UNIVERSITY OF BIRMINGHAM

TAG Updates and News

Software and Computing Workshop, 30th November, 2010

Mark Slater, Birmingham University

Current TAG Usage

TAG files are *small (~1KB per event), metadata files* with ~200 keywords/variables that can be used to make a basic pre-selection

They can then be used to *'point back' to their parent AOD, ESD or RAW files*

TAGs in ATLAS are generally used in one of three ways:

- *Applying a TAG query to a TAG DQ2 Dataset*

This involves using the TAG datasets as input and specifying the TAG query and input file type (AOD, ESD, RAW) you wish to run over in the job options

- *Using ELSSI to produce a TAG file*

This uses ELSSI to generate a TAG file that only contains the references to events and files that pass your TAG query. This is then used in the grid jobs.

- *Using the Skimming Service in ELSSI*

Again, use ELSSI to generate a TAG file but then have the built in Skimming Service run the appropriate grid jobs to skim out your own AODs/ESDs

UNIVERSITY OF BIRMINGHAM

TAG DQ2 Datasets

This workflow is designed to save time on occasions where a user has *a small number of events in each file of a dataset* and so will generally end up 'touching' each one

TAG Datasets are *automatically generated and replicated* by DDM and can be submitted to as a *normal dataset* (provided the JOs are setup correctly)

Marcin from the TAG group has just produced a tool that can link *TAG datasets to their AOD/ESD/RAW counterparts* at submission which solves the only issue specific to this workflow

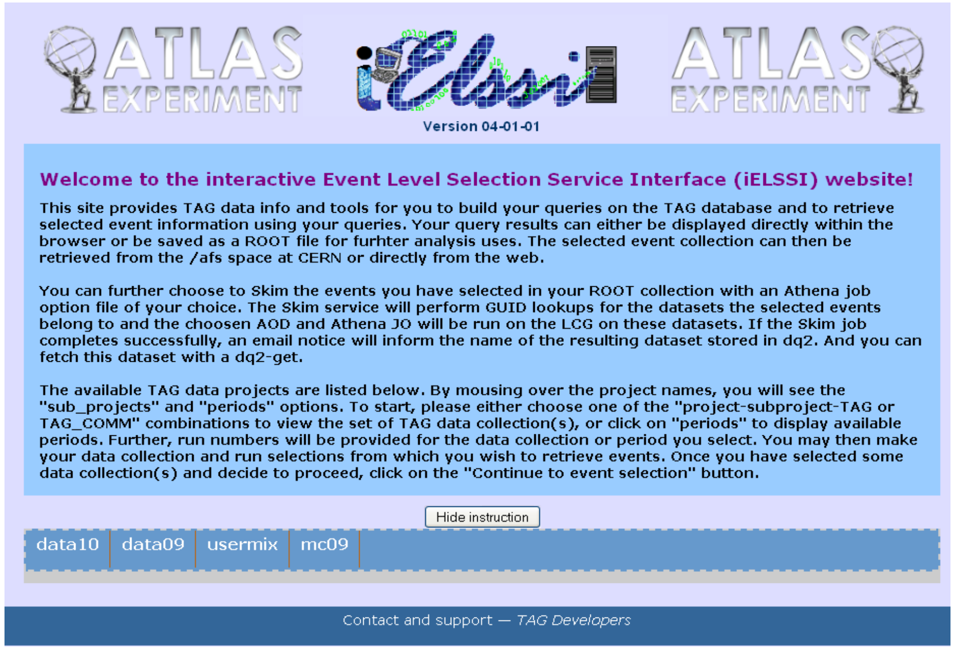
With the exception of some improved interface support, I don't see any more significant development required here

ELSSI/User Produced TAG files

This workflow is designed to save time on occasions where a user has *events over a small subset of a dataset* and so will only need to run over a few files

Using the ELSSI web interface, users can *run queries selecting on TAG properties* over various streams of the data

A TAG file is produced from this selection that *only references the events* that pass the TAG query – *see talk on Thursday by Elisabeth Vinek*



Version 04-01-01

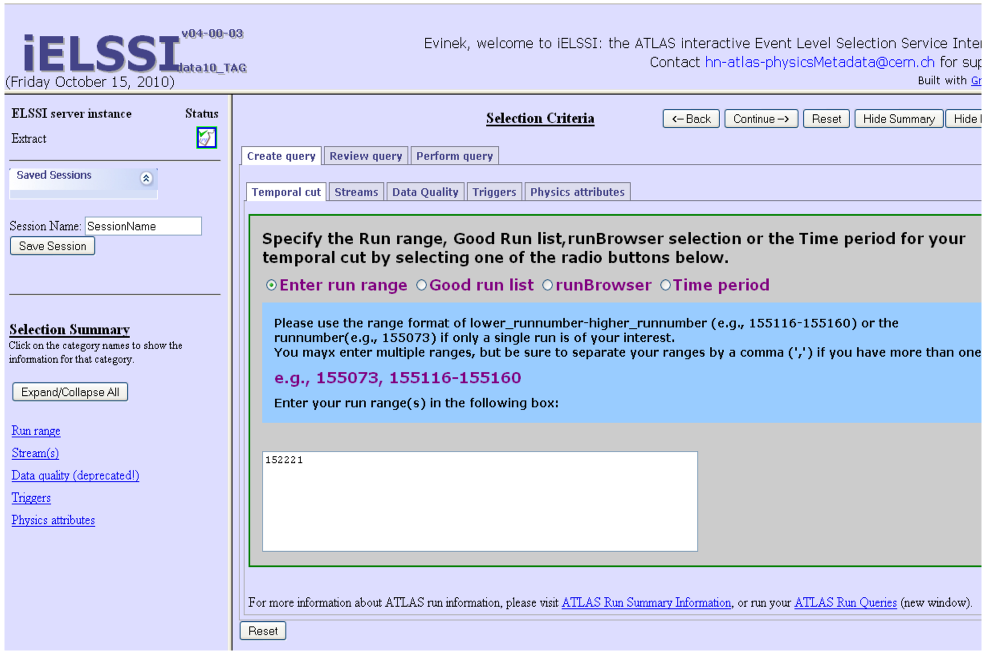
Welcome to the interactive Event Level Selection Service Interface (iELSSI) website!

This site provides TAG data info and tools for you to build your queries on the TAG database and to retrieve selected event information using your queries. Your query results can either be displayed directly within the browser or be saved as a ROOT file for furhter analysis uses. The selected event collection can then be retrieved from the /afs space at CERN or directly from the web.

You can further choose to Skim the events you have selected in your ROOT collection with an Athena job option file of your choice. The Skim service will perform GUID lookups for the datasets the selected events belong to and the choosen AOD and Athena JO will be run on the LCG on these datasets. If the Skim job completes successfully, an email notice will inform the name of the resulting dataset stored in dq2. And you can fetch this dataset with a dq2-get.

The available TAG data projects are listed below. By mousing over the project names, you will see the "sub_projects" and "periods" options. To start, please either choose one of the "project-subproject-TAG or TAG_COMM" combinations to view the set of TAG data collection(s), or click on "periods" to display available periods. Further, run numbers will be provided for the data collection or period you select. You may then make your data collection and run selections from which you wish to retrieve events. Once you have selected some data collection(s) and decide to proceed, click on the "Continue to event selection" button.

Contact and support — *TAG Developers*



(Friday October 15, 2010)

Evinek, welcome to iELSSI: the ATLAS interactive Event Level Selection Service Inte
Contact hn-atlas-physicsMetadata@cern.ch for su
Built with Gr

ELSSI server instance Status
Extract

Session Name: SessionName

Selection Summary
Click on the category names to show the information for that category.

Specify the Run range, Good Run list,runBrowser selection or the Time period for your temporal cut by selecting one of the radio buttons below.

[✓ Enter run range] Good run list runBrowser Time period

Please use the range format of lower_runnumber-higher_runnumber (e.g., 155116-155160) or the runnumber(e.g., 155073) if only a single run is of your interest.
You mayx enter multiple ranges, but be sure to separate your ranges by a comma (',') if you have more than one

e.g., 155073, 155116-155160

Enter your run range(s) in the following box:

152221

For more information about ATLAS run information, please visit ATLAS Run Summary Information, or run your ATLAS Run Queries (new window).

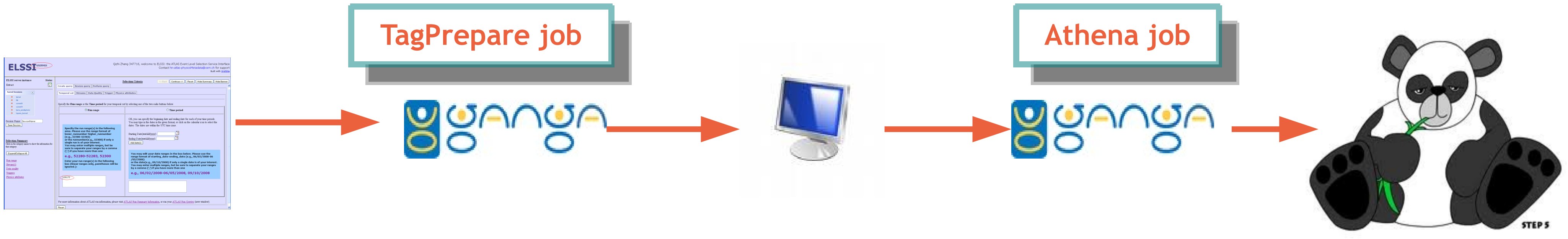
UNIVERSITY OF BIRMINGHAM

ELSSI/User Produced TAG files

Obviously to run over these specific events on the grid, you not only need to know the referenced files at submit time but *also at run time for each job*

This adds a certain amount of overhead to both the *submission* (when finding the references) and the *file transfer* (shipping the TAG files themselves)

For Ganga, we get around this by using a *pre-submit step* that sorts out the links between the TAG file and the data files and splits the TAG files up so they only reference the events used by a job



UNIVERSITY OF BIRMINGHAM

ELSSI/User Produced TAG files

For small files (~10MB) this system works perfectly well. However, it's very easy (and probably necessary) to produce TAG files *a lot larger (~1GB)*

To try to get around this problem, we have introduced a *compression step* to the TagPrepare job that gives a *compression of factor of ~100* to the split TAG files

This compressed TAG file(s) can then be sent with the job easily (and re-used if necessary) where it gets inflated on the worker node

This could in principle *be adapted to pathena* by either the addition of a new tool or performing the above compression at submission time

However, this could add *1-10 mins to the submission though...*

UNIVERSITY OF BIRMINGHAM

Possible Improvements

There are two proposed improvements to this system:

- *Find references to TAG datasets and send the query*

If ELSSI could also provide GUID info for the TAG datasets as well as AOD, etc. the query could be sent and run over these datasets like normal TAG DQ2

- *Reconstruct the TAG file in the build job*

Have the build job re-run the query (using ELSSI) and use that info to reconstruct the TAG file appropriate for the jobs associated with it

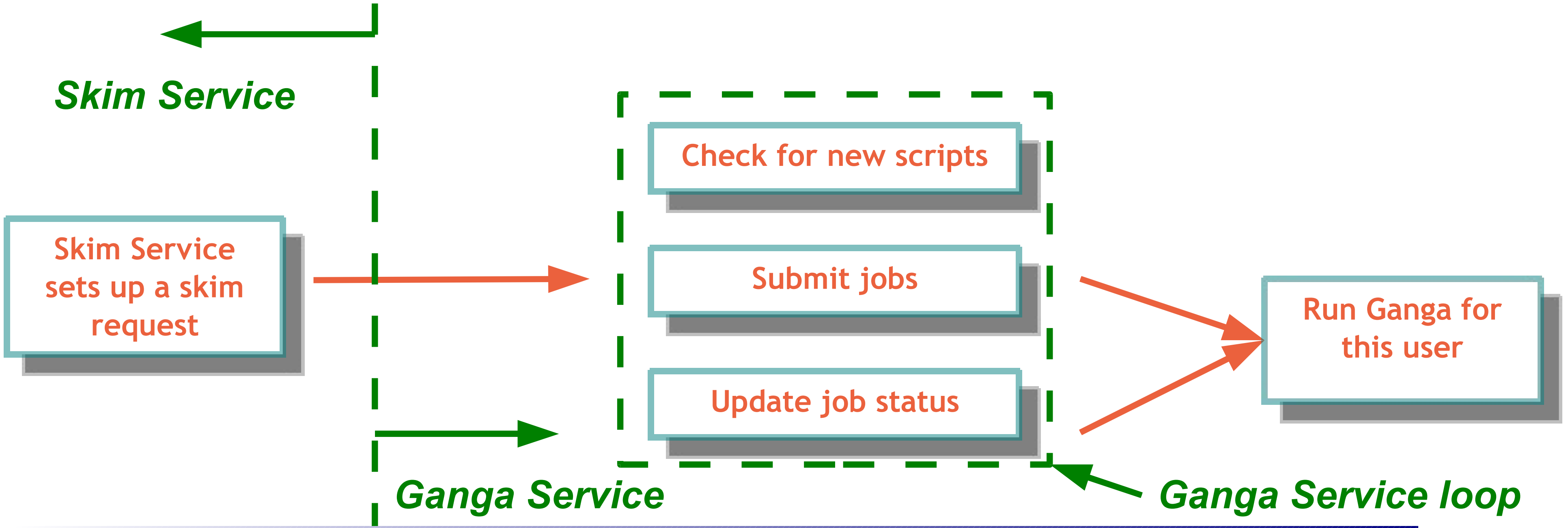
I plan to investigate these options in the next month or so in order to simplify the interface for the user

The ELSSI Skimming Service

Finally, there is the Skimming Service that is part of the ELSSI web interface which can be used to *create personal AOD/ESD skims* of the selected data

Though the Ganga Service is generic, in this case it is only provided *specific 'skimming' job options* given the ELSSI extraction output and other options provided by the user through the web interface

The system can *handle multiple users* and provides a *web page* that tracks the jobs and *emails the user* when the skim is complete



Usage and Planning Overview

Current usage is small but increasing:

- Judging from DAST mails, the most used at present are the TAG Datasets
- However, more are starting to use ELSSI generated TAG files
- I believe a much greater move will happen next year with more data
- In addition, there is increased interest in the Skimming Service

Now that the workflows are tested, I plan to develop the 'interface':

- *Improving site selection for Direct Access jobs*

Direct Access is essential for TAG and finding working sites is tricky. More sites need to support it (for TAG at least) and submission needs to take this into account

- *Decide on and implement a simplified 'TagPrepare' system*

The most efficient and user-friendly process needs to be found and implemented

- *Make it easier to make JOs 'TAG-ready'*

Users have difficulty setting the JOs properly to access TAG (esp. with inputFilePeeker). The tools should help with this.