistical properties of the resulting catalogue.

After identification, sources can be deblended using a multi-thresholding technique, and then fitted with elliptical Gaussians.

In the longer term, the TKP plans to make all its source finding routines available to the community as a standalone package. It is likely that future development will supplement the Python-based system with a faster implementation of the same algorithms in C++.

4. Database

The database is central to processing throughout the LOFAR transients system. In its simplest form, it is a repository of lightcurve information: it will store (and make available for data-mining) information on all the sources observed by the RSM over its lifetime. While the RSM is running, this will result in up to 10 MB/s of data being added to the archive (the growth rate per year obviously depending on how much observing time is allocated for RSM observations).

It is immediately clear that storing and accessing this amount of data pushes the capabilities of standard database management systems. The TKP has therefore been working with the Centrum Wiskunde & Informatica (CWI) in Amsterdam, developers of the unique, high-performance MonetDB database[3] [3]. By introducing innovation at all levels of the database stack, from a column-oriented data storage model to a vectorized query execution system, MonetDB has a proven track record in high-performance astronomical applications [9].

This solid foundation provides an excellent base not just for data-mining, but for extending pipeline processing into the database. TKP members have developed systems for automatically *associating* sources in the database: that is, when a new source measurement is inserted into the database, an automatic routine will determine what other detections are of the same object, and combine them all to build a lightcurve. The database can automatically keep track of all lightcurves, monitoring them for variability, and notify the rest of the pipeline when a new transient or variable source is discovered or when a known object begins to behave in a scientifically noteworthy way.

5. Classification and Response

After a transient or variable source has been identified—either via image differencing, or via a statistical analysis of its lightcurve in the database—an attempt will be made to identify and classify it based on the properties of its lightcurve. This is essential not only for future data-mining of the

[3] `http://monetdb.cwi.nl/`

lightcurve archive (attempting to find scientifically relevant information in the terabytes of data collected would otherwise be impossible), but also to make it possible to respond to ongoing events in real-time. Such responses could include re-running the pipeline with a different configuration, scheduling a follow-up observation with LOFAR, or broadcasting a notification of the event to the community at large (see Section 6).

A classification system is being developed which will automatically classify lightcurves as they are stored and updated in the database. Classification will depend on a list of simple parameters (a "feature vector") which can be derived from the lightcurve: quantities such as flux, variability, dispersion measure, spectral index, and so on. Many of these quantities can be automatically computed as the lightcurve is updated by the database engine itself; for some, more intensive computation in an external pipeline process is required. As part of the process, it is necessary to account for partially sampled or otherwise incomplete lightcurves.

Based on the feature vector, a number of classification techniques will be applied. For maximum performance and generality, a machine learning approach is being investigated and a library of routines for transient classification based on a random forest of decision trees has been developed [4]. However, such a system will need human-defined training data before it becomes useful, and there will always be a requirement for identifying events of particular interest to specific astronomical cases. Therefore, provision is also being made for astronomer-defined classification steps to be inserted into the pipeline.

As classifications and derived quantities are calculated, they are written back to the database, and can thus be referred to in future pipeline runs. As more data becomes available, the classification will become increasingly refined.

6. Notifications and VOEvent

Since it will explore a new parameter space, it is hard to predict the rate of transient discovery by the RSM. However, estimates in the range of tens to hundreds per 24 hours of observing are likely conservative. And LOFAR is just one of a range of new "transient machines", including such facilities as Pan-STARRS and LSST, which will be coming online over the next few years. The deluge of transients being detected will quickly grown beyond the abilities of humans to analyse them all.

Further, by combining the software-driven nature of LOFAR with the real-time monitoring made possible by this pipeline, it is possible to respond to events in near-real-time, potentially catching the most scientifically valuable results. In order to best take advantage of this, however, it is obviously necessary to remove the requirement for human intervention.

The TKP is therefore incorporating an automatic system for both generating and receiving alerts of transient events as quickly as possible. This is based on the VOEvent standard [12, 14], which provides a structured, machine-readable way of representing information about events as an XML document. VOEvent is transport-agnostic: it can be delivered to a recipient system in any way that is convenient. We anticipate both private event channels being developed with partner facilities (for example, LIGO and MAGIC) as well as broadcasting LOFAR-derived events to as wide a community as possible. Test messages are already being successfully exchanged with LIGO, and planning of appropriate follow-up methodologies is underway.

7. Conclusion

The LOFAR Transients Key Science Project plans to exploit the unique capabilities of LOFAR to regularly monitor the sky for transients. In the process, a completely new parameter space will be explored.

In order to enable this project, a number of new technologies and techniques have been developed, which are outlined above. Many of these may also be relevant to other projects. In particular, the pipeline framework has already been adopted by other LOFAR science pipelines, the source finding code will be made publicly available, and the database systems are potentially useful for many large astronomical catalogues. We will be making as much as possible of this software publicly available, and we are keen to develop these techniques in collaboration with other projects.

References

[1] A. Alexov, *The LOFAR Known Pulsar Data Pipeline*, in proceedings of *The ISKAF2010 Science Meeting*, 2010, PoS(ISKAF2010)060.

[2] A. Alexov et al., *LOFAR Data Format ICD: LOFAR Sky Image*, Document ID LOFAR-USG-ICD-004, Revision 0.13, April 2010.

[3] P.A. Boncz, *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*, Ph.D. Thesis, University of Amsterdam, 2002.

[4] T. Coenen, *Automatic LOFAR Transient Classification*, Masters Thesis, University of Amsterdam, 2008.

[5] R.P. Fender and the LOFAR Transients Key Project, *The LOFAR Transients Key Project* in proceedings of *The VI Microquasar Workshop: Microquasars and Beyond*, 2006, PoS(MQW6)104.

[6] R.P. Fender, *LOFAR Transients and the Radio Sky Monitor* in proceedings of *Bursts, Pulses and Flickering: wide-field monitoring of the dynamic radio sky*, 2007, PoS(Dynamic2007)030.

[7] G. Heald, *Recent LOFAR imaging pipeline results* in proceedings of *The ISKAF2010 Science Meeting*, 2010, PoS(ISKAF2010)057.

[8] A.M. Hopkins et al., *A New Source Detection Algorithm using the False-Discovery Rate*, *AJ* **123**:1086–1094, 2002.

[9] M. Ivanova, N. Nes, R. Goncalves and M. Kersten, *MonetDB/SQL Meets SkyServer: the Challenges of a Scientific Database*, in proceedings of *The 19th International Conference on Scientific and Statistical Database Management*, 2007.

[10] K. van der Schaaf, C. Broekema, G. van Diepen and E. van Meijeren, *The LOFAR Central Processing Facility Architecture*, *Experimental Astronomy* **17**:43–58, 2004.

[11] B. Scheers, Ph.D. Thesis, University of Amsterdam, 2010 *in prep*.

[12] R. Seaman et al., *Sky Event Reporting Metadata (VOEvent)*, version 1.11, International Virtual Observatory Alliance, 2006, http://www.ivoa.net/Documents/PR/VOE/VOEvent-20060810.html.

[13] J.N. Spreeuw, *Search and Detection of Low Frequency Radio Transients*, Ph.D. Thesis, University of Amsterdam, 2010.

[14] J. Swinbank, *Standards and Systems for Transient Response*, in proceedings of *The 8th International e-VLBI Workshop*, 2009, PoS(EXPReS09)022.