Gemini Science Archive: Meta-Data Database Definition

Colin Aspin, GSA Scientist 4 January 2003

This document should be considered a preliminary overview of the requirements for the MDD and hence, should form the basis for further discussion.

1. Overview of Data Pathway

The Gemini Science Archive, the GSA, will involve a flow of data and associated information, from the instruments/data-handling system (DHS) plus numerous other sources into the archive. The meta-data database, the MDD, located at the Gemini Observatory will collect the ancillary (non-scientific data) data associated with the science datasets. The first-pass contents of the MDD are described in detail below. The MDD should be replicated from Gemini to the Canadian Astronomical Data Centre (CADC) frequently (at least daily, perhaps more often) where it will be ingested into the CADC GSA and associated with each dataset. The facility should exist update already existing MDD data and re-associate it with the appropriate dataset. The GSA software will only allow datasets to be made available to the community once the Gemini-defined proprietary period has expired and the data is considered ‘public’. Data whose proprietary period has not yet expired will not be visible to general users of the GSA but will be available, in a secure manner, to Gemini Observatory staff. Publically available data will be searchable via a web interface and users will be able to obtain previews of all available datasets on-line upon request. The above data pathways define, and detailed at least in part, in the GSA Phase II development documents.

2. The Meta-Data Database, the MDD

The MDD is a fundamental component of the GSA. It will contain all relevant information needed to fully define a Gemini dataset and hence will bring together information from many areas of operation from data headers, to instruments, telescope and onto the environment. The MDD contents will be collected from several sources and will be stored automatically within the MDD. The MDD will be a Sybase database and will be replicated to CADC where ingest into the archive will take place and associations made between meta-data and Gemini datasets. A list of first-pass meta-data entries in the MDD is given below:

i) FITS header information	(defines the observations/dataset)
ii) Gemini Engineering Archive Data	(environmental, telescope info)
iii) Observing logs	(user info, comments etc)
iv) Observing programs	(Observing Tool files)

Item i) above is obtained from the data FITS files themselves. Item ii) is obtained from the Gemini Engineering Archive, the GEA. Item iii) is obtained from the Remedy Nightlog/Observing Log database. Item iv) is obtained from the Observing Tool (OT) database.

Examples of meta-data that will be stored in the database include:

Telescope and Instrument configurations: these are read from the dataset FITS headers, interpreted using the data dictionary and combined with additional parameters from the GEA on telescope setup, wave-front sensor (WFS) setup, dome configuration etc.

Environmental data: these are taken from a combination of data from the Gemini weather tower (e.g. temperature, wind speed/directions, humidity, barometric pressure etc) internal dome sensors (e.g. dome temperature, M1/M2 temperature etc), WFS seeing estimations, and external weather servers (e.g. JCMT water vapor measurements, CFHT transparency measurements, radar/IR maps of the Big Island etc). *All this data should be obtained from the GEA archive.*

Proposal information: these data are from the OT database and include specifics of the way the data were acquired (e.g. mosaic patterns) and the on-line pipeline data reduction (e.g. reduction procedures, calibration used etc).

Nightlog and obslog files: These are from the night's observing and define in more detail the sequence of data acquisition during the night including comments on such things as data quality and useful facts associated with the data.

3. MDD data sources, subsystems and interfaces

Below we detail the sources of the data to be ingested into the MDD together with relevant information on the sub-system involved and software interfaces requiring to be accessed.

FITS Header Information:

Each data FITS image coming from a Gemini facility instrument will have a set of header items that are required to be ingested into the MDD and associated with the dataset. These items will also define time instances that will serve as timestamps for the association of other MDD information from other sources (e.g. the GEA). The FITS headers should be accessed via the DHS database and should be allowed to be updated in case of FITS file header modification post-observing via re-ingestion into the DHS database (once corrected). This functionality needs to be included in the DHS code and a mechanism made available update the MDD version of the FITS headers and subsequently the additional supportive data (e.g. GEA data). The MDD access to the DHS database should be transparent to CADC since they wrote the DHS code. Additional modifications may be needed to this code under the DHS support contract currently being negotiated or in the GSA Phase II contract.

The Gemini Engineering Archive

It is envisaged that all data defining the environment of the observations e.g. weather information, internal environment information, telescope sub-system performance information, be included into the GEA at the time of data acquisition. This results in the requirement that the MDD need only interface to the GEA for this data (this is my vision, I need to discuss it with the software team and management but I don't foresee any major problems). The GEA is a Sybase database that can be queried with standard SQL queries. A complete list of items required to be stored in the MDD from the GEA (currently existing and yet to be implemented) is not yet available, however, the access method will be the same, i.e. via SQL queries.

- It is clear that the MDDB will have to poll the GEA for new data. This can be done using a datetime field in the GEA given the following assumption.
- What are you referring to when you say "poll"? Is this polling the GEA on a pre-defined time interval to get the required information? If yes then I would say MDDB does not need to poll since GEA will do the polling and store all required data internally. MDDB just needs to access the correct information in GEA once per day or whatever interval is reasonable. In fact, Craig Allen, the GEA software engineer seems quite taken with the idea of GEA populating the MDDB, think about that one.
- We are making the assumption that data within the GEA is never modified. If this assumption is wrong, then the update mechanism will have to be more complex and may require changes within the GEA to support the MDDB.

- GEA may be modified. We would have to modify GEA to indicate in some way that an entry has been updated and MDDB should re-access/grab it. If the GEA updated MDDB then this could be done automatically.
- Please add the GEA database table schema to this document.
- Apparrently no real database table schema, Craig says it is a narrow table format with time stamp. I'll attach a document describing GEA, this might help.

Please add the list of currently existing GEA items as an appendix to this document~~.~~

- This is very incomplete at present, we need to define and add all MDDB required GEA elements. This should be internally done by GEA software engineers.

The Observing/Nightlogs

The current nightlog, written by the nighttime astronomer is ingested into the Remedy system at the end of each night. This narrative log is stored via Remedy in a Sybase database as multiple database entries (rather than just one text file). These entries not only relate the timeline of the night's observing but detail such things as time lost to faults, seeing conditions and weather related problems. The Remedy database can also be accessed directly via SQL queries.

- Please add the nightlog database table schema to this document.
- We would like the nightlog to be stored in MDDB as a single document and not accessible to the users but accessible to the Gemini staff. This should simplify the matter of nightlog ingestion. The nightlog would be an ascii file.
- Please identify which parts of the nightlog are to be transferred to the GSA (i.e. are there fields that Gemini feels are internal to Gemini).
- See above
- Are nightlog entries ever edited after the initial data entry? If so, is there a query-able field in the schema that allows one determine when an entry was last updated?
- They are sent to Remedy, modified to show the correct time lost stats and then left. The modified nightlog is the one we would want in MDDB.

The observing log or obslog, is currently a handwritten text document that lists the observations taken and instrument setup independent of the FITS header information. It is, at this time, a standalone text document located in a specific directory on the Sun Solaris computer network. Useful items from this obslog include such things as comments on setup on particular astronomical targets (e.g. offsets to get objects down slits etc) and instrument performance information (e.g. Slit positions on the science detectors).

- Are these files ever edited after creation? If so, we will need a mechanism to detect modifications?
- Yes, soon after creation they are edited to reflect the correct observation sequence. After that, no, they are not edited.
- How can we get the start and end date/time of an observing log? Is in the filename, in the directory name or in an easily-parse-able location within the file?
- UT date is in the filename, should be all that is required.
- Is there information within a file that Gemini feels are internal to Gemini and should never be copied to the GSA?
- Obslog, no, details are public, nightlog, yes, all internal information.

The format and construction of the obslog (and possibly the nightlog) will change when the Observatory Control System (OCS) is implemented at the telescope for control of data acquisition at many levels. The exact nature of these changes are, as yet, unknown due to the continued development of the OCS. Once possible scenario is that, the obslog (and possibly nightlog) will become part of the observing program database (see below).

- We cannot design to an unknown spec. If the changes are documented then we will need to see the design. Otherwise we have to exclude this from the initial MDDB work and build a mechanism based on the obslog files described above.
- If you just expect to ingest into MDDB two ascii files then that is all you need to do. No spec information should be needed just ingest two ascii files and associate a date to them so they can be associated with that night's data.

The Observing Program

The observing program is an xml document that defines the observation resulting in the archive dataset. Currently these are stand-alone text files that are executed manually via sequence executor software. In the OCS software, these will be stored in a Java database and selected for execution via a program browser. The GSA MDD should be able to access the Java database to archive the xml sequence responsible for each dataset. The program ID will be associated with the xml sequence and will also be in the FITS header of the dataset.

- If the OCS implementation is going to be deployed soon and the following questions can be answered, we can design a mechanism to use the Java database:

    - What kind of database is this (Objectivity, …)?
    - What tools does Gemini have (that the GSA could use) to retrieve and possibly parse an Observing Program?
    - Can we query the database to identify a completed observing program for a given program ID (those are the ones that get copied to the GSA)?
    - Does an observing program ever get modified after it is completed? If so, we will need a mechanism to detect modifications?
- Or do we design to ingest the stand-alone files:
    - Does the filename contain the program ID or do we have to extract if from the xml file?
    - What tools does Gemini have to extract a program ID from the file?
    - Where is the repository of these files? Is it the archive or a special directory within Gemini?
    - In answer to all these questions, it seems best to ingest the xml observing programs at the end of each semester, once observing on them (and modifications to them) are complete. The programs could easily be obtained from a disk directory repository where each program has an xml file identified by the observing program ID. This would remove all the worries about Java databases and updating programs etc. How does this sound?

4. Polling and the MDD

It is not entirely clear that the MDD will need to poll any Gemini sub-system for data. The GEA is already setup to poll sub-systems, specifically, at present, Epics records on user-defined time intervals. It would seem appropriate that the GEA polls such items as environment data, weather information etc and store it at the frequency defined by the observations and information update. This suggestion has not been approved as yet but makes sense from the point of view of completeness of the GEA. The MDD then would only need to poll the GEA for this type of data and not external resources such as satellite map sites and weather towers.

5. Additional information: future additions

Additional items considered required in the MDD are, for example, publication information. Since this is an external resource, i.e. the ADS reference and citation indices, I consider them extraneous to the current workscope and should form part of GSA advance capabilities unless this, and another external resource can be archived by the GEA and hence, entered into the MDD via GEA SQL queries~~.~~

From Severin.Gaudet@nrc.ca Mon Feb 17 14:31:47 2003
Date: Thu, 6 Feb 2003 12:10:49 -0800
From: "[ISO-8859-1] Séverin Gaudet" <Severin.Gaudet@nrc.ca>
To: Colin Aspin <caa@gemini.edu>
Cc: "[ISO-8859-1] Jean-René Roy" <jrroy@gemini.edu>,
     David Schade <david.schade@nrc.ca>,
     David Bohlender <david.bohlender@nrc.ca>
Subject: Re: Hi

[The following text is in the "ISO-8859-1" character set.]

[Your display is set for the "US-ASCII" character set.]

[Some characters may be displayed incorrectly.]

Colin

We have finally had everyone in the same room to discuss your document received on January 29. We are now comfortable with the scope of the MDDB work and judge it to fit within the time allocated in the proposal. We will meet with our contracts officer this afternoon to begin moving the process toward signature.

We have not been totally inactive on the GSA front. We interviewed candidates for the senior GSA position and we will be interviewing candidates for the junior position next week.

The next task is to address the DHS contracts. I assume the priority is the support contract as opposed to the development contract - is this correct?

The target date for a first visit to discuss GSA is March but we will know more once the hiring process is completed.

Séverin