Data Management Plan for Gemini Observatory

Data and Research Products

Gemini Observatory generates astronomical images and spectra, their required calibrations, metadata[1] for data header entries such as observing conditions (cloud cover, sky brightness, etc.) and engineering data. To date, more than 36 Tbytes have been produced and archived.

Science data are typically published in the peer-reviewed literature and in conference proceedings. Gemini monitors all publications that make use of data obtained at the Observatory and records these in a publications database that can be queried online[2]. Principal Investigators (PIs) are responsible for publishing their results, and for acknowledging the Observatory and its funding agencies. Gemini tracks all publications, and Partners may use this information in their national time allocation system.

Data Rights Policies and Public Access

Gemini Observatory's current Data Rights and Proprietary Periods are described on its web page[3]. In general Principal Investigators retain exclusive use of data taken at Gemini for 12 months, with exceptions for Fast Turnaround Proposals (6 months), System Verification/Demo Science (2-3 months depending on the specific Call for Proposals), and a few other limited cases. Thereafter, the data become publicly available. Gemini's observing system requires appropriate calibrations to be defined along with science targets and produces records of data quality and sky conditions. The archive is therefore broadly useful to the scientific community beyond the original investigation team. Moreover, the Gemini Observatory Archive[4] (GOA) can be queried by various parameters (including target name, target position, date, instrument, etc.) to identify science observations of interest. These policies and provisions are independent of the archive system used.

Data Format and Integrity

All Gemini science data are stored as Flexible Image Transport System (FITS) files, the astronomical standard. Gemini data formats are stable, documented[5] and compliant with the FITS standard. Gemini FITS headers are likewise stable, a prerequisite for legacy usage via the science archive. By stable, we mean that once a facility instrument is in operations, the FITS headers change very little (occasionally a header item may be added, e.g., we added Principal Investigators' requested conditions to the headers).

Data quality assessment and observing condition flags are set by Gemini staff and allow external users of data in the science archive to assess the quality of Gemini data. Web pages are available to assist scientists in using archive data[6].

Gemini Observatory Archive

The GOA stores and distribute data to users, with protected and authenticated access to data during the proprietary period, and open access thereafter. The GOA is provided through an AURA/Gemini-developed web interface deployed on a virtual Linux server, with bulk data storage on Amazon Web Services (AWS) Simple Storage Service (S3). Science data and raw calibration data are being transferred in essentially real time to the AWS S3. Processed calibrations are stored, along with data taken for engineering purposes. The user-friendly interface provides

---

[1] Metadata includes information directly from instrument FITS headers, additional derived science metadata, such as spectral range, spectral resolution, and Galactic coordinates, and publication metadata input by Gemini staff.

[2] http://www.gemini.edu/apps/publications-users/

[3] https://www.gemini.edu/sciops/observing-gemini/data-rights-and-distribution

[4] https://archive.gemini.edu/

[5] http://www.gemini.edu/sciops/data-and-results?q=node/10794

[6] http://www.gemini.edu/sciops/data/dataIndex.html

automatic association of calibration files with science observations, and independent searches on calibration files are also possible.

The live storage at S3 is redundantly stored across multiple facilities and multiple devices in each facility. Data are validated on upload and retrieval to detect any corruption. The S3 system is designed to sustain the concurrent loss of data in two facilities. Additionally, to further guard against data loss, we use the AWS Glacier facility to create and maintain off-line backups of the archive data stored on AWS S3 with similar redundancy, multi-site protection, and verification.

Local Data Storage

In addition to the GOA, Gemini maintains multiple local data backups. First, all local on-line science data is stored on RAID arrays to protect against data loss from individual hard disk units.

The FITS storage system backs up all FITS data catalogued at Gemini to LTO tape. The system purges old data from local data stores only when the data are known to have been successfully written to two separate LTO tapes, stored in separate locations.

Additionally, disk-to-disk “mirroring” of all Tier I (NetApp) storage is carried out between the base facilities and the summits in both the north and south.

Finally, Gemini operates enterprise-level backups, which include all operational summit systems along with administration, engineering, etc., and these enterprise-level back-ups also capture science data. Access to all enterprise backup systems and data is restricted to select staff members of the Software/Information Systems group.