LSST:UK DAC: An initial analysis of requirements from Transient and Variable Science

WP2(Data Access Centre)
Task 2.3 (Support of Phase A Dev work)

Project Acronym	LUSC-A
Project Title	UK Involvement in the Large Synoptic Survey Telescope
Document Number	LUSC-A-09 <Contact George.beckett@ed.ac.uk for number>

Submission date	25/06/2019
Version	1.0
Status	Final
Author(s) inc. institutional affiliation	Andy Lawrence, Dave Morris, Roy Williams (ROE; Stephen Smartt (QUB); Mark Sullivan, Sebastian Hoenig (Southampton); Carole Mundell (Bath); Phil Lucas (Herts); Tim Naylor (Exeter)
Reviewer(s)	Bob Mann (Edin)

Dissemination level	
Confidential	*The prime purpose is for informing other members of the consortium. In due course it may be suitable to be public.*

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	28/04/2017	Outline and introductory section.	A.Lawrence, Smartt, Sullivan, Mundell
0.5	14/11/2017	Full draft.	Lawrence, Smartt, Sullivan, Mundell, Hoenig, Lucas, Naylor, Morris, Williams
1.0	25/6/2019	Final version. Only minor changes since version 0.5; this final version is essentially considered a tidied up "snapshot" as of Nov. 2017	Lawrence, Smartt, Sullivan, Mundell, Hoenig, Lucas, Naylor, Morris, Williams

LSST:UK DAC: An initial analysis of requirements from Transient and Variable Science

Andy Lawrence[1], Stephen Smartt[2], Mark Sullivan, Carole Mundell, Sebastian Hoenig, Phil Lucas, Tim Naylor, Dave Morris[1], Roy Williams[1]

Abstract
We briefly summarise the key implications for the LSST:UK DAC set by the needs of several science areas involving the study of transient events and highly variable objects - in particular Supernovae, Tidal Disruption Events, extreme AGN variables, Gamma Ray Bursts, Gravitational Wave Events, microlensing events, accreting binary stars, and eruptive stars. Other UK DAC documents provide an overview of LSST transient processing, previous experience with PanSTARRS and ATLAS, and the design of the Lasair Broker. **Note:** This document is a "snapshot" of the initial requirements planning as at November 2017

[1] *Institute for Astronomy, University of Edinburgh, ROE, United Kingdom*
[2] *Astrophysics Research Centre, Queen's University Belfast, United Kingdom*

Contents

1. Introduction

1.1 Context

This document encapsulates the initial requirements discussions concerning transient and variable science, for Phase A of the LSST:UK project. Minor updates have been made at the end of Phase A (June 2019) but the bulk of the material is exactly as the document was written in November 2017. Following this analysis, the Lasair Broker project began, other variability related projects were developed as part of the Phase B proposal process, and transient and variable science requirements have been further developed in the context of the project Science Requirements Document.

1.2 Motivation

The LSST:UK consortium is expecting to construct a Data Access Centre (DAC) to supply the needs of UK scientists working on LSST data. To a first approximation, this may be a clone of the US DAC. However, the priorities of UK scientists may be somewhat different to the US community, so as a minimum we must make sure that the UK DAC addresses those priorities. In addition, we have the opportunity in the UK for an organised approach to "User Generated" software and data products, previously known as "Level 3" products, and known within the UK project as the DEV programme. The DEV technical work will develop tools and pipelines that interact with the UK DAC, and that will facilitate UK science. The DEV work is formally distinct from the DAC infrastructure, but in practice we expect each DEV team to work closely with the DAC to implement their tools, and the DAC must anticipate what the DEV tools may be. Finally of course, although UK priorities may be different from US ones in their relative balance, there is a large overlap with US activities; we hope that innovative work undertaken in the UK may be fed back into the international community project.

One of most important priorities in the UK is time domain studies - and especially transient events and highly variable objects. As well as being of widespread scientific interest, it is an area where UK workers have a considerable amount of technical experience, which we should build on. In Section 2 we survey how things work in some key scientific areas.

In Section 3 we try to draw out some general conclusions and lessons. In Section 4 we make some suggestions for the development of the UK DAC, which are explored more fully in a separate document.

1.3 Definitions

Note that throughout this document, *DAC* refers to the anticipated UK Data Access Centre, which will receive data from the LSST project and serve it to registered UK users. To refer to the different kinds of pipeline and data product, we use the LSST construction-phase terminology in place when this document was first written - i.e. Levels 1,2, and 3.

- *Level-1* refers to the transient alert stream emerging from LSST, and the associated pipeline.
- *Level-2* refers to the annual stacked images and associated catalogues, and the associated pipeline and potential products, such as light curves.
- *Level-3* refers to community-supplied software and pipelines which interact with the DAC to undertake further scientific data processing. In UK terms, Level-3 software is being organised through the "DEV" workpackages of the LSST:UK consortium.

In the operations phase, these categories are now referred to as *prompt data products*, *data release products*, and *user generated products*

1.4 General points

(i) We present an extremely simplified and brief summary of the key scientific points. In most cases, much more detail is available in the LSST Science Book, and in due course, the various Science Collaboration Roadmaps. Our aim is to draw out the simplest implications for DAC and DEV design.

(ii) We do not discuss moving objects here, even though many of the same issues recur.

(iii) The basic structure of the LSST data flow is considered fixed, although there of course some very relevant live issues, such as the cadence.

(iv) Nearly all the science areas rely on an initial level of QA filtering, removing junk from real objects.

1.5 Related documents

This UK-DAC document analyses the UK community science requirements in areas involving transient events and extreme variables which are transient like. It relates to several other UK DAC documents: **(Update with current doc numbering...)**

- LUSC-A-02 sets out an overview of the LSST approach to Level-1 alerts
- LUSC-A-03 summarises the QUB experience in processing the PanSTARRS and ATLAS alert stream
- LUSC-A-09 (this one) analyses the UK time domain science requirements
- LUSC-A-08 Describes the design of the Lasair event broker
- The LSST:UK Science Requirements Document

Key related LSST project documents include:

- LSE-163 Data Products Definition Document
- DMTN-102 LSST Alerts- Key numbers

Further LSST:UK documents anticipated early in Phase B include a survey of the relevant time domain technology landscape, and a plan for UK DAC development work in this area.

2. Key science areas

In this section we summarise and analyse the science requirements in each of a number of areas. We do this in a qualitative and descriptive manner. In Section 5 we set out an outline of how to collect relevant information in a uniform and quantitative manner.

2.1 Supernovae

Supernovae occur within galaxies and the precursors are invisible before the event. They are therefore discovered by difference imaging, and are part of the Level-1 alert stream. LSST will be a primary *source* of SNe, but can also provide context for SNe discovered elsewhere, if they go off in the right bit of sky. They typically rise in days and decay over weeks, although fast-evolving supernovae with much shorter rise and decay time scales (order 10 days) have recently been identified ("fast-and-blue"; FaB). Follow-up within hours or days (spectroscopy, other wavelengths) is therefore needed, which requires even faster decision on whether a transient is a supernova or not, what type of supernova it is, and whether it is interesting enough to follow up. Supernovae evolve in colour as they decay, in a manner which depends on the type. This has some cadence implications, but we won't discuss those here. To make these decisions, scientists need the *context* of the event - precursor LSST imaging and photometry; the LSST light curve as it evolves; and the external context - imaging and photometry from other wavelengths/telescopes.

In recent supernova focused projects such as PESSTO, the balance between automated and manual filtering has evolved in the direction of automation, including for example machine-learning based light-curve classification. However, at the downstream end, SN scientists still like a level of manual decision, based on assembling the contextual information, adding annotations, and manually moving "tickets" to various queues or archived regions. (This is what the "PESSTO Marshall" does).

2.2 Tidal Disruption Events (TDEs)

TDEs are in many ways similar to SNe - rising in days and decaying in weeks to months, and emerging from invisibility in galaxies. However unlike SNe, they will always be at the centres of galaxies, and candidates studied so far seem to show little colour evolution. Requirements for studying TDEs are therefore fairly much the same as for SNe, and indeed the contextual information is needed to distinguish TDEs and SNe. Contextual information is also needed to separate true TDEs from extreme AGN variables (see below) - there should be no previous AGN activity. As well as there being no previously known AGN, examination of the pre-discovery LSST data should reveal no variability. Checking this will need access to past pixel data and/or forced photometry as well as past routine photometry. It can be that this will occur at leisure in later analysis; but it might also be needed promptly to take decisions on whether to act on a candidate TDE.

2.3 AGN extreme variables

Large numbers of AGN are already known, and we can expect that LSST will produce many more, in particular by combining colour selection and variability information. The simplest requirement is therefore for the DAC to provide methods to easily link light curves to Level 2 database sources, and enable efficient data mining which goes beyond simple SQL queries. Scientists will want some standard attributes of those light curves - variance, structure function slope, etc - but these will probably be easily derivable from the light curves.

A minority of AGN show extreme variability - blazars, changing look quasars (CLQ), switch-on/switch-off sources, and high amplification microlensing events. Some of these will be known objects, and of course scientists will organise campaigns to monitor some of these at multiple wavelengths. However, often, they will want dedicated campaigns to respond within days when an object does something interesting - e.g. a blazar starting an outburst, a microlensing event rising to a peak, a suspected CLQ taking a nosedive, or a previously non-AGN galaxy showing AGN activity. This will need a user-supplied watchlist, the ability to characterise and classify a light curve, and the ability to *forecast* what will happen next. The duration of variability ranges from as little as days for some blazars to years for many CLQs, microlensed outbursts, or the onset of accretion.

Rather harder is the problem of how to spot and act on such interesting events in objects which are not already known, and are just a tiny-fraction of Level-2 database. They will not necessarily be part of the Level-1 alert stream because they vary too slowly. We need something like an automated gradient detector, or a Δm flag over some chosen timescale, or possibly something much more sophisticated.

2.4 Stellar microlensing

Some stellar microlensing events will come out of invisibility; others will be existing stellar sources that begin to increase in brightness. They come and go by several magnitudes over days to weeks. This is fast enough that they may appear in the Level-1 alert stream for the main survey, but it will be very hard to recognise them for what they are. The Deep Drilling Fields have a high enough cadence that light curve characterisation and forecasting as discussed above for AGN can recognise stellar microlensing events. However the LSST cadence will not be fast enough to complete the job; an event alert has to be sent out to other observatories for round the clock coverage, especially near the peak. The microlensing community is already used to the idea of forecasting and adapative cadence; the key requirement is for LSST to recognise them quickly (within hours) and distribute alerts.

2.5 Eruptive stellar variables

A number of stars undergo dramatic outbursts - e.g. FU Orionis stars, Novae, Cataclysmic variables. Like with AGN, some are already known, but we don't need to monitor them continously at other wavelengths - rather, we need to filter according to a user-supplied watch-list, and to classify, characterise, and forecast based on the filtered light curves. Typically such objects show a relatively predictable pattern, but very unpredictable timing. Like with stellar microlensing, new events have to be recognised quickly (within days), so that alerts can be issued. As with the AGN, although there will be pre-known objects, there will be great interest in finding new objects. This will require tracking huge numbers of objects and triggering on interesting changes.

2.6 Flare stars and general stellar variability

Flare stars, like stellar microlensing events, eruptive variables, and extreme AGN variables, can be seen as transient-like, but may not be in the Level-1 alert stream, depending on timescales and previous history. Likewise, a small number could be in user-supplied watchlists, but most will be in otherwise anonymous stellar objects, requiring triggering from an extremely large pool. However, what makes this area even harder than eruptive variables and stellar microlensing is deciding when to trigger - general variability is very common, and has a complex variety of behaviours. Would an automated trigger alert us to something like Boyajian's star?

2.7 Periodic variables

Periodic variables include pulsating stars like Cepheids and RR Lyrae, traditional eclipsing binaries, and low-amplitude eclipses from exoplanets. It is not expected that there are sudden changes that will trigger action, and so no obvious need for a watch-list. There will be great interest in discovering new such objects. It could be that this is purely a science-exploitation issue rather than an infrastructural one. However, there may be an argument to supply extra light curve attributes or data products (for example periodogram) to facilitate such science.

2.8 Gravitational wave events and gamma-ray bursts

We now know that neutron-star mergers have an optical counterpart, which lasts a few days-weeks. It is still not clear whether black hole mergers have an EM counterpart, but this means that searching for such counterparts remains a high priority. These events are rare enough (once a day for GRBs, a few times a year for GWEs) that LSST will not usually be looking at the right part of sky, but when it is, it could play a crucial role . The requirement would be to query the history of recent transients by time and estimated distance. This would preferably be an automated search responding to an incoming request.

The GRBs are thought to be highly beamed, but the accompanying optical transient is not. LSST could therefore be a source for mis-directed GRBs, which should be many times more common. They will be in external galaxies, and invisible before the event. Like SNe, they should therefore appear in the difference-imaging Level-1 alert stream. The issue will be in early characterisation, in order to recognise that the event is not a normal supernova, and give it a high priority for issuing an external alert.

2.9 Fast radio bursts

Some progenitor theories for Fast Radio Bursts predict short-duration signatures at optical wavelengths, and others do not. The origin of the burst may not be precisely localised on the sky – the Parkes beam full-width half-maximum (FWHM) is 14.40 arcminutes. The LSST data can thus be searched in this small area either immediately after, or long after the FRB.

3. Analysis of infrastructural implications

Some recurring themes come through the preceding analysis of science requirements.

(1) LSST can be both a *source* of events, and a *context* for external events. The DAC needs to accept and act on both internal alerts and external alerts. It also needs to be able to emit our own alerts that others can act on. Emitted alerts need to be in a standard open format, because the recipients are in general unknown.

(2) We clearly want our own *broker*, in the sense of a running agent that accepts one or more event streams, and emits one or more event streams of its own. Even if this is a simple re-direction, the term "broker" here is appropriate by analogy with financial services etc - the expectation is that end-users consume our version of the event stream, rather than going to the source. Of course this makes more sense if the broker *adds value* - filtering events; adding context; stripping out or adding metadata fields; providing an interface for interrogating the event stream; providing a toolbox for users to make customised use of the event stream.

(3) Filtering the event stream is a multi-stage process - for example, (a) removing junk; (b) separating candidate object types; (c) separating into priority lists etc; (d) producing a follow-up action list. It will probably be clearest if this cascade is made apparent to the user, each stage with an associated *trigger*, allowing an event to pass to the next stage, so we can perhaps refer to a Stage 3 trigger etc.

(4) The first of the above stages is universal and can be hard-coded into the DAC. The rest are science-case dependent, and so should not be hard wired. Rather, the DAC and/or associated DEV project should provide some kind of *toolbox* so that scientific projects can customise their triggers. Discussion with users suggests that it may be best to provide both a menu-driven web-style interface, and a library of routines for scripted access. Using the menu shows the user the equivalent script, encouraging them to move from menu to writing script directly.

(5) There is more to life than the Level-1 alert stream based on difference imaging. Many potential science projects will want to trigger on light curves emerging from the Level-2 process. Of particular concern in this respect are slowly evolving events and/or variable objects on multi-year timescales where not only short-term Level-2 data would be required, but some form of access to LSSTs full Level-2 archive. This may be computationally challenging, especially with regards to storage and access.

(6) There is also a recurring need for the ability to monitor a user-supplied watch-list, and to trigger action based on this. A user-supplied watch list may be for example a few hundred galaxies that have emitted very high energy gamma rays; or the DAC itself may implement a watch-list of 10^8 galaxies in the local universe, since many users would be interested in events that are associated with a host galaxy.

(7) Triggering on light curves requires the ability to rapidly classify and characterise them - either by supplying standardised computed attributes, or more generally to feed them through an algorithm that takes decisions and makes forecasts. Some of this problem will be dealt with at the downstream science exploitation end. However, the DAC needs to supply a framework or toolbox that makes it easy to flexibly construct customised triggers, and a way for users to apply these tools to incoming events. Furthermore, the best algorithm won't always be clear in advance, so this is a machine learning problem very appropriate for DEV development.

(8) Most scientific areas need contextual information of various kinds. Internal information includes previous LSST history, and before-and-after postage stamps. External information includes photometry, astrometry and images from other surveys, and information derived from those surveys, such as redshift. In the past such external information has been hardwired in to the interfaces for projects such as FGSS or PESSTO. It would be better to allow both: the DAC makes annotations from touchstone archives such as SDSS or 2MASS, as many users want these; but also to allow each science project to customise its interface, to provide an *annotation toolbox*. Such a toolbox would be built with established VO standards and registries to maximise the reach.

(9) Although we need to automate as much as possible, there is also a need to add manual annotations at various stages, especially at the downstream end. Such annotations could be simple text comments, but they could also include the additional of numerical data, links to images, and so on.

(10) So far we have not identified a pressing need for response (e.g. sending out alerts or triggering follow-up) within minutes, but there sometimes is a need within hours.

4. Possible design implications for the UK DAC

Detailed designs will come in due course. Here we draw some very general lessons for the UK DAC.

The requirements analysed in this document have implications for both DAC and DEV work. Within LSST:UK, the agreed distinction between DAC and DEV work is that DEV work requires astronomical knowledge whereas DAC work does not. DAC and DEV work will need to interact quite closely of course. The primary job of the DAC will be to do the initial stages of processing of the data emerging from LSST; the secondary job will be to provide a framework, and tool boxes, to enable the DEV teams to add value. Fig. 4 gives a general overview of the LSST dataflow/processing stages, and who is responsible for what.

Users will want to see the event stream in two distinct ways: *past* and *real time*. The past stream will be preserved in a database, and can filtered or replayed through algorithms and triggers in the development process, as well as being used for data-mining in that archive. Once a trigger is built and optimised, the user will want to work with the real time stream, so that the trigger code is executed immediately upon arrival of an event, and the possibility of further actions taken. The DAC will need to make this transition possible for the user, from past to real time streams.

In Fig. 4 we illustrate the idea that there may be multiple filter/trigger stages, and that these are mostly supplied by the DEV teams and/or community. Things that the DAC will need to supply are probably:

- A menu driven web interface
- A python-scripting notebook interface
- Docker-style containers with pre-packaged “pipes”
- A light curve toolbox
- An annotation toolbox
- Facility for interacting with watch-lists

5. Quantitative requirements

In this section we set out an outline for distilling key quantitative features of various science areas in a more uniform and quantitative manner, in the expectation that this will drive the design of the UK DAC/DEV system. **Note:** The values below represent an initial attempt at quantifying these features, as of November 2017. It is expected that we will return to this analysis during Phase B development.

Fast Radio Burst
Expected frequency: Unknown in the optical
Expected duration: seconds (bursts are milliseconds in the radio)
Desired response time: minutes (less than 1hr)
Primary data source: Radio surveys
Secondary data source: LSST fast transients discovered in the survey

Gamma-ray burst afterglows
Expected frequency: 1 per day
Expected duration: minutes to hours
Desired response time: minutes (less than 10 mins)
Primary data source: LSST Level 1 stream
Secondary data source: Gamma and x-ray missions

Kilonovae
Expected frequency: 0.1 - 1 per week
Expected duration: hours to days
Desired response time: minutes (less than 10 mins)
Primary data source: LSST Level 1 stream
Secondary data source: LIGO-Virgo

Supernovae
Expected frequency: 100/day
Expected duration: : hours to months
Desired response time: : minutes (less than 1hr)
Primary data source: LSST Level 1 stream
Secondary data source: External transient surveys

Tidal Disruption Events
Expected frequency: 1/day
Expected duration: days-months
Desired response time: minutes (less than 1hr)
Primary data source: LSST Level 1 stream
Secondary data source: External transient surveys

AGN Extreme Variables (both slow and fast evolving)
Expected frequency: XXXX
Expected duration: : weeks-years
Desired response time: days-weeks
Primary data source: LSST Level-1 stream and LSST Level-2 data; possibly more than two latest releases!
Secondary data source: multi-wavelength surveys from radio to high energies

Stellar microlensing
Expected frequency: XXXX
Expected duration: : XXX
Desired response time: XXX
Primary data source: XXX
Secondary data source: XXX

Eruptive Stellar Variables
Expected frequency: XXXX
Expected duration: : XXX
Desired response time: XXX
Primary data source: XXX
Secondary data source: XXX

Flare stars and general stellar variability
Expected frequency: XXXX
Expected duration: : XXX
Desired response time: XXX

Primary data source: XXX
Secondary data source: XXX

Periodic Variables
Expected frequency: XXXX
Expected duration: : XXX
Desired response time: XXX
Primary data source: XXX
Secondary data source: XXX

Acknowledgments

A wide variety of useful comments came from members of the LSST:UK SWG, both by email and in various project meetings.

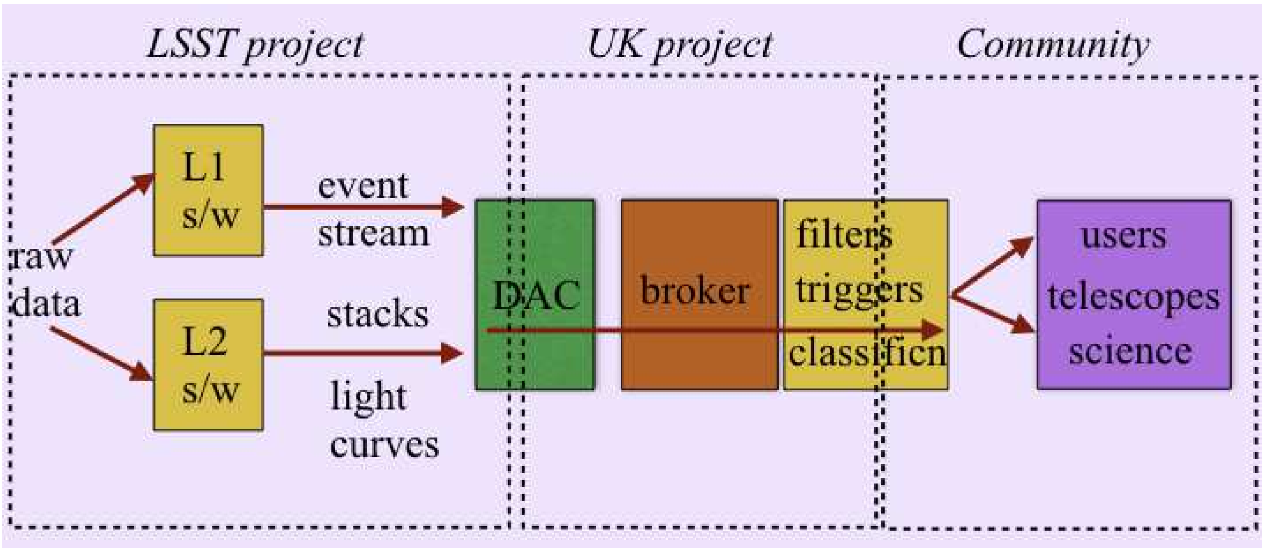

Figure 1. Overview of data flow, with an indication of the division of responsibilities

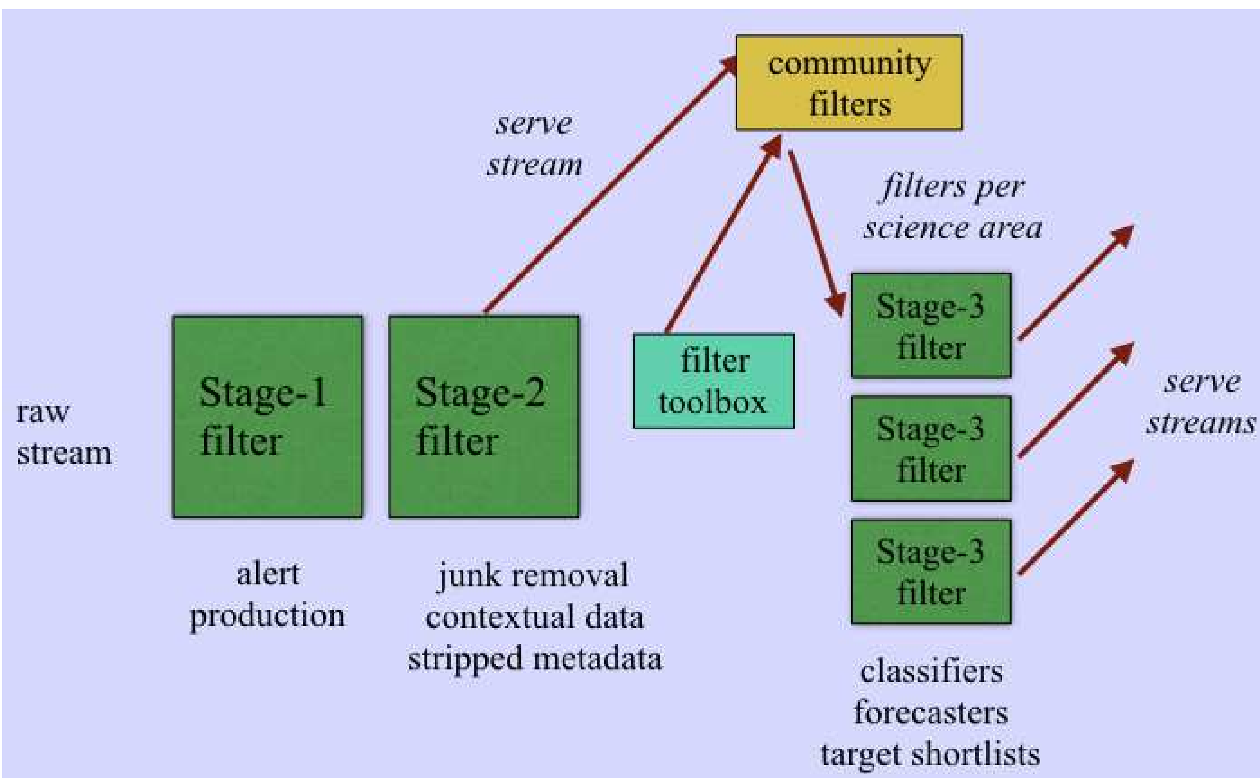

Figure 2. Possible stages in the filtering/triggering process.