GammaLib - Change request #2120

Improve access time of keyword values in GFitsHeader

05/30/2017 12:43 PM - Cardenzana Josh

Status:	Closed	**Start date:**	05/30/2017
Priority:	Low	**Due date:**	
Assigned To:	Cardenzana Josh	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	1.3.0		

Description

While looking into the simulated data challenge, I've noticed that it takes a very long time to loop through all of the data files. Particularly, it has taken over 40 minutes to run csobsselect on my machine when passing it the full list of observations. I've tracked this speed issue down to the reading in of the FITS header keywords, specifically the keywords associated with the Monte Carlo source name and ID (MMN##### and MID#####). Since there are over 1500 models and each observation contains the full list of source names and IDs, this read in takes a very long time.

Currently, the keyword values from a FITS file are stored in the GFitsHeader class as a vector of GFitsHeaderCard objects. This means that when a given keyword is needed the class has to loop through the entire vector until it finds the desired value. For most implementations where there are only a few simulated models in an observation this is reasonably fast, however in the event that a given observation has MANY keywords (as in the case above) this can slow down the loading of observation files. Changing the storage of the keywords from 'std::vector<GFitsHeaderCard>' to 'std::map<std::string, GFitsHeaderCard>' has the potential to speed up the lookup time, since std::map stores a sorted version of its entries and uses a binary search algorithm to lookup items.

After talking with Jürgen, he has changed how the observations store the simulated source names and IDs so that each observation FITS file only contains header keywords associated with the models actually simulated within that observation. This will probably dramatically improve the time issue I'm seeing.

History

#1 - 05/31/2017 05:13 PM - Cardenzana Josh

- Status changed from New to Pull request

- % Done changed from 0 to 50

I've added a new member to the GFitsHeader class:

std::map<std::string, GFitsHeaderCard*> m_keyname_map

which stores a reference to a given GFitsHeaderCard in 'm_cards' which is accessed by the card's keyname. It should be noted that 'm_cards' is capable of storing multiple instances of a given keyword (if that keyword is "COMMENT" or "HISTORY"), however 'm_keyname_map' will only have a single instance of these keywords. Dispite this, because the original implementation worked by looping over the entries in 'm_cards' until a given keyname is found, the behavior should be unchanged. The only change is when accessing the header keyname value via the keyname itself. Accessing the value by index still directly queries the 'm_cards' vector.

Comparing the runtime of 'csobsselect' between the current 'devel' branch and the above update on the recent iteration on the data challenge GPS observations gives the following results:

devel:

```
real    43m35.192s
user    43m9.987s
sys     0m12.940s
```

#2120 update:

```
real    4m7.695s
user    3m54.809s
sys     0m6.732s
```

#2 - 06/06/2017 04:16 PM - Knödlseder Jürgen

- Status changed from Pull request to Closed

- Target version set to 1.3.0

- % Done changed from 50 to 100

Merged into devel