Gemini Science Archive: Top Level Requirements

Colin Aspin, GSA Scientist

4th January 2003

1. Summary

The Gemini Science Archive, the GSA, is the primary method of access to Gemini Observatory non-proprietary facility instrument data by the worldwide astronomical community. All datasets, together with all necessary ancillary information required to fully define a specific observation will be archived and readily accessible to users via a simple web-based interface. The archive will collect and store information from many different sources and will associate and cross-reference this 'meta-data' with each observation for quick and easy access. The meta-data database is the most important and fundamental aspect of the GSA in that it brings together all possible information about an observation to completely describe it and the conditions of all monitored systems around it. It is a goal to populate the archive automatically and hence requires minimal daily support and maintenance. The top-level goal of the archive is to make Gemini data easily and quickly accessible by the community so the best use of the information contained therein can be made.

2. Overview of Data Pathway

The GSA will involve a flow of data and associated information, from the instruments/data-handling system (DHS) plus numerous other sources into a meta-data database, the MDD, located at the Gemini Observatory. The contents of the MDD are described in detail below (section 3). The MDD will be replicated from Gemini to the Canadian Astronomical Data Centre (CADC) daily where it will be ingested into the CADC Gemini archive and cross-referenced to associate each dataset with all relevant meta-data. The GSA software will only make datasets 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. In addition, complete datasets, including science data, calibration data and ancillary information in the MDD, will be cross-referenced by discovery agents (section 4) and made available to GSA users upon request either via 'hard-media' (see section 5) or, for smaller datasets, via network access.

3. The Meta-Data Database

The MDD is the fundamental component of the GSA and upon which all GSA features will be built. It will contain all relevant information needed to fully define a Gemini dataset bringing together information from many areas of operations from data, to instruments, telescope and onto the environment. The MDD contents will be collected from numerous sources and stored automatically within the MDD. The MDD will be replicated at CADC where ingest into the archive will take place and associations made between meta-data and Gemini datasets. A list of proposed meta-data entries in the MDD is given below:

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

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

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

a) 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.

b) 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).

c) 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).

4. Discovery Agents

The GSA will utilize the MDD using a set of discovery agents to search the database for associations using Gemini defined rules. An example of such a discovery agent and its rules might be the association of calibration data with observational datasets. It can be envisaged that this agent would scan data from the previous night automatically and try to associate bias images, dark images, flat-field images, standard star images, spectroscopic/telluric standard images etc by using rules that define, for example, the same instrument configuration must have been used for both and science targets be defined as OBJECT while calibrations be defined as BIAS, DARK, FLAT or CAL. A number of discovery agents can be envisaged to simply cross-reference and archive the datasets with the MDD content.

5. Hard-Media Distribution

Hard-media consists of media written by the archive to to sent (physically) to the GSA user. The current hard-media we consider necessary are:

CD-ROMs	(for datasets < 600Mb)
DVD-Rs	(for datasets 600Mb -- 4Gb)
DDS-3/4 DAT tapes	(for datasets > 4Gb)

6. Critical Requirements

There are several areas where there exist critical path items for the implementation of the GSA.

6.1 The GEA: work on the GEA is proceeding but, as yet, the first pass of items to be included in the GEA (from both Epics records and external sources) is incomplete. To expedite this, we can provide a first-pass set of monitoring items for the GEA to make the GSA functionally useful for the GSA. In addition, a way of triggering the capture and storage of GSA specific items in the GEA would be a very useful feature since many GEA items need to be time critically related to a dataset.

6.2 Data Dictionary: this needs to be defined from the current list of FITS header items in the data files from each facility instrument.

6.3 Nightlogs/Observing Logs: A standardized way of producing nightlogs (a narrative text of the night's progress with comments on the observations acquired) and observing logs (a time sequence of observations taken with telescope/instrument parameters recorded) is required. The nightlog part of this requirement is being addressed by the web-based nightlog system soon to be tested at Gemini North. The observing log is currently defined at the end of each night by a standalone process that scans data images. This can be auto-created by either the MDD or, pseudo- on-line by a system process.

6.4 OT Database: This is being addressed as part of the observatory control system (OCS) and will soon (timescale?) be the way in which observing programs are defined, stored, retrieved and executed at the telescope. Access to this database from the MDD server is required.

6.5 DHS Database: Only selected FITS headers items from the data images are currently ingested into the DHS system. Ideally, this ingestion needs to be extended to include all FITS header keywords. Some additional work on both the DHS software and hardware is required to do this efficiently.

7. Cross Database Activities

We require the GSA MDD to cross several database boundaries in its operation. These are:

i) The GEA Database
ii) The Remedy Database
iii) The Observing Database
iv) The DHS database