Gemini Science Archive

Operational Concept Definition Document

(Revised Initial)

Version 04

21 November 2000

Norman Hill, Séverin Gaudet, Daniel Durand, David Schade, David Bohlender

National Research Council of Canada
Herzberg Institute of Astrophysics
Canadian Astronomy Data Centre

Felipe Barrientos, Felipe Richardson

Proyecto Gemini-Chile
Comisión Nacional de Investigatión Científica y Tecnológica

Table Of Contents

Gemini Science Archive

Chapter 1 *Introduction*

1. Purpose

This document describes the operational concept for the GSA. It comprises the science cases which drive the design, the operational scenarios which describe how an astronomer would retrieve the data required to do each of the science cases, and the user requirements which arise from the science cases. The focus of this document is on the functionality an archive user wants and needs from the GSA.

It is important to note what this document does **not** address. It is not intended to address implementation and design issues like cost, practicality, and how a given function should be implemented. A detailed examination of the design, implementation and cost issues raised by the user requirements will be done in later documents. Also this document is not intended deal with the requirements defined by Gemini, they will be covered in the detailed description of the functional and performance requirements which will be given in [7].

This document is an amalgamation and expansion on several other documents, including [4], [9], [10], and [11].

2. Overview

This document is divided into the following chapters:

- Chapter 1 is an introduction to this document.
- Chapter 2 gives a general description of the GSA, its reason for existence, and describes how it fits within the Gemini Telescope systems.
- Chapter 3 describes the science cases justifying the creation of the GSA. These science cases are based on current uses of astronomical archives, and the expected uses of the GSA if new capabilities are added. The science cases from were used to create the operational scenarios in Chapter 4 showing the interaction with the archive required to do the science.
- Chapter 4 describes the operational scenarios for the GSA, showing the interaction between the GSA and an archive user, required to collect the data needed to do the science described in Chapter 3. The operational scenarios were then used to derive the science requirements for the GSA described in Chapter 5. Tables are included in Chapter 4 linking the operational scenarios with the science cases upon which the scenario is based.
- Chapter 5 lists the user requirements that arise from the operational scenarios described in Chapter 4. Tables are included which link the requirements with the operational scenarios that depend on the requirement.
- Chapter 6 list the reviewer comments and our responses given during the OCDD review video conference held on May 26, 2000.

3. Abbreviations and Acronyms

API	Application Programming Interface.
ASCII	American Standard Code for Information Interchange.
BDS	The abbreviation used for the Gemini Science Archive bulk data storage subsystem.
CADC	Canadian Astronomy Data Centre.
CAT	The abbreviation used for the Gemini Science Archive catalogue subsystem.
CD-R	Compact Disc Recordable.
CD-ROM	Compact Disc Read Only Memory.
CoI	Co-Investigator (see glossary).
DAT	Digital Audio Tape.
DBMS	Data Base Management System.
DHS	Data Handling System.
DI	The abbreviation used for the Gemini Science Archive data ingest subsystem.
DP	The abbreviation used for the Gemini Science Archive data processing subsystem.
DPR	Data Processing Recipe (see glossary).
DR	The abbreviation used for the Gemini Science Archive data Retrieval subsystem.
DVD	Digital Video Disk (sometimes referred to as Digital Versatile Disk).
DVD-R	Digital Video Disk (sometimes referred to as Digital Versatile Disk), recordable.
EMSS	Einstein Medium Sensitivity Survey.
ESA	European Space Agency.
FITS	Flexible Image Transport System.
FPRD	Gemini Science Archive Functional and Performance Requirements Document (see glossary).
FTP	File Transfer Protocol.
GB	Gigabytes (1073741824 bytes).
GEA	Gemini Engineering Archive (see glossary).
GSA	Gemini Science Archive (see glossary).
GSC	Gemini Science Committee.
GSCG	Gemini Software and Controls Group.
GSRS	Gemini Software Requirements Specification.
GUI	Graphical User Interface.
HIA	Herzberg Institute of Astrophysics.
HROS	High Resolution Optical Spectrograph.
HST	Hubble Space Telescope.
HTTP	HyperText Transfer Protocol.
ICD	Interface Control Document.
IRAS	Infrared Astronomical Satellite (see glossary).
ISO	International Standards Organization.

IVO	International Virtual Observatory (see glossary).
MB	Megabytes (1048576 bytes).
MEF	Multi-Extension FITS file.
NED	NASA/IPAC Extragalactic Database (see glossary).
NGST	Next Generation Space Telescope.
NIRI	Near InfraRed Imager (see glossary).
NVO	National Virtual Observatory (see also IVO).
OCDD	Gemini Science Archive Operational Concept Definition Document (see glossary).
OT	Gemini Observing Tool.
PI	Principal Investigator (see glossary).
PSF	Point Spread Function (see glossary).
RDBMS	Relational Database Management System.
S/N	Signal to Noise ratio.
SDD	Software Design Description.
SIMBAD	Set of Identifications, Measurements and Bibliography for Astronomical Data (see glossary).
SP	Science Program (see glossary).
SQL	Structured Query Language (see glossary).
TB	Terabytes (~1.1×10^{12} Bytes).
TBD	To Be Determined.
UI	User Interface. In this document, this refers to the archive user interfaces to the GSA.
URL	Universal Resource Locator.
WCS	World Coordinate System.
WPD	Work Package Description.
WWW	World Wide Web.

4. Glossary

Altair — A natural guide star adaptive optics system being built for the Gemini north telescope.

Archive — In the GSA documents, the term archive is the set of scientific astronomical data, intended to be accessible to the general astronomical community. An archive provides the tools necessary to allow an astronomer to easily identify and retrieve those data which will contribute to a specific area of study.

Associated Data Superset — This is used to refer to the concept of using data processing to merge several inputs into a single output. For example, several datasets containing data for the same field could be averaged into a single associated data superset. Within the CADC, we tend to refer to associations as image data of the same field, however we are trying to generalize the concept to include other associations, such as mosaics of adjoining images, averaging of spectral data of a target (where the spectra cover the same band), and concatenation of spectra of a target (where the spectra do not cover the same band). An associated data superset may be a virtual concept, with only a database entry describing how members could be combined, or it may be a FITS file which contains the results of data processing. See also Data Superset.

Association — See “Associated Data Superset”.

Astronomical Server URL — A standard proposed which will allow standardized access to astronomical catalogues over the world wide web. (See http://vizier.u-strasbg.fr/doc/asu.html.)

Calibration — This is the processing of raw data to remove instrument or environmental signatures. This is usually the first processing step towards making raw data scientifically useful.

Catalogue — There are two types of catalogues which are significant to the GSA.

- An archive catalogue contains descriptor values for each scientific observation in an archive, and it is the primary tool used by astronomers when identifying data for retrieval. The descriptor values are properties of the data as a whole, e.g. WCS information, S/N for the data, primary target, etc.
- An object catalogue is a list of astronomical objects, with associated descriptors. In this type of catalogue the descriptors describe the objects. An object catalogue may be derived from the data in a single archive (e.g. an HST object catalogue is under development at the CADC), from a single project (e.g. EMSS), or they may be compiled from a variety of sources (e.g. SIMBAD, or NED). Entries in an object catalogue could have links to an archive catalogue, identifying the data from which the object descriptors were derived.

Classical Observing — This is a mode of observing where the execution of a science program is overseen by a PI or a CoI who is present at the telescope while the science program is being executed. See also Queue Observing, Remote Observing, and Service Observing.

Client — A computer program which uses a **“Server”** to carry out some action which it cannot otherwise do itself. The client commands the server and uses it as a system resource.

Co-Investigator — A person collaborating with a Principal Investigator (see glossary) on a project.

Configuration — A description of the state of one or more components of a system. For example, a Gemini instrument controller configuration might contain the location of the filter wheel, position of the grating etc. A state described by a configuration can be dynamic. For example an instrument could be continuously changing the state of the chopper in one “configuration”.

Context diagram — A special case of a “Data Flow Diagram (DFD)” in which a single bubble represents the entire system. All boxes in a context diagram represent objects in the outside world.

Cross catalogue query — A query that correlates the content of two or more catalogues. An example of a cross archive query would be “find all NIRI data from the Gemini Science Archive where there also exists at least one observation in the JCMT SCUBA catalogue with an overlapping field of view”. (See also Multi-archive query.)

Cross archive search — See “Cross catalogue query”.

Data Handling System (DHS) — That part of the Gemini Control System responsible for displaying, processing, saving and archiving data in a standard way.

Data Flow Diagram (DFD) — A diagram showing how data is handled by a software system.

Data label — A unique label which is used to identify a dataset created by the Gemini Telescopes. A data label can be used to uniquely identify the dataset for as long as the dataset exists.

Data Mining — The extraction of information or knowledge from large databases.

Data Package — The assembly of data packaged and distributed together on removable media (CDs or DVDs) as either science data, facility data, or baseline calibration data. A data package includes all relevant observational data (pixel data and meta-data) and may also include relevant PR images and information. A data package may contain data from one or more observations. A data package may require one or more volume of the removable media.

Data Processing Recipe (DPR) — A recipe describing how to process a particular type of data.

Data Quality — A single descriptor which gives an indication of the usability of a dataset. Tentatively, the image quality for data from the Gemini telescopes should categorize the image quality as follows:

- The data does not meet Gemini standards and has no scientific value. This data will not be visible to GSA users.
- The data meets Gemini standards, but does not meet the standards required for the science program.
- The data meets Gemini standards and meets the standards required for the science program.

Data superset — A data superset is either a single dataset, or an associated data superset (see "Associated Data Supserset").

Dataset — Data created from a single OBSERVE command. A dataset consists of header data describing the dataset and one or more frames. Each of the frames has a separate header containing any information specific to the individual frames. All data from a dataset is eventually stored in a single FITS file, with each frame stored in a FITS extension.

Descriptors — The data which describes a dataset. This would include data from a dataset's FITS file headers, and data from other sources, such as weather data, and proposal information.

DHS database — The DHS database contains records of datasets collected by the telescope, along with a selected set of header parameters, the history logs from the various principal systems, the data dictionary, information about archive media created by the DHS, etc.

External Requirements — These are requirements which are beyond the scope of the GSA implementation and operations, and the Gemini project as a whole. See also Functional Requirements and Usability Requirements.

File Transfer Protocol (FPT) — A widely available protocol for transferring data files over the internet.

FITS extension — A FITS extension is part of a FITS format file. FITS extensions are used to allow more than one image or table to be stored in a single FITS file. Each FITS extension has a separate header area, which contains meta data describing the data content of the extension.

Note: The content of a FITS extension could be meta-data further describing the data in the primary FITS data, or describing the data in another FITS extension.

FITS format file — A file obeying the Flexible Image Transport System (FITS) standard (see http://fits.gsfc.nasa.gov). For the GSA project, this document the term “FITS format file” will refer to a file written to disk.

FITS header — A FITS header as written to a FITS format file. This consists of a series of 80-byte ASCII card images. Each record contains a unique 8 character keyword identifying the data item, the value of the item encoded as an ASCII string, and an optional comment describing the item. The value of a data item must be a single integer, floating point or string value.

Functional Requirements — These are the requirements the GSA must meet in order to make possible and practical, all of the science described in the GSA OCDD. See also Usability Requirements and External Requirements.

Gemini Engineering Archive (GEA) — The purpose of the GEA is to collect data from all of the Gemini systems and to allow efficient retrieval of the data for the purpose of analysing the system behaviour. To avoid duplication, the GEA may contain pointers to data in existing databases such as the DHS permanent store and Observing database. The DHS permanent store contains all image data, image processing histories and recipes, and completed science programs. The Observing database contains planned, active and completed science program information.

Gemini Science Archive (GSA) — An archive system intended to preserve scientific data collected by the Gemini Telescopes. After the proprietary period for the data has expired, the data is made available to the general astronomical community.

Gemini Science Archive Functional and Performance Requirements Document (FPRD) —

This document describes the functional and performance requirements to be included in the Science Archive, and describes functional and performance requirements required from other Gemini Systems.

Gemini Science Archive Operational Concept Definition Document (OCDD) — This document describes the science cases for which the Science Archive is designed, contains the scenarios which show how a GSA user would interact with the GSA in order to collect the data required for each of the science cases, and lists the user requirements which arise from the scenarios.

Gemini Science Archive User — For the purposes of these documents, GSA users are considered to be professional astronomers, or students in a university level astronomy program. The GSA is not intended to provide data to the general public.

Hipparcos — Hipparcos is a pioneering space experiment dedicated to the precise measurement of the positions, parallaxes and proper motions of the stars. The intended goal was to measure the five astrometric parameters of some 120 000 primary program stars to a precision of some 2 to 4 milliseconds of arc.

Infrared Astronomical Satellite (IRAS) — An infrared space telescope. It conducted an all-sky survey at wavelengths ranging from 8 to 120 microns in four broad-band photometric channels centred at 12, 25, 60, and 100 microns. Some 250,000 point sources were detected. A catalogue of small (< 8’) extended sources gives the characteristics of some 20,000 objects. An atlas of images covering the entire sky gives the absolute surface brightness at each of the four survey wavelengths.

Instrument — A device, consisting of a detector and a means of dispersing and/or focusing light onto that detector, designed to extract a particular type of information from the radiation gathered by the Gemini telescope. An instrument is specially designed for astronomy research. The wavefront sensors on the Gemini telescope can be regarded as an instrument.

Instrument Control System (ICS) — That part of the Gemini Control System which is responsible for controlling and operating a scientific instrument. There is a separate Instrument Control System for each available scientific instrument.

Interactive Observing — A mode of observing where details of the sequence of events are decided in real-time while performing the observation. See also Planned observing.

Interface Control Document (ICD) — A document which either describes a general interface between Gemini systems, or which describes in detail the interface between two specific Gemini systems.

International Virtual Observatory (IVO) — Also known as the National Virtual Observatory (NVO). This is a proposed virtual observatory which will incorporate catalogues and archive data from a variety of surveys and observatories.

ISO 9660 — The international standard defining the CD-ROM data format.

Logging — The act of keeping a record of happenings during an observing session. “Logging” encompasses the observing log, system history records and any engineering logs made for the day crew.

Meta-data — Meta-data is data which describes data. This can be a recursive concept in that the header of a FITS file contains keywords and values which are Meta-data for an image stored in the file, but there could also be Meta-data describing the FITS keywords and values (permitted range or values, units, description, etc.).

Multi-catalogue query — A query which spans more that one catalogue, but does not directly correlate the content of the two catalogues. An example of a multi-catalogue query would be “find all images of a specified object in the Gemini catalogue, in the HST catalogue, or in the CFHT catalogue. (See also cross archive query.)

Multi-archive search — See “Multi catalogue query”.

NED — NED is a catalogue built around a master list of extragalactic objects for which cross-identifications of names have been established, accurate positions and redshifts entered to the extent possible, and some basic data collected. Bibliographic references relevant to individual objects have been compiled, and abstracts of extragalactic interest are kept on-line. Detailed and referenced photometry, position, and redshift data, have been taken from large compilations and from the literature. NED’s data and references are being continually updated, with revised versions being put on-line every 2-3 months.

NICMOS — An HST instrument for both imaging and multi-object spectroscopy.

NIRI — A Gemini north instrument designed for near infrared imaging and spectroscopy.

Non-conforming instrument — An “Instrument” which does not conform to the Gemini hardware or software requirements. Such an instrument is only allowed as a **“Visiting instrument”.**

Observation — An observation is the smallest scheduable event for the Gemini telescopes. A Science Program may consist of multiple observations, and an observation may consist of several datasets, each with a different instrument and/or telescope configuration.

Observing log — The observing log is a list of all observations taken and key attributes of the observations, including time of observation, observer comments, object of observation, etc.

Observing Modes — As defined in the Gemini Software Design Description, p5-46, the observing modes are: interactive, queue-based, service and remote.

Observing Tool — The primary astronomer interface of the Gemini control system.

Planned Observing — Planned observing is a science program where the details of the observing process are planned in advance. Planned observing is required for both Queue Observing and Service Observing, and is strongly encouraged for classical users. See also Interactive Observing.

Point Spread Function — A measure of image quality.

Preview data — A representation of a dataset intended for display to an archive user. The intention of preview data is to allow the archive user to quickly view an image or spectral plot, giving an indication of the data quality and usefulness. Preview data is a calibrated, compressed (possibly with lossy compression) version of the data.

Principal Investigator (PI) — The Gemini user identified as being the person who has sole access to the science archive data associated with his/her science program for the proprietary period.

Proprietary Period — The period in which data in the GSA is not released to the general public. From the point of view of the GSA operations, the initial public release date for each dataset is set by Gemini, and may be changed by Gemini at any time.

Prototype — An implementation of the functionality of some aspect of the GSA prior to the delivery of the fully finished system. The purpose of the prototype is to clarify an aspect of the design.

Queue Observing — In this mode of observing, an astronomer submits a science program to an observatory, and the science program is added to a list of science programs waiting to be executed. The execution of the science program is overseen by observatory staff. The scheduling of the program execution can by based on many factors, including time in the list, current observing conditions, and the conditions required by the science program. Queue mode observing also requires that the science program consists of planned observations. See also Classical Observing, Remote Observing, and Service Observing.

Recipe — See Data Processing Recipe.

Remote Observing — This is similar to Classical observing, except that the Observer oversees the execution of the science program from a remote site (usually their home institution). See also Classical Observing, Queue Observing, and Service Observing.

Rock Ridge — An extension to ISO 9660 to allow Unix-specific information such as links, file attributes and native Unix naming to be stored along with the ISO 9660 information.

ROSAT — ROSAT is an X-ray space telescope which went into operation in June 1990. As the primary objective ROSAT has performed in a scan mode the first all-sky survey with imaging telescopes in the soft X-ray band of 0.1 keV - 2 keV (corresponding to wavelengths of 100 Å - 6 Å) as well as in the adjacent extreme ultraviolet region of 0.04 keV - 0.2 keV (corresponding to wavelengths of 300 Å- 60 Å). More than 60,000 X-ray sources have been detected with the ROSAT all-sky survey, larger by almost two orders of magnitude than the 840 sources of the catalogue of the previously largest all-sky survey of the HEAO-I satellite.

Science program — A formal description of an observer's plan for using a Gemini Telescope. Science programs are produced by the Gemini Observing Tool. A science program is suitable for near-automatic execution. It usually consists of an unordered group of Observations.

Server — Any computer program which carries out some form of service for one or more client computer programs. Typically, a server will be on a remote machine, and therefore it has direct access to resources not directly available to client programs. A server is completely subservient to its clients. It responds to commands from the clients but does not initiate dialogue with clients.

Service Observing — This is similar to queue observing, in the sense that the execution of a science program is overseen by a member of the observatory staff, instead of one of the a members on the science team who submitted the science program. Service observing also requires that the science program consists of planned observations. See also Classical Observing, Queue Observing, and Remote Observing.

SIMBAD — The SIMBAD astronomical database provides basic data, cross-identifications and bibliography for astronomical objects outside the solar system. SIMBAD can be queried by object name, coordinates and other criteria (filters).

Status — Attributes which define the state of a hardware or software system.

STIS — An HST instrument for both imaging and spectroscopy.

Structured Query Language (SQL) — A standard language, common in commercial database management systems, used for specifying how to search a relational database for items matching certain criteria. The criteria can involve arithmetic and logical combinations of record properties (tuples).

Telescope — In the Gemini software documentation the word "Telescope" is used to refer to the Gemini 8m telescope together with all of its associated peripheral devices. These include the telescope mount, the primary and secondary mirrors, together with their support and control structures, the Cassegrain rotator, and all the devices associated with acquisition and guiding and with the adaptive optics. Some of the latter devices can also be regarded as instruments.

Transportable media — Media that will contain the results of an observation program for transport to the observers home facility, and to the archive centre. This is base-lined to be either DAT or EXABYTE tapes, or CD-ROM.

Universal Resource Locator (URL) — The unique address of an internet resource, such as a hypertext document, gopher page, mail address or a computer. The address adheres to a communications protocol described in RFC1738 (http://www.cis.ohio-state.edu/htbin/rfc/rfc1738.html).

Unix — A standard host-level operating system. The Gemini baseline for "Unix" is "Solaris" marketed by Sun Microsystems.

Usability Requirement — This is a requirement which does not enable any new functionality, but which make the GSA easier to use.

Visiting instrument — A self-contained **"Instrument"** which is brought to the Gemini telescope for temporary use. This type of instrument is intended only for the use of the instrument owner.

WFPC2 — An HST imaging instrument.

Work package — A well-defined unit of work, described by a workscope document. A work package usually consists of the development of a subsystem or a layer of utilities, which can be devolved to a collaborating group.

World Coordinate System (WCS) — A mapping of the pixel column/row (x/y) coordinate system of an image into a coordinate system that has meaning in the real world. In astronomy, the would coordinate system is most often right ascension and declination, but could also include other coordinate systems, such as galactic coordinates. A WCS could also be associated with a spectral dataset, where one of the world coordinates would be wavelength.

World Wide Web — A part of the internet designed to allow easy navigation of the network of computers through the use of GUI's and hypertext links between different URL's or addresses.

5. References

[1] *Engineering Archive Requirements*, Phil Puxley, Dayle Kotturi, Jim Oschmann, Bret Goodrich, Gemini, Daniel Durand, David Schade, David Bohlender, Séverin Gaudet, HIA, http://www.gemini.edu/documentation/webdocs/spe/spe-c-g0073v02.pdf

[2] *Gemini Archive Position Paper*, Andy Woodsworth, January 3, 1992, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/References/cadcPosition.pdf

[3] *The Gemini Science Archive: A Proposal by the Canadian Astronomy Data Centre*, Dennis R. Crabtree, September, 1995, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/References/proposal.pdf

[4] *Gemini Science Archive Conceptual Design Study Work Scope* No. 9414257-GEM02012, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/Workscope/workscope.pdf

[5] *Gemini Science Requirements*, SPE-PS-G001, Version 2., http://www.gemini.edu/documentation/webdocs/spe/spe-ps-g0001.pdf

[6] *Gemini Software Design Description*, SPE-C-G0037, Gemini 8m Telescopes Project, http://www.gemini.edu/documentation/webdocs/spe/spe-c-g0037.pdf

[7] *GSA Functional and Performance Requirements Document*, gsa_FPRD/01, Norman Hill, Séverin Gaudet, Daniel Durand, David Schade, David Bohlender, NRC, Felipe Barrientos, Felipe Richardson CONICYT, in preparation

[8] *IEEE Guide to Software Requirements Specification*, IEEE Std. 830 - 1984

[9] *Recommendations of Science Archive Workshop*, http://www.gemini.edu/sciops/data/scia-arch-rec.html

[10] *The Scientific Case For a Gemini Data Archive*, David Schade, Daniel Durand, Jean-Rene Roy, Maria-Teresa Ruiz, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/References/science_case.pdf

[11] *Summary of Science Archive Workshop Presentations*, http://www.gemini.edu/sciops/data/scia-arch-wkshop.html

Created:February 7, 2000
Modified:June 5, 2000

Gemini Science Archive

Chapter 2 General Description

1. History

The Gemini Science Archive was originally proposed in a position paper [2] written by the CADC in early 1992. This paper described the value of a Gemini Archive, and described how it could be implemented. This was followed in September 1995 by a proposal [3] by the CADC to design, implement, and operate a Gemini Science Archive based on the existing archive infrastructure maintained at the CADC.

At its April 1998 meeting the Gemini Science Committee (GSC) reviewed the science cases for a Gemini archive presented in [10], and endorsed the following resolution concerning a Science Archive for Gemini:

> *"The GSC thinks a Gemini science archive is scientifically compelling. Such an archive would provide the scientific community with tools for effective on-line access to all Gemini science data and supporting information in order to promote further scientific exploitation of those data. The Gemini Science Archive should guarantee that the valuable datasets obtained with the Gemini Telescopes are usable by future generations for research and education. The Project should further study the implementation and the resources required for a Gemini science archive."*

The Gemini Board has endorsed the GSC views on a Gemini Science Archive, and agreed that the Project should study the specifications and requirements for a science archive

In order to provide a basis for the science archive conceptual design activities, Gemini sponsored a workshop to assess the feasibility of a viable Gemini Science Archive and recommend prioritized specifications and requirements for the science archive. The Science Archive Workshop was held on Monday 14 September 1998, in the Gemini Hilo Base Facility, Hilo, Hawaii. The workshop recommendations can be found in [9]. A summary of the workshop presentations can be found in [11].

The workshop recommendations described in [9], formed the basis of the Gemini Science Archive Conceptual Design Study work scope [4], and this document has been produced as part of that work scope.

2. Relation to Current Projects

2.1 CADC Archive Development

The CADC currently supports archives for several telescopes, including the HST, the CFHT, JCMT, and several others. It is expected that the development of the GSA by the CADC will have benefits to these other archives, and that development efforts for these other archives will in turn benefit the GSA. In particular, the GSA may benefit from the current development efforts at the CADC, aimed at creating an astronomical data mining system.

The other archives at the CADC will also be an advantage in attempting to initiate the unified catalogue required for multi-archive and cross archive searches. The CADC can represent the interests of all of the archives supported by the CADC, including the GSA, which we believe will increase the chances of creating a successful unified archive. It is also possible to create a very useful unified archive consisting only of the archives supported by the CADC.

2.2 The Gemini On-line Data Processing System

The Gemini Telescopes currently plan to implement a On-Line Data Processing system (OLDP) to be used for near real-time evaluation of data quality during telescope operations. We hope that the GSA will benefit from the existence of this data processing system by re-using the data processing recipes, and possibly the data processing infrastructure that must be developed to support the OLDP. It is also feasible that the implementation of the OLDP could be done as a shared task with the GSA development.

3. Relation to Predecessor and Successor Projects

3.1 The CADC Archive Infrastructure

The GSA will build on the software, hardware, and staffing infrastructure created by the CADC for existing archives. The experience gained, lessons learned and software developed during the development of these archives will be applied to the development of the GSA.

3.2 The Gemini Engineering Archive

The Gemini Telescopes have currently implemented a Gemini Engineering Archive (GEA), which records much of the data required by the GSA. The existence of the Engineering archive will simplify the task of gathering data required for the GSA.

3.3 The Gemini Data Handling System

As part of a separate work package, the CADC developed the Data Handling System (DHS) for the Gemini telescopes. As part of the development of the DHS, the CADC developed an intimate knowledge of the data handling infrastructure of the Gemini telescopes, and considerable knowledge of other elements of the Gemini telescopes. This knowledge will be extremely valuable in the design and implementation of the GSA.

3.4 Ongoing development of the GSA

After the initial design and implementation of the GSA is complete, an ongoing development process should be instituted as a part of the GSA operations. This ongoing development will allow advanced capabilities to be added to the GSA. These advanced capabilities may be capabilities documented in [7] which cannot be implemented within the initial GSA development budget, or they may be capabilities that are not yet identified.

4. Function and Purpose

The Gemini Science Archive should provide the scientific community of the partner countries with on-line access to all Gemini scientific observations to ensure full scientific exploitation of those data. The Gemini Science Archive should be considered an integrated part of the planning, observation, calibration, data reduction, and data distribution processes that occur at Gemini. The Gemini Science Archive should guarantee that the valuable datasets obtained with the Gemini Telescopes are preserved for use by future generations for research and education.

A well-designed and properly implemented Science Data Archive, as part of the capabilities of the Gemini observatories, would be a major contribution toward the full exploitation of the unique characteristics of the Gemini telescopes. An effective archive would boost scientific productivity and would ensure that maximum value was extracted from the expensive-to-obtain observational data. Archiving requirements are consistent with, and would help to optimise, queue mode observing requirements. In the long term, a science archive would enable many of the high level science goals outlined in the Gemini Science Requirements document [5]. Some specific goals— for example, understanding the relationship between quasars and their host galaxies, probing variable phenomena in stars and galaxies, dynamics of galactic nuclei, would be unattainable without archival access to the complete set of Gemini observations. The collective impact of the full Gemini Science Archive would far exceed what could be produced by the programs of single observers or teams of observers. It would span a wider range in physically important parameters such as redshift and luminosity of active galactic nuclei or galaxies. Furthermore, many high-priority Gemini projects would use the Gemini data in conjunction with Hubble Space Telescope, HIPPARCOS, ROSAT, IRAS, NGST and other archival data.

Access to Gemini data by the wider community, after the proprietary period, would ensure that the data are fully exploited and that maximum scientific value is returned to the astronomical communities and, ultimately, the citizens that support Gemini. The true legacy of the Gemini observatories, in addition to the initial discoveries would be the collection of excellent observations produced by its innovative instruments and this legacy, in the form of the **first scientifically effective archive of ground-based observations,** would play an important role in astrophysical research long after Gemini itself ceases operations.

5. Mission Statement

Following is the mission statement for the GSA:

The Gemini Science Archive should provide the scientific community with tools for effective on-line access to all Gemini science data and supporting information in order to promote further scientific exploitation of those data. The Gemini Science Archive should guarantee that the valuable datasets obtained with the Gemini Telescopes are usable by future generations for research and education.

6. User Characteristics

The users of the GSA are primarily, but not limited to, astronomers from the Gemini member countries. It is expected that the GSA users will use the archive for the following general purposes:

- Applying new techniques to re-analyse data obtained by the original observer.
- Creating sets of data larger than could be practically created with PI observations.
- Verification of the findings of the original observer.
- Augmenting original observations collected with either the Gemini Telescopes, or with other telescopes.

- Evaluation of the performance and capabilities of Gemini Telescopes and instruments by potential users.
- Planning for future observations of the same or similar fields of view.
- Creating new datasets by spanning observing runs.
- Examining historical behaviour of time varying phenomena.
- Proposal preparation for PI observations at Gemini and other telescopes.
- As a teaching or learning tool.

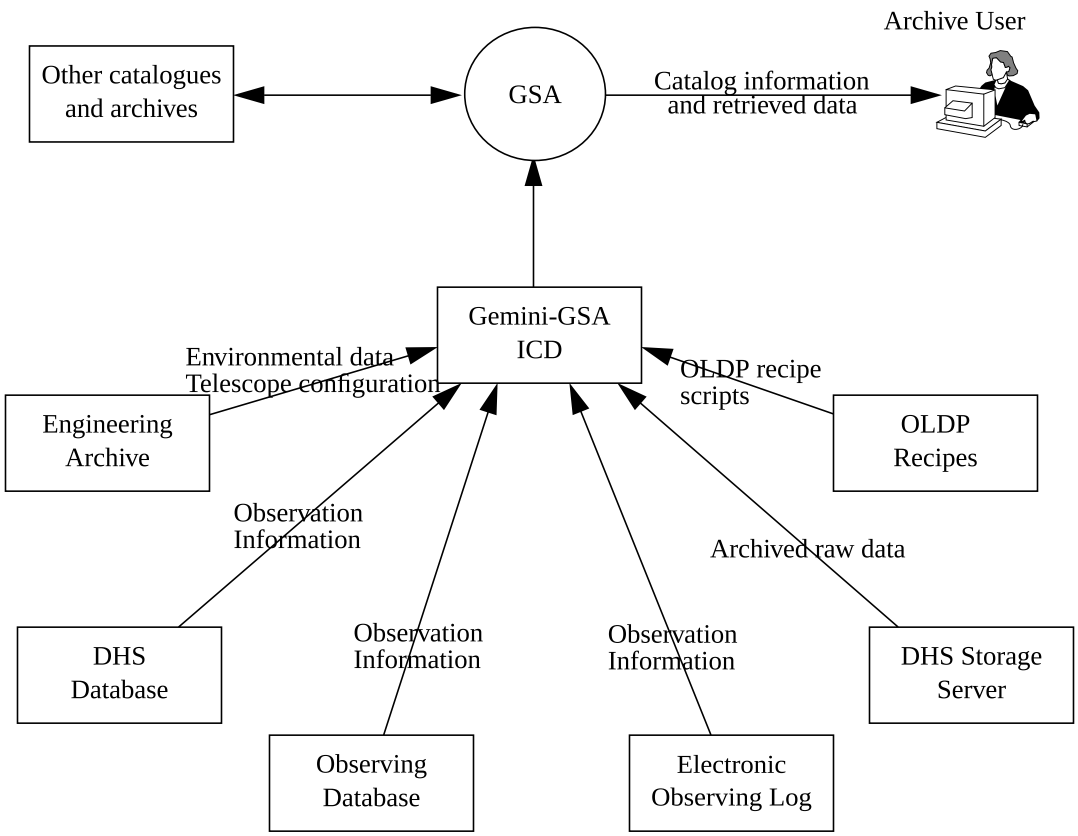
7. Relation to Other Systems

The GSA uses data from a variety of other Gemini systems as input, and will therefore need interfaces to these systems. These systems include:

- The Gemini DHS Storage Server will produce removable media containing the raw datasets which will be used by the GSA.
- The Gemini Engineering Archive will record telescope activity and environmental data which will become meta-data for the raw datasets, and which will help identify calibration data associated with the raw data.
- The Gemini DHS database will provide an independent list of raw datasets which will be used to verify that all expected data is received on removable media.
- The Gemini Observing Database will provide information about the structure of the observations, the data processing done, and the public information from the original proposal.
- The Gemini Electronic observing log will describe the actions of the observer during the observing run, which will become meta-data for the raw data.
- Data processing recipes and rules for identifying calibration data, used by Gemini for data quality analysis will be the basis for the GSA automatic data processing.

The interface to each of these data sources will be described in an ICD. The context diagram for the GSA is shown in Figure 1 on page 15.

FIGURE 1. GSA Context Diagram



Gemini Science Archive

Chapter 3

Science Cases[1]

1. Introduction

Space-based observatories have produced scientifically effective archives for over two decades. Data from IUE, IRAS, Einstein and other missions have made clearly important contributions to progress in astronomy. Hubble Space Telescope (HST) has broken new ground in the development of archives of optical data. The observations are all saved, pipeline processed and calibrated, catalogued and distributed. The HST archive has only recently begun to be heavily exploited and will be a valuable resource for decades to come. Hanisch (1998, SPIE, vol. 3349) recently reported that the data retrieval rate from the HST archive is now higher than the rate at which new data are being ingested; he also pointed out that, up to present, ten times more International Ultraviolet Explorer (IUE) data have been extracted from the IUE archive than were originally put in it.

Some large astronomical projects like the Palomar Sky Surveys, the Sloan Digital Sky Survey, or the 2Mass Survey are themselves archive projects. For example, The Sloan Digital Sky Survey (SDSS) — using a telescope and instruments dedicated to that project— will contain photometric, spectroscopic, and morphological parameters for several hundred million objects. Archiving is taking its place as one the most important resources that serve the astronomical research community.

It is important to appreciate the difference between a safe store for observatory data and a useful science archive. The science archive requires careful cataloguing and effective search and retrieval tools as well as the capability of reliable calibration; see for example the AstroBrowse Web site at `sol.stsci.edu/~hanisch/astrobrowse_links.html`. These are the features which allow an archive to produce science and they are absent from basic data storage systems.

1.1 Archive Research Opportunities

There are at least three classes of archive research project. The first consists of cases where the data are used for an entirely different scientific project than for which they were obtained. The second is the case where new, improved, or otherwise different and more effective methods of analysis are brought to bear on the data. The third, and perhaps most important class exploits the collective effect of the archive where a larger and more comprehensive dataset (consisting of all of the archive observations taken to date) spanning a wider range in some important parameter is available to the archive researcher than could ever be available to an individual proposer. **The whole of the archive dataset is worth far more than the sum of the parts**, and the linkages across archives and across wavelength regimes adds still more value to archive data.

1. Most of this chapter is extracted from [10], but new science cases have been added which illustrate the uses of the GSA in a wider cross section of astronomy. The description of the GSA originally included in [10] has been incorporated into Chapter 2, and the requirements originally included in [10] have been incorporated into Chapter 5.

An excellent illustration of the effective use of archive resources is “The Demography of Massive Dark Objects in Galaxy Centres” (Magorrian et. al. 1997, astro-ph/9708072). Nearly all of the leading workers in this field have collaborated to produce a study which uses imaging data from at least 6 HST programs and incorporates kinematic information from more than 10 separate ground-based observational programs. Clearly, this approach of combining many years worth of observational effort into a large homogeneous dataset is extremely effective. The existence of good archive facilities makes this type of substantial scientific progress possible.

As a second illustration, the CFHT archive was searched for observations of NGC 1068, an AGN which displays spectroscopic and photometric variability. The search took only a few minutes of effort and returned 189 exposures from 8 separate programs spanning 7 years. Spectra and images in the optical were obtained in 6 programs and infrared observations were made in 2 additional programs. The long time baseline makes this a very valuable archival dataset. A search of the JCMT archive revealed 613 observations of this object in the sub-millimetre regime from numerous programs. These could be combined with the 283 HST observations taken over a period of 6 years with 6 different instruments in 22 separate programs. All of these data are available from a single archive site at CADC. Over 50 observations are available from 6 different X-ray and gamma ray missions through the HEASARC archive. This is a very well-observed object but many other sources have been observed at multiple wavelengths at multiple observatories over long time baselines and archiving preserves the value of these data for future research.

The range of published archival research is impressive. Koesterke et al. (1998 A&A 330,1041) combine HST and IUE archival spectra to study mass loss in four PG1159 stars, Cagnoni et al. (1998 ApJ 494,54) use archival ASCA observations to evaluate the contribution to the X-ray background of discrete sources in the 2-10 keV energy band. Sodemann & Thomsen (1998 A&A Suppl. 127,327) perform an extension and re-analysis of earlier crowded-field photometry in M32 from archival HST imaging. Archival data from the Burst and Transient Source Experiment (BATSE) is reprocessed by Kommers et al. (1997 ApJ 491,704) to achieve higher sensitivity and is searched for low-significance transient signals in the 50-300 keV range. Ciliegi & Maccacaro (1997 MNRAS 292,338) study time and spectral variability of Einstein Extended Medium Sensitivity Survey (EMSS) active galactic nuclei (the EMSS itself was an archival research project based on the Einstein archive). Serendipitous asteroid trails are detected by Evans et al. (1997 AAS 29.0701) by examining 30,568 frames from the HST Wide-Field Camera and 96 moving objects are found. Rigopoulou, Lawrence, & Rowan-Robinson (1996 MNRAS 278,1049) combine archival and other data from the sub-millimetre (JMCT) to the X-Ray (ROSAT) for a set of ultra-luminous IRAS galaxies. IRAS archival data is used by Noriega-Crespo et al. (1997 AJ 113,780) to produce a survey for bow shock structures around OB runaway stars, and follow-up work uses re-processed IRAS high-resolution maps.

It is evident from these examples—all involving space-based instrumentation in a key role— that extremely valuable science over a wide range of subject areas can be done with a properly-implemented Science Data Archive and that much of this science would not be possible in its absence.

1.2 Ground-Based Astronomy Archives

There is no fundamental reason why a Science Data Archive from a ground-based facility should be more difficult to implement or less valuable than an archive of data from a space-based observatory. There have historically been differences in facility design motivated by the fact that a higher level of planning and automation is required in space where real-time human intervention in observing procedures is much more difficult than on ground-based telescope. **The Gemini observatories are being designed to operate in a mode that very much parallels that of space-based observatories and these design requirements will allow an effective archiving**

to be created. The motivation behind Gemini's design decisions is not the impossibility of human control (although the effect of Mauna Kea's altitude argues for a minimum of real-time decision-making), but the desire for optimum observatory performance which requires detailed planning of observations and queue-mode observing.

What are the unique difficulties of ground-based observing and archiving? The salient feature is varying weather conditions. The solution is to monitor and log transparency and seeing and maintain links between this information and the data. Queue-mode observing allows the observatory to respond to changing conditions by executing programs that are best-suited to those conditions thus optimising the scientific productivity of the Gemini facilities.

A number of other challenges faced by existing archives of ground-based observations are related to instrument and facility design as well as deficiencies in observing and logging procedures. Extensive experience with the archive of the Canada-France-Hawaii Telescope (CFHT) has brought these problems into sharp focus and has showed how to solve them. There have been four basic deficiencies. First, at CFHT there is no guarantee that adequate calibration material is obtained. Second, there is no requirement for adequate or uniform logging of observations and weather conditions. Third, there is no guarantee that data headers include **ALL** of the instrument, telescope, and other system configuration information that is essential to understand and reliably calibrate the data; furthermore there is no guarantee that some key components (e.g. filters) are in the right place, because they are not all encoded and monitored. These are the reasons that the CFHT archive—currently the best archive of ground-based data in existence—has not realized its full potential scientific productivity. Archive users, in general, simply lack sufficient information and confidence about what occurred during the execution of the observations to produce reliable science-quality data from the archive.

Gemini has been designed along the lines of a space-based observatory and this guarantees that most of the problems cited in the case of the CFHT archive are automatically resolved. An archive has been envisioned as an integral part of the Gemini facility. The success of the archive requires, above all, this element of integration with day-to-day operations of the observatory scientists and engineers, effective interactions with the instrument teams, and the contributions of the user community. In the ways it has designed its telescopes and instruments, the manner it has constructed its engineering archive and the plans it is setting for its operational modes, Gemini has already laid the foundations that are needed for the effective operation of a Science Data Archive.

1.3 Motivation for a Gemini Science Archive

The main argument in favour of allocating resources to a Science Data Archive is that it would increase the quality and the quantity of the science that is produced by the Gemini observatories. It would also preserve the scientific value of the Gemini data far into the future, not as part of an historical record but, rather, as data that would continue to contribute actively to scientific progress for decades to come.

A successful Gemini Science Data Archive would ensure that all scientists in all countries of the partnership would have access (following an appropriate proprietary period) to all of the data produced by the observatory. This would represent important added value to the Gemini partnership. The archive would ensure that the Gemini observations are fully exploited and that opportunities for doing Gemini research would be as widely distributed as possible. An archive would also be extremely valuable both for educational purposes and for public outreach activities (where HST has excelled). Furthermore, there is an issue of public accountability. The allocation of resources to an archive would demonstrate to the taxpaying public that all efforts were being taken to ensure that maximum value was being extracted from the Gemini facility and that these astronomical data were highly-valued and needed to be protected to preserve future research opportunities.

2. Science with the Gemini Archive

Gemini's first and subsequent generations of instrumentation will provide unprecedented observational capabilities and will open up new opportunities for study in fields as diverse as planetary searches and high-redshift clusters of galaxies. The uniqueness of Gemini science makes the provision of an archive of these data a compelling priority.

In the following sections, we discuss a representative (but very incomplete) subset of science cases which show how a Science Archive would help to realize several specific scientific goals of the Gemini observatories. In some cases the main benefit is derived from the larger and more comprehensive database represented by the Gemini observations accumulated over a period of time. Sometimes the increased time resolution is important, for example, for the study of variable phenomenon and proper motions. In some cases, Gemini observations will be combined with those from other facilities to provide a wider baseline in time or in wavelength. In some examples observations will be used for a completely different scientific purpose from that for which they were obtained and sometimes they will be used for the same purpose, but the new results will be due to fresh viewpoints and more effective methods of analysis used by archival researchers. Gemini archival observations will be used as the basis for new proposals to Gemini and other facilities, and also will be employed in conjunction with data from sister archives. In all of these cases, the Gemini archive would help ensure that the best is made of the available information content.

Each science case is labelled with a unique identifier (e.g. SC 1.1) which will be used when referencing specific science cases in the later chapters of this document, and from other documents.

2.1 Star Formation

Star formation occurs in molecular clouds in regions with large amounts of extinction and the spatial scales involved are small (1-10^4 AU which translates to sub arc second even for nearby regions). High spatial resolution and infrared capabilities are most important. Imaging and moderate to high spectral resolution (requiring good light gathering power) are needed to answer questions about environmental effects on the initial mass function, the physical state of molecular clouds and grain chemistry. Accretion and outflow processes around Young Stellar Objects can be investigated, and the basic parameters of circumstellar disks in nearby regions will be measurable using a combination of adaptive optics, coronagraphy, and polarimetry.

The Gemini archive could be useful in several aspects of the study of star formation. For example:

SC 1.1 Detection of newly born stars

In 1986 Reipurth & Bally (Nature 320, 236) detected the sudden appearance of a highly variable nebula (Object 50) in the southern part of L1641 cloud in Orion. The nebula, which is associated with a bipolar molecular flow and an IRAS source, was not present on the Palomar Schmidt plates taken in 1955. During the period, the nebula experienced dramatic variations, including the appearance of a jet containing several emission knots.

The authors believe they had witnessed the first light emerging from newly born star. The superb IR capabilities of Gemini will often lead astronomers to point at star forming regions, and any phenomena like the one described above should be readily detected. The use of archival data of the same regions will be crucial in the interpretation of the observations leading to a better understanding of the stellar formation process.

SC 1.2 Birth of Herbig-Haro systems

Using VLA archive data, Rodríguez et al. (1989; ApJ, 346, L85) found that the outer components of the triple source in Serpens FIRS1 are moving away from the central object. The Gemini archive could be helpful in identifying and determining the kinematics of outflow phenomena at even earliest stages after its onset and certainly of the kinematics of Herbig-Haro objects at later stages of evolution.

SC 1.3 Motions of jets and knots

The archive data could allow investigation of the wiggling motions (changes in flow direction) of the highly collimated jets associated with recently formed stars.

SC 1.4 Variability of pre-main sequence objects

Pre-main sequence stars are known to be highly variable objects, particularly at X-ray wavelengths. Of course the Gemini archive could be useful in assessing their variability at IR wavelengths, particularly in the H_2 luminosity from the surroundings thought to be associated with the X-ray luminosity.

SC 1.5 The fate of very massive stars

Massive stars evolve through the Wolf-Rayet phase to explode as supernovae. It is believed that the most massive ($M \geq 40\ M_{\odot}$) go through a relatively short ($\leq 10^5$ yr.) but active phase of instability after leaving the main sequence. Stars in this stage—known as luminous blue variables (LBVs)—are very rare (Figure 1 on page 22); the most luminous infrared source in the Galaxy, η Carina, for example erupted between 1837 and 1860. Such events are probably more frequent in starbursting galaxies, where many very massive stars are formed in super stellar clusters. Because of insufficient image resolution and because these events are most conspicuous in the infrared, very few LBV events have been studied. With Gemini's excellent imaging capability in the infrared, an archive would easily allow the search for, and subsequent study of, erupting LBV stars in massive star forming regions within $D \leq 10$ Mpc.

2.2 The Interstellar Medium

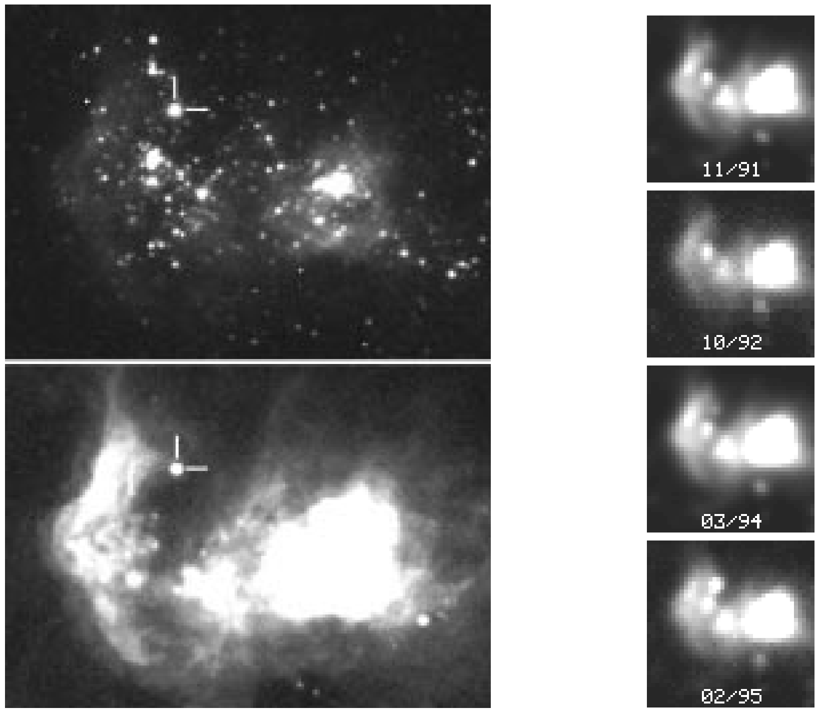
Gemini South's HROS (and a possible pier-mounted high-resolution spectroscopy facility) will provide member country astronomers with the ability to reach unprecedented high S/N, high resolution spectroscopy of many classes of objects. Many of these will involve investigations of small-scale variability in O and B stars such as nonradial pulsations, variability due to surface inhomogeneities, shock phenomena in the dense, radiatively-driven winds of Wolf-Rayet and O stars, or measurements of photospheric velocity fields.

The large spectral coverage of HROS provides a multiplexing advantage for the above studies but also provides a plethora of "free" information on the local interstellar medium (ISM).

SC 2.1 Diffuse interstellar bands

Recent catalogues now list almost 300 diffuse interstellar bands (DIBs), with most of these occurring between 5500 and 8000A. In other words, a typical spectrum of a moderately reddened star will contain a DIB every 10A. Some of the sharper DIBs display possible molecular-band

FIGURE 1. A new LBV in the giant HII region NGC 2363 located at the SW of the magellanic irregular NGC 2366 (D = 3.7 Mpc). a) HST WFPC2 B(F439W) and Hα (F656N) images of NGC 2363 in 1996 January 8. The Luminous Blue Variable star is identified. The field of view is 13” × 9”. b) Time sequence of ‘archival’ R-Band ground-based images of NGC 2363. The LBV appeared in early 1994, and it is now the most luminous object in the galaxy NGC 2366. (Drissen, L., Roy, J.-R., & Robert, C. 1998, ApJ, 474, L35)



sub-structure, and it is now generally accepted that the DIBs are caused by complex, carbon-based molecules in the gaseous phase. There is also increasing evidence that these may include carbon-chain, cyclic and large spherical molecules, with sizes extending all the way to complexes large enough to possibly account for the general interstellar extinction. Understanding the nature of the DIBs is obviously of fundamental importance to many areas of astronomy. However, in spite of intensive efforts there has yet to be a convincing identification of a specific molecular carrier for any single DIB since their discovery in the 1920’s.

While there will no doubt be Gemini programs geared specifically towards observations of these DIBs on a few sight-lines towards bright, standard stars, an analysis of Gemini archival data for all early-type stars will provide a dataset much richer than a few individual observing programs could hope to accumulate. Observations of DIBs towards all the OB stars in the archive could be used to search for statistical trends of individual DIBs and to examine the suggestion that there exist families of DIBs which may be generated by a single carrier. The different sight lines will also sample many diffuse clouds with a wide range of physical properties such as density and temperature. Correlations of these properties with the strengths of the DIBs will provide fundamental constraints on the nature of the DIBs carriers.

SC 2.2 Chemical evolution of the Galaxy

Archival data obtained with resolutions greater than 100,000 (e.g., the pier-mounted spectrograph) could also be used to measure the strengths of known molecular features such as CH, CH^+, C_2, C_3 in many different interstellar environments. CH^+, which does not suffer from chemical separation that other molecules like CO are susceptible to, is of particular interest since it is the perhaps the best molecule with which to measure the $^{12}C/^{13}C$ isotope ratio in the local ISM. This is a fundamental measurement of the chemical evolution of the Galaxy.

2.3 Magellanic Clouds Research

There seems little doubt that Gemini South will frequently be directed towards targets in the Magellanic Clouds. High-resolution, high-S/N spectra of Magellanic Cloud, Stream and Bridge O, B, and A stars obtained with HROS will provide unprecedented accuracy in the measurements of the abundances and mass loss rates of stars in environments considerably different than that of the Milky Way. These spectra in turn will provide critical constraints on the importance of metallicity on such fundamental astrophysical phenomena as radiatively driven winds, radiative diffusion, and convection.

SC 3.1 Interstellar medium in the Magellanic Clouds

As for observations of similar objects in the Milky Way, archival HROS spectra of massive Magellanic Cloud stars will also provide a wealth of additional data on the gaseous component of the interstellar medium of the Magellanic Clouds. Measurements of the interstellar features (both atomic and molecular, the latter in the form of the diffuse interstellar bands, DIBs) in these stars will enable archive researchers to produce detailed maps of the atomic and molecular column depths in the Magellanic Clouds. Somewhat surprisingly, observations to date suggest that depletion patterns in the ISM of the Magellanic Clouds and the Galaxy are quite similar, despite significant differences in overall metallicities, dust-to-gas ratios, and UV extinction curves in the three galaxies. A more detailed investigation of these correlations should therefore provide information on the physics of grain formation, as well as on the conditions necessary for the formation of the DIBs and their presumably molecular carriers.

2.4 AGN Unification and Quasar Studies

One of the more thorny astrophysical problems is AGN unification. By this we mean deciding which classes of AGN objects are in fact identical, but simply viewed from different orientations, and also which classes differ in only minor respects and should be considered part of a spectrum with minor changes in some parameter. Opinions on this topic have swung wildly from the days when the (innumerable) classes of AGN were developed, to recent times when many people argue that all AGN are explicable by a single model with only minor changes.

In order to arrive at the true answer (most probably somewhere between the above extremes), one needs data of comparable quality on a wide range of AGN types, covering a wide range in redshift and luminosity. Such a large dataset would never be produced by a single observer because of time allocation limitations, *but would inevitably accumulate in a Gemini archive*.

There are many promising unification candidates including: Broad absorption line and normal radio quiet QSOs, or BL Lac objects and low luminosity radio galaxies. Only by combining data from many different Gemini proposals will a large enough sample be obtainable. One can then start to assemble the entire picture, and finally, determine how many physical parameters are really needed to determine what an AGN will appear like to us. Being able to analyse uniformly NIRI imaging for all AGN, for example, would be a powerful use of a Gemini archive.

SC 4.1 Radio loud QSOs and luminous radio galaxies

Both AO and non-AO imaging with NIRI will allow comparison of the host galaxies of the two classes. IFU spectroscopy with GMOS and NIRS will show gas motions, both inflow and outflow over the galaxy nuclei, while polarimetry with NIRI and GMOS will show locations where scattered nuclear light may be found. To be successful, a unification model would have to explain any observed difference in the properties of the host galaxies or nuclear gas. It would also be supported by the discovery of strong scattered nuclear light in objects where it is not seen directly.

SC 4.2 Seyfert 1 and 2 galaxies

Several classic cases are known of Seyfert 2 galaxies which show broad lines when studied in polarised light. This has led some to claim that all Seyfert 2 galaxies are in fact Seyfert 1 galaxies, where the nucleus is obscured. There are however a number of problems with this simple hypothesis. Large well defined samples of Seyferts are needed, with high resolution imaging and IFU spectroscopy (NIRI, GMOS) to test this hypothesis.

SC 4.3 Multi-wavelength studies

Quasars and AGN often attain new interest as result of discoveries in varied wavelength regimes, for example, the X-ray and radio. Figure 2 on page 25 shows a set of data from various telescopes at various wavelengths that can be combined to derive information about the AGN and its host galaxy. Variability in some band, lensing, membership in large scale structure, or the presence of foreground absorbers all represent auxiliary information that might not be available when original Gemini observations are made, but which would renew interest in those observations. An archive thus plays a crucial role in maintaining the value of Gemini data for AGN science.

2.5 Dynamics of Galaxy Nuclei

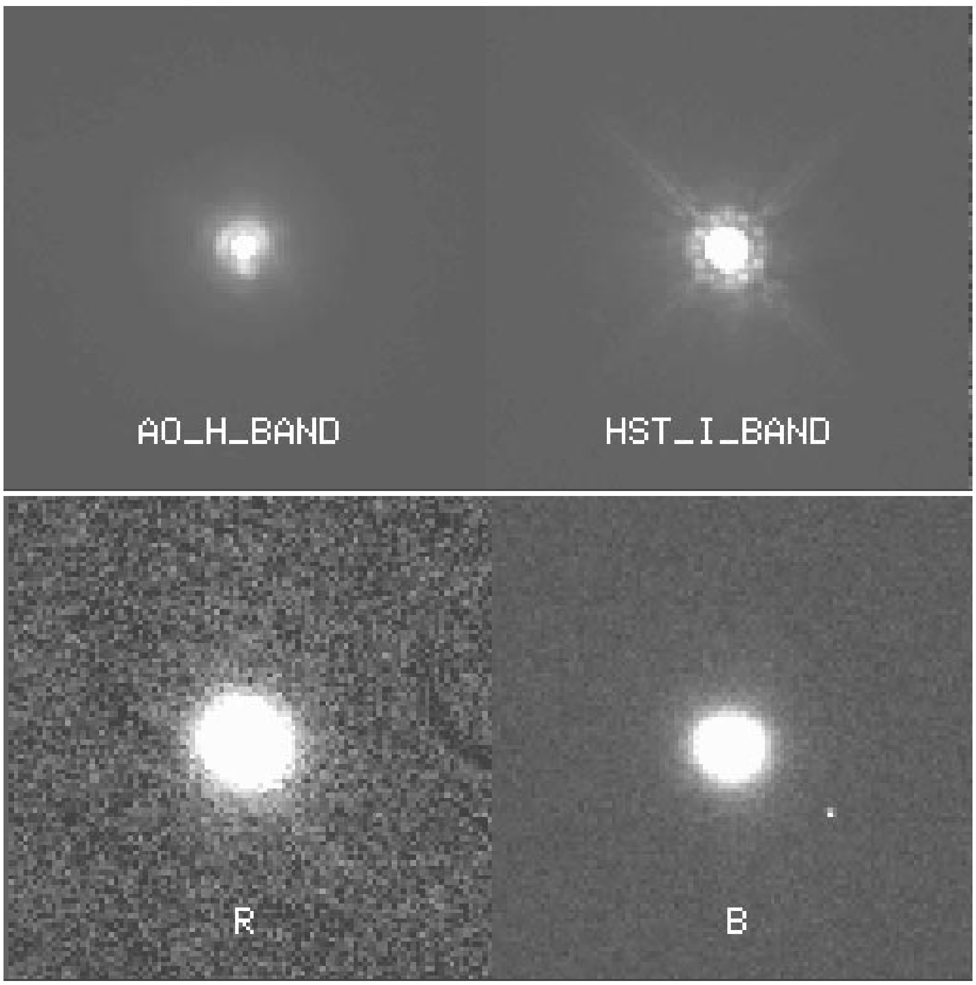
High-resolution spectroscopy of nearby galaxy nuclei frequently provide evidence for the presence of massive black holes. Light gathering power and high spatial resolution give Gemini a unique capability to extend these studies in terms of data quality and in terms of target distance. Gemini will produce much larger samples of observations of galaxy nuclei, and permit the reliable assessment of the true frequency of the black hole phenomenon and its relation to galaxy properties. Multiple groups will be involved in galaxy nuclei investigations, and one can expect many more studies like that of Magorrian et. al. (1997) mentioned is Section 1.1 on page 17 to use archive data in the future.

SC 5.1 The Galactic centre

The Galactic Centre provides an unprecedented laboratory for investigating the central regions of a late-type spiral galaxy. Early studies with single-element infrared detectors revealed a number of bright point sources, hinting that the stellar content near the Galactic Centre may be different from the surrounding bulge. Using state-of-the-art instrumentation, such as the CFHT Adaptive Optics system, it has been possible to obtain near-diffraction limited images of the central parsec of the Galaxy, such as the image shown in Figure 3 on page 26. It is now recognised that there is a population of young stars near the Galactic Centre that are centred around a supermassive object, corresponding to the radio source SgrA*. Efforts to study the region around SgrA* are confounded by (1) crowding, which limits efforts to resolve stars fainter than K~ 14, even with angular resolutions corresponding to the diffraction limit of 4-metre class facilities (Davidge et al. 1997, AJ, 114, 2686), (2) the potentially complicated kinematics of stars in this region, which introduces uncertainties in dynamical studies using only radial velocities, and (3) contamination from foreground (and possibly even background) disk objects, which becomes progressively more significant towards fainter magnitudes.

The second and third problems can be overcome by measuring proper motions, and archival data from Gemini (as well as other 8 metre telescopes) will provide the means of obtaining homogeneous stellar positions over moderately long time scales. When combined with radial velocities, proper motions can be used to deduce true space velocities, so that the orbits of stars about SgrA* can be constructed, and an accurate mass determined. Data of this nature will permit an extension

FIGURE 2. An example of multi-wavelength archive research combining Adaptive Optics with other existing data. This is an example of an X-ray selected AGN at z=0.037 observed at 4 different wavelengths from near-infrared to B-band (Schade & Crampton 1998 in preparation). All of these data are or will be available through archives. At left is an Adaptive Optics image in the H-band (pixel size 0.0375 arcseconds) taken at CFHT and next to it is an HST PC I-band image of the same object (scale 0.0455 arcseconds/pixel). These are displayed so as to show the diffraction rings in both images. The next two images are R-band and B-band images taken in seeing of 1.5 arcseconds (with pixel size 0.31 arcseconds) at La Palma with the JKT 1 meter telescope. The 1-meter integrations show the low surface brightness outer regions of the galaxies while the high-resolution images probe the inner structures with a good wavelength baseline. The boxes are each 101 pixels on a side. These data were simultaneously analyzed to obtain structural parameters of the inner region of the AGN and of the host galaxy.

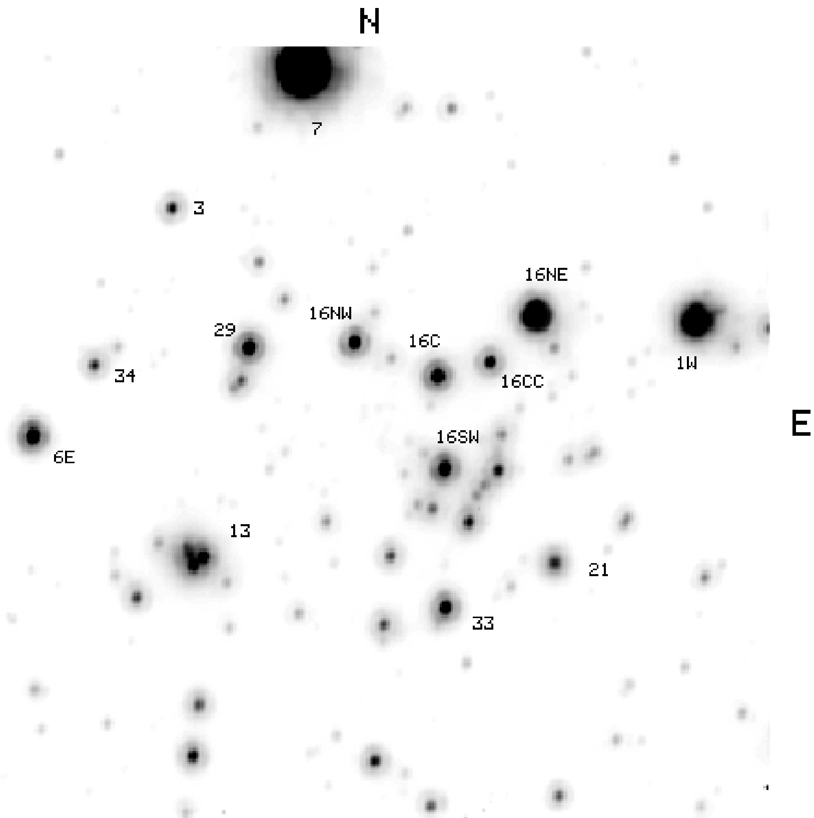


of the pioneering work of Eckart & Genzel (1996, Nature, 383, 415). Proper motion measurements can also be used to distinguish between bulge and disk stars, and thereby establish whether or not the faint blue objects detected by Davidge et al. 1997) are bona fide main sequence turn-off stars associated with SgrA, or foreground interlopers.

SC 5.2 High-order bars

Bars within bars are recognised to be a possible mechanism to fuel AGN. Stellar bars allow gas to crowd and collide, thus losing angular momentum and sinking to the centre. The search for secondary and tertiary bars in the centres of disk galaxies (Wozniak et al. 1995, A&AS, 111, 115; Friedli et al. 1996, A&AS, 118, 461) is ideally served by archives. Shlosman et al. (1989, Nature,

FIGURE 3. The central 12.4 x 11.9 arcsec field of the Galaxy, as observed in K with the CFHT AOB. The Airy pattern can be seen around the brightest sources. SgrA* is at the approximate center of the image. Various IRS sources are labelled.



338, 45) has predicted that a sequence of bars of different orders may cascade down in scale, in ratios such as 7:1. Large bars have typical size of 5-7 kpc, while secondary bars are ≤ 1 kpc. Bars are better seen in the near-infrared, probably because of reduced absorption at longer wavelengths. High-order (e.g., tertiary) bars would allow gas to get closer to the nucleus, e.g. $r \leq 20$ pc for a tertiary bar. The large number of disk galaxies that will be observed with Gemini in the infrared will provide a priceless source of images for searching for tertiary bars at the scale of 100 pc and for establishing the statistics of such features with respect to AGN properties.

2.6 Evolution of Cluster and Field Galaxies

Gemini will be heavily involved in surveys for high-redshift clusters and field galaxies and in spectroscopic and imaging follow-up for these surveys. Many images of faint galaxies will be obtained. Adaptive Optics will provide superb spatial resolution for both imaging and spectroscopy.

SC 6.1 Einstein crosses

In the infrared, Gemini will produce the best ground-based imaging ever done and wide field is not a requirement for discoveries of small scale gravitational lensing phenomena. An example of the type of serendipitous discoveries that should be expected comes from HST archive data. Figure 4 on page 27 and Figure 5 on page 27 show two cases where archival data was used to discover new gravitational lenses producing Einstein crosses. In the first case (Fischer et al. 1998 in preparation) the lens is a luminous cluster elliptical galaxy at z=0.46. The required lensing mass is consistent with the velocity dispersion implied by fundamental plane considerations if the source (most likely a quasar) is at z=~ 3. In the second case (Ratnatunga et al. 1995 ApJ 453, L5) the lens is an isolated field galaxy at z=0.81, and the source shows an emission feature interpreted as Lyα at z=3.4 (Crampton et al. 1996 A&A 307, L53). These discoveries were made as a result of re-processing of archive data with more sophisticated techniques than were used by the original observers. The Einstein crosses were not apparent until good modelling and subtraction of the lensing galaxy image was performed rendering the multiple images obvious.

FIGURE 4. Einstein cross discovered in archival HST data. (HST PI was Dressler, analysis by Fischer, Schade, Barrientos 1998 in preparation.) This object is in a cluster at z=0.46 and the lens is a luminous elliptical galaxy. The lensing was not detected in the original investigation but was revealed by applying new processing techniques. The right portion of the frame shows the image with the modelled galaxy profile removed.

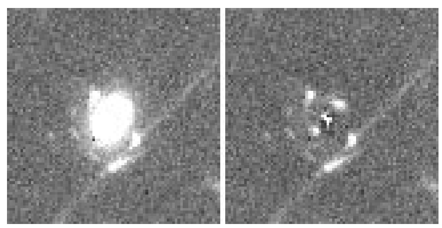
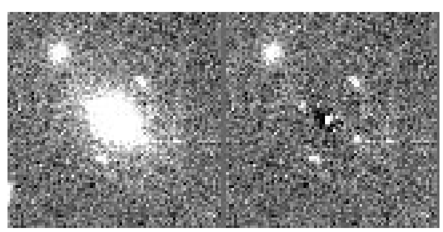


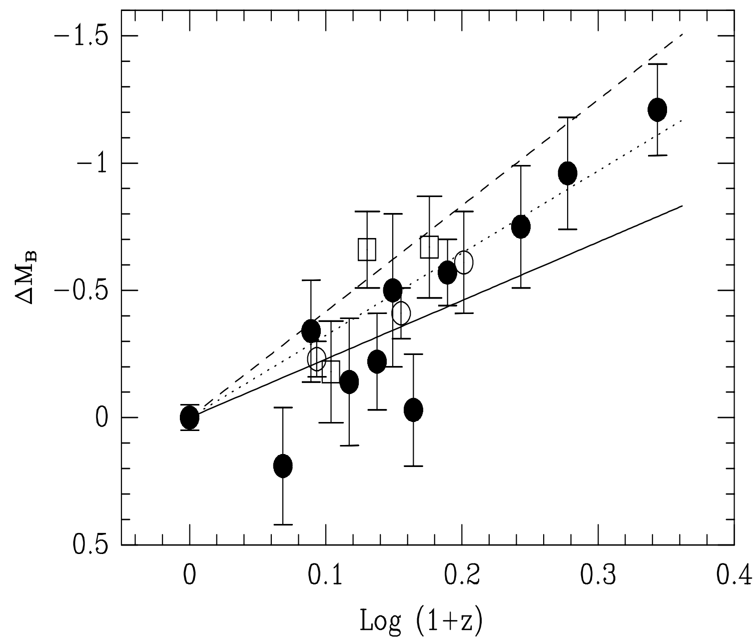
FIGURE 5. A second case of lensing discovered in archival data is this Einstein cross produced by a very luminous field elliptical galaxy at z=0.81. (HST PI Groth, analysis by Ratnatunga et al. 1995 ApJ 453, L5.) The source in this case is a quasar with redshift z=3.4. The right portion of the frame shows the image with the modelled galaxy profile removed.



SC 6.2 Luminosity evolution of elliptical galaxies

A second example of archive science illustrates the case where the aggregate effect of accumulated archive observations provided greater opportunities for discoveries than those available to the original proposers of the observations. Figure 6 on page 28 shows the evolution in luminosity (ΔM_B) of elliptical galaxies in clusters as a function of redshift (Schade et al. 1997 ApJ 477, L17). A collection of nine clusters (from seven distinct observational programs), spanning the redshift range $0.17 < z < 1.21$, was extracted from the HST archive. A uniform analysis technique was applied to trace the shift in luminosity of the sequence of giant elliptical galaxies with increasing look-back time. Each point on this diagram represents not a single galaxy, but a sequence of galaxies in a single cluster (from 4 to 28 individual galaxies). The results indicate that elliptical galaxies undergo luminosity evolution consistent with models of the passive aging of a single-burst stellar population formed at $z > 3$.

FIGURE 6. Luminosity evolution of elliptical galaxies from archival HST data (Schade et al. 1997 ApJ 477, L17). ΔMB is the shift in luminosity at a given size as measured from the sequence of giant elliptical galaxies. Each point represents not a single galaxy but the shift of a sequence of galaxies at a given redshift or in a given cluster. Solid symbols are for cluster elliptical galaxies using *HST* imaging and open symbols are from ground based CFHT fields. Open circles are cluster E's and open squares field E's.



SC 6.3 Gravitational optics

A third example is the study of spectrophotometric evolution of distant galaxies with cluster gravitational optics: cluster lenses are “high z filters” which select background galaxies and distort them, making their shapes easy to identify while foreground galaxies are undistorted. Only objects with $z \geq 2$ *zlens* can be selected with gravitational “telescopes” (Figure 7 on page 29 and Figure 8 on page 30). For each lensing cluster, a set of arclets gives a subsample of a global population of faint field galaxies. Thus a large number of clusters, observed under fine imaging conditions, are needed to be able to follow the spectrophotometric evolution of all morphological types of galaxies. Imaging and spectroscopy of many high-redshift clusters will be conducted with the Gemini telescopes, and this will provide an opportunity to assemble a statistically reliable sample of high redshift lensed galaxies within a reasonable time frame.

2.7 Science sensitive to target selection criteria

Many studies which could use data in the GSA are sensitive to the target selection criteria used when designing an observing program. These selection criteria are a chain of selections which include both conscious choices made by observers, and implicit selections caused by the resources (catalogues, the availability of data in archives, published observations of other researchers, etc.) used when observers select the targets to be observed. The result of these selections is that the population of observed targets may be in some way biased, in a way which is understood by the PI not to effect his/her science.

FIGURE 7. HST image (F702W) of galaxy cluster A2218.

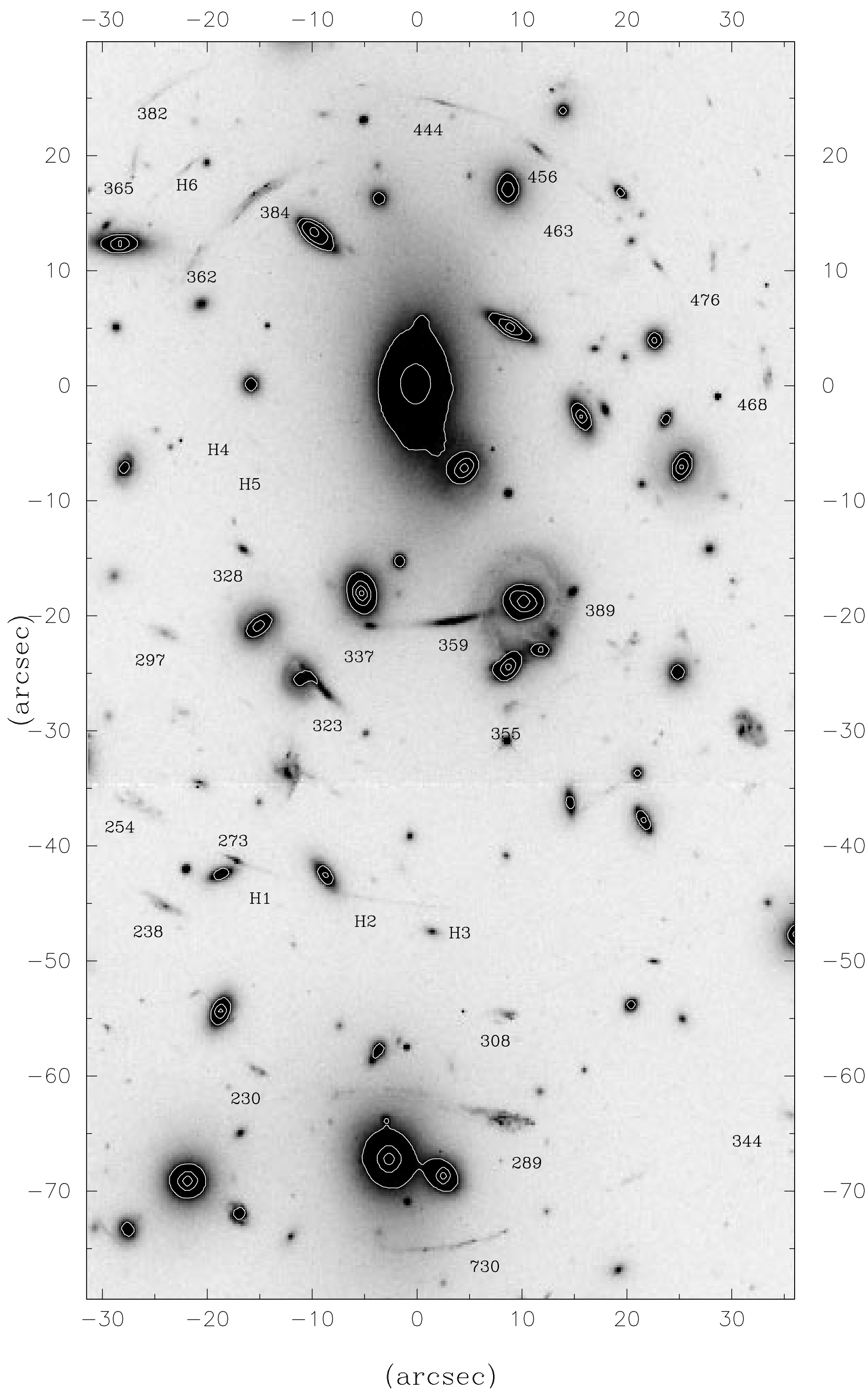
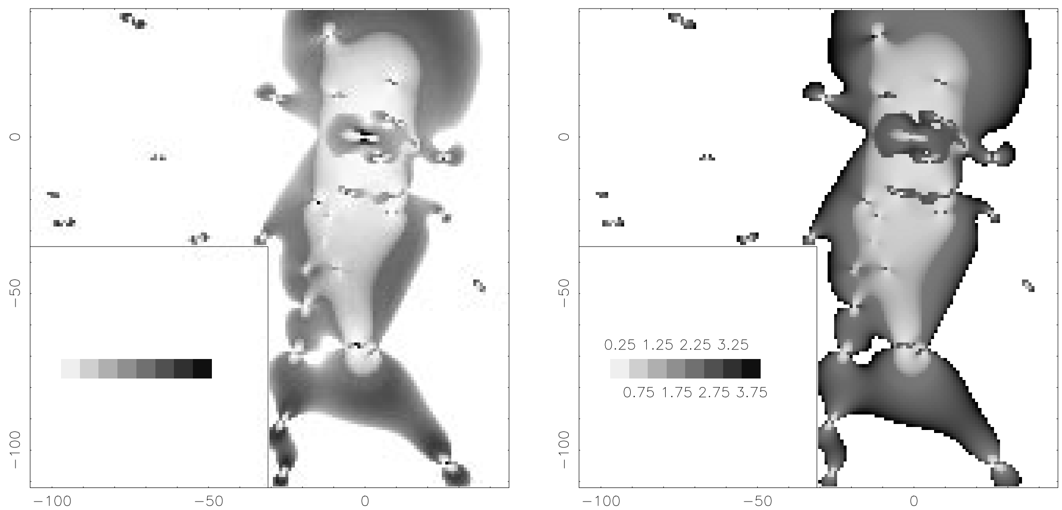


FIGURE 8. Left: Density of arclets in cluster A2218 (with B ≤ 24.5) for a given model of mass distribution. Right: Mean redshift of arclets in cluster A2218; more distant “arc” galaxies appear further away from the cluster center than less distant ones. (Bezecourt et al. 1998).



If an archive researcher knows that his/her research is sensitive to certain types of target selection bias, then the observations which are subject to this bias must be removed from the set of archive observations used.

Ideally an archive should contain information describing the target selection criteria used for each observation, and should provide the tools required to allow an archive user to exclude observations where the target selection criteria would effect the results of his/her study.

SC 7.1 AGN host galaxy studies

The GSA could be used as a source of a large number of images of AGNs and their host galaxies. An archive researcher could use these images to study the statistical distribution of the properties of an X-ray flux limited sample of AGNs. A set of observations could be selected from the GSA by joining the GSA catalogue with the Einstein Medium Sensitivity Survey (EMSS).

Assume that one of the observing programs used in the archival study evolved as follows:

1. An X-ray flux limited sample was compiled and optical counterparts identified (e.g., the EMSS).
2. A sample of all AGNs in this survey with $z < 0.15$ (120 objects) was selected for HST imaging, resulting in successful observations of 76 objects.
3. A Gemini observer creates an observing program to study 10 of the AGNs from the HST sample, all selected because they have elliptical galaxies as hosts.

If the user is only aware of the initial X-ray flux selection, and if the above science program is the only one in the GSA, then she/he could conclude that all X-ray selected AGN hosts are at $z < 0.15$ or that all AGN hosts are elliptical galaxies. If this science program was only one of many in the GSA, then the results will be biased in a much less obvious way.

Both the HST and the Gemini archive data can be used effectively for these types of studies, but only if the chain of selection is known.

SC 7.2 Stellar populations in the Milky Way

A great deal of imaging will be available at various galactic latitudes and longitudes that would be suitable for studying the structure of the Milky Way. Infrared imaging data has the advantage of low sensitivity to galactic extinction.

Gemini Science Archive could be used to study galactic structure through the study of star counts in various directions on the sky. In this situation the GSA would be used as a source of stellar field images, unbiased by any image selection criteria used by or original observer.

2.8 Education

The Gemini Science Archive will enable undergraduate and postgraduate educators a wealth of material to stimulate and train the next generation of astronomers in a range of activities from proposal generation, to data processing and analysis.

A few possibilities are discussed below.

SC 8.1 Proposal generation

Many PhD projects require original observations at state-of-the-art facilities such as Gemini. The Gemini Science Archive will provide young (as well as experienced) researchers with essential data required to produce the compelling observing proposals required to acquire time on today’s heavily oversubscribed facilities. Archive users will be able to determine if data relevant to their proposal may already exist in the Science Archive and modify their program as required. The retrieval and processing of observations of similar targets acquired with similar instrument configurations can also be used to evaluate instrument performance as it pertains to his or her particular science requirements, and to evaluate various observing strategies.

SC 8.2 Re-processing of archived observing runs

Advanced undergraduate and beginning graduate students could acquire data from the GSA, and subsequently process the data with the appropriate tools. For example, an instructor could retrieve a number of executed proposals from the archive and assign a proposal to each student. In its simplest form, an assignment might consist of the retrieval of the unreduced observations associated with the proposal from the archive and the subsequent processing of these data with IRAF (or other software). The end result could then be compared to the data produced by the Archive’s pipeline processing scripts.

Archived observing runs could also be used as a source of test data for use by astronomers creating new tools or algorithms for data reduction.

SC 8.3 Archival research training

With the exponential growth rate in the quantity of archival astronomical data, students will soon require training simply to learn how to search, relate, and make productive use of such huge amounts of information. With advanced search capabilities available in the Gemini Science Archive it will make a logical starting point for archival research training.

2.9 Targets of Opportunity

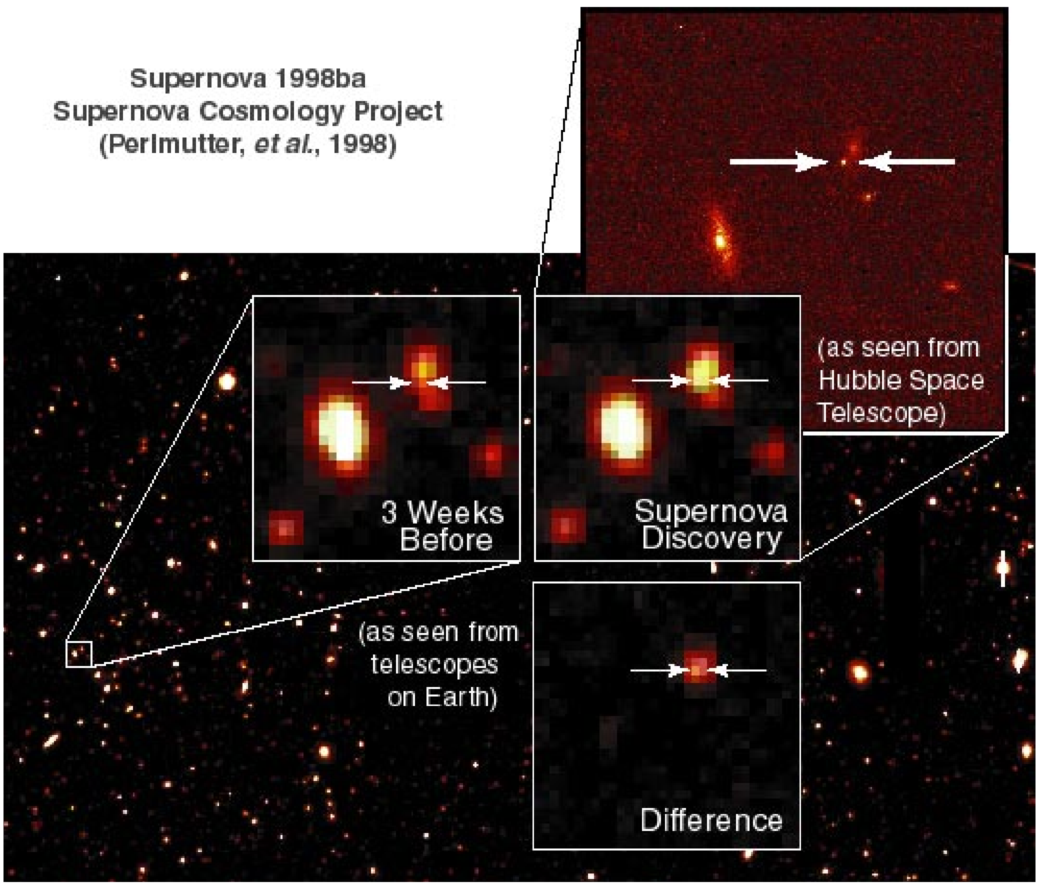
Many astronomical phenomena are transient in nature, and will be serendipitously discovered by astronomers doing unrelated studies. The GSA will provide a resource which will help to study the evolution of these transient phenomena. The GSA may also allow an observer to retrieve historical data for a field, which will help him/her to determine if a target is truly transient.

SC 9.1 Supernova study

The luminosity evolution of a supernova not only helps to understand supernova explosions, but also establishes the cosmological scale. Because all type Ia supernovae have the same intrinsic light curves (a measurement of the rise and fall of their light output), they can act as clocks over cosmological distances. Theories of the Big Bang have successfully predicted the abundances of light elements. The first stars were composed of hydrogen, helium and a very small amount of lithium and beryllium and almost nothing else. Some of these stars became supernovae and distributed the “heavy” elements, made in their interiors, into the interstellar environment. Subsequent generations of stars have further increased the proportion of “heavy” elements such as carbon, oxygen, phosphorus and iron. Studies of supernovae are then quite important for cosmological research.

Supernovae occur at a frequency of about 1 per galaxy per 100 years, and so opportunities to study them cannot be overlooked. Usually, supernovae are serendipitous discoveries, often when they are close to their maximum luminosity. It is very important to be able to go back in time, i.e. to an archive, to be able to trace the shape of the luminosity curve of the precursors. In this case, it is essential that archive users have a way of gaining permission to access proprietary data since the window of opportunity to study the supernova is well within the proprietary period. Figure 9 on page 32 shows images of supernova 1998ba both before and after the supernova occurred. This supernova was discovered by the Supernova Cosmology Project (http://www.lbl.gov/supernova).

FIGURE 9. SN 1998ba



SC 9.2 Transient target detection

Another scenario could be that an observer notices immediately after acquiring an image, the presence of an unfamiliar feature or object. The feature could be a new supernovae, Gamma Ray burst, asteroid, or any other transient event. The observers first reaction should be to look in the GSA (and other archives) for previously acquired exposures of the same field, in order to verify the discovery and be able to justify an **immediate** change to his/her observing program. It is thus important to insure that all non-proprietary observations in the GSA be available right away, at any time, and reasonably promptly.

3. Other Considerations

3.1 Feasibility Checks, Proposal Planning and Optimization

The Science Data Archive would allow proposers to evaluate quickly the feasibility of their program by study of actual data retrieved from the archive. There is no substitute for seeing results from the same instrument and configuration that an observer is considering using. Access to real data would allow proposers to estimate reliably the required integration times, and allow them to evaluate the effect of different atmospheric emissivity and image quality on the data needed for their science.

3.2 Public Outreach

While certainly not a driver of the requirements for the Gemini Science Archive, the use of the Archive as a tool for Public Outreach warrants serious consideration.

The Canadian Astronomy Data Centre (CADC) provides a good example. The CADC currently has close to 1200 registered users of their facilities, the vast majority of whom are professional astronomers who are required to register in order to retrieve data from the CADC-supported archives or use facilities such as the CADC's interface to the SIMBAD astronomical database.

In order not to inflate usage statistics the CADC does not normally provide registration ID's and passwords to the casual browser. However, such requests from the general public, *including educators at all age levels*, generally exceed the number of "legitimate" requests from the astronomical community. The general public obviously has a very strong interest in seeing astronomical images obtained at world-class facilities and the CADC encourages them to view the preview images provided by their implementation of the HST and other archives. Other "free" services such as the Digital Sky Survey are also extremely popular.

Provided that care is taken to provide an intuitive user interface, the Gemini Science Archive can readily be designed to accommodate the needs of both the professional astronomer (who requires the ability to make complex queries of the GSA, possibly combined with other archives) and interested lay person (who may simply want to view previews of the latest public images). Working in the opposite direction, some thought should also be given to providing links from any Gemini Public Relations images, once the data are no longer proprietary, to the Gemini Science Archive. Archival researchers, with a few clicks of a button, could then retrieve all of the calibrated data associated with the press release for their own investigations.

The very existence of the Gemini Science Archive should also be widely advertised in order to demonstrate to the taxpayers just how valuable the international astronomical community considers Gemini telescope data.

3.3 Quick-look Data Processing

The intention of Gemini quick-look data processing requires the existence of good calibration material and automated processing in real time. Thus a processing pipeline needs to be defined for each instrument where quick-look tools are implemented. Minimal calibration steps should include bias correction, flat-fielding and wavelength calibration. Additional effort would provide mosaicing, co-adding, image distortion correction, atmospheric refraction correction and fluxing. These requirements directly benefit the GSA in that the automatic association of science data with calibration data, the automatic re-processing of data, and the ability to provide a reasonable set of calibration data are all requirements of an archive.

3.4 Queue Scheduling

Simulations (Mountain, Simons, & Boroson, 1995 RPT-PS-G0053) provide evidence for substantial scientific gain from queue scheduling; MSB discuss models where between 70% and 100% of observing time is dedicated to this mode of operation. The most compelling science is often the most challenging technically, and the observations that fully exploit the best atmospheric conditions are likely to be obtained in queue mode.

Queue-mode observing automatically satisfies the requirements of a Science Data Archive. The data are obtained and accompanied by a) sufficient calibration material, b) electronic logs of the events that took place during data acquisition, and c) weather monitoring information. These products delivered to the proposers are sufficiently complete and reliable to allow them to carry out the desired analysis with confidence. There is no difference between delivering these products to the original proposer several hours after they were obtained and delivering them, following the proprietary period, to an archive user.

3.5 Calibration

In order that data be “archive-able”, observations must always be accompanied by sufficient calibration material and logs of weather and other events. **The definition of specific calibration requirements, which largely determine the value of the Gemini science archive, requires ongoing consultation with the user community.** The information should be complete enough to redo the observations in exactly the same way. This requirement applies equally to queue-mode and classical-mode observing. There are several benefits to the classical-mode observer of obtaining the minimum calibration data. First, instrument teams and observing staff who use the instrument frequently possess a great depth of expertise and would advise which calibration procedures are necessary and would implement those procedures correctly and efficiently so that the minimum necessary amount of time is expended on calibration. Secondly, calibration data should be treated as “shared” data. This will often result in further reductions in calibration effort, since each observer is not required to obtain a complete calibration dataset where redundant material already exists. An additional benefit is that the quality of the calibration provided to the classical (as well as the queue-mode) observer would be as good as or better than they would otherwise have obtained. Finally, the calibration material would be tuned for integration into both the quick-look evaluation tools and into the processing pipeline that would exist for each instrument.

3.6 Pipeline Processing

The observer in either queue or classical mode would benefit from pipeline processing whether they use it to produce their final data products or whether they use it as a benchmark to evaluate the results of their own processing software. The production of pipeline software is an integral part of the archive process, but it is also an integral part of the implementation of quick-look tools for real-time evaluation of data quality. As is the case for calibration material, the instrument teams, and Gemini staff astronomers, will have considerable depth of expertise on processing of data from each instrument and would develop processing algorithms in consultation with users. The standard pipeline processing software will not satisfy all telescope users but, on balance, it would save most observers a considerable amount of time normally spent on routine data reduction and software development.

4. Conclusions

A Science Data Archive would represent a major contribution to the scientific productivity of the Gemini observatories in a number of ways. First, we have given a number of examples where it would enable first-rate scientific research that would never be done in the absence of an archive. Second, a data archive carries benefits for proposal preparation, instrument performance verification and optimization, queue-mode observing, and Gemini operations in general. This is because it is more efficient to operate in an environment where information is managed and processed with the best existing technological tools. An archive is a major component of such an environment. Third, it would help to keep us competitive. The VLT and Subaru projects realize the power of effective archiving and are investing large resources in archive development. We know how to develop and run a highly productive science archive with a fraction of the resources these projects are expending. Finally, the Gemini Science Archive would distribute scientific opportunities among astronomers in the partner countries, would help to inform the public about the excitement and importance of astrophysical research, and would demonstrate to the taxpayers who support Gemini that we were acting responsibly in ensuring that these valuable data are being handled with the greatest care to ensure that they are fully exploited and that their value is being preserved for future generations of scientists.

Gemini Science Archive

Chapter 4 Operational Scenarios

1. Introduction

This chapter describes the operational scenarios which would enable a GSA user to do the science described in Chapter 3. These operational scenarios are then used to define the user requirements, which are listed in Chapter 5. Table 1 on page 37 and Table 2 on page 39 link the operational scenarios with the relevant science cases from Chapter 3.

TABLE 1. Operational scenario to science case cross reference

OS 1 Search by position
SC 1.1 Detection of newly born stars
SC 1.2 Birth of Herbig-Haro systems
SC 1.3 Motions of jets and knots
SC 1.4 Variability of pre-main sequence objects
SC 1.5 The fate of very massive stars
SC 5.1 The Galactic centre
SC 8.1 Proposal generation
SC 8.3 Archival research training
SC 9.1 Supernova study
SC 9.2 Transient target detection
OS 2 Retrieve proprietary data associated with an known proposal
This scenario doesn't have a specific science case attached to it. The requirement to provide data for PIs and CoIs has emerged from discussions between the GSA design team and Gemini staff.
OS 3 Search by position, qualified by image descriptors
SC 1.1 Detection of newly born stars
SC 1.2 Birth of Herbig-Haro systems
SC 1.3 Motions of jets and knots
SC 1.4 Variability of pre-main sequence objects
SC 1.5 The fate of very massive stars
SC 3.1 Interstellar medium in the Magellanic Clouds
SC 4.1 Radio loud QSOs and luminous radio galaxies
SC 4.2 Seyfert 1 and 2 galaxies
SC 4.3 Multi-wavelength studies
SC 5.1 The Galactic centre
SC 8.1 Proposal generation

TABLE 1. Operational scenario to science case cross reference

SC 8.3 Archival research training
SC 9.1 Supernova study
SC 9.2 Transient target detection
OS 4 Search by class of object, qualified by data superset and object descriptors
SC 1.2 Birth of Herbig-Haro systems
SC 1.3 Motions of jets and knots
SC 1.4 Variability of pre-main sequence objects
SC 1.5 The fate of very massive stars
SC 2.1 Diffuse interstellar bands
SC 2.2 Chemical evolution of the Galaxy
SC 3.1 Interstellar medium in the Magellanic Clouds
SC 4.1 Radio loud QSOs and luminous radio galaxies
SC 4.2 Seyfert 1 and 2 galaxies
SC 5.2 High-order bars
SC 6.1 Einstein crosses
SC 6.2 Luminosity evolution of elliptical galaxies
SC 6.3 Gravitational optics
SC 7.1 AGN host galaxy studies
SC 8.1 Proposal generation
SC 8.3 Archival research training
OS 5 Selection of data unbiased by field selection criteria
SC 6.2 Luminosity evolution of elliptical galaxies
SC 7.1 AGN host galaxy studies
SC 7.2 Stellar populations in the Milky Way
SC 8.3 Archival research training
OS 6 Cross archive search
SC 4.3 Multi-wavelength studies
SC 8.3 Archival research training
OS 7 Multi instrument and multi archive search
SC 4.3 Multi-wavelength studies
SC 8.3 Archival research training
OS 8 Search by proposal
SC 2.1 Diffuse interstellar bands
SC 2.2 Chemical evolution of the Galaxy
SC 3.1 Interstellar medium in the Magellanic Clouds
SC 4.1 Radio loud QSOs and luminous radio galaxies
SC 4.2 Seyfert 1 and 2 galaxies
SC 4.3 Multi-wavelength studies
SC 5.2 High-order bars

TABLE 1. Operational scenario to science case cross reference

SC 6.1 Einstein crosses
SC 6.2 Luminosity evolution of elliptical galaxies
SC 6.3 Gravitational optics
SC 7.1 AGN host galaxy studies
SC 7.2 Stellar populations in the Milky Way
SC 8.1 Proposal generation
SC 8.2 Re-processing of archived observing runs
SC 8.3 Archival research training

TABLE 2. Science case to operational scenario cross reference

SC 1.1 Detection of newly born stars
OS 1 Search by position
OS 3 Search by position, qualified by image descriptors
SC 1.2 Birth of Herbig-Haro systems
OS 1 Search by position
OS 3 Search by position, qualified by image descriptors
OS 4 Search by class of object, qualified by data superset and object descriptors
SC 1.3 Motions of jets and knots
OS 1 Search by position
OS 3 Search by position, qualified by image descriptors
OS 4 Search by class of object, qualified by data superset and object descriptors
SC 1.4 Variability of pre-main sequence objects
OS 1 Search by position
OS 3 Search by position, qualified by image descriptors
OS 4 Search by class of object, qualified by data superset and object descriptors
SC 1.5 The fate of very massive stars
OS 1 Search by position
OS 3 Search by position, qualified by image descriptors
OS 4 Search by class of object, qualified by data superset and object descriptors
SC 2.1 Diffuse interstellar bands
OS 4 Search by class of object, qualified by data superset and object descriptors
OS 8 Search by proposal
SC 2.2 Chemical evolution of the Galaxy
OS 4 Search by class of object, qualified by data superset and object descriptors
OS 8 Search by proposal

TABLE 2. Science case to operational scenario cross reference

SC 3.1 Interstellar medium in the Magellanic Clouds	
OS 3	Search by position, qualified by image descriptors
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 4.1 Radio loud QSOs and luminous radio galaxies	
OS 3	Search by position, qualified by image descriptors
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 4.2 Seyfert 1 and 2 galaxies	
OS 3	Search by position, qualified by image descriptors
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 4.3 Multi-wavelength studies	
OS 3	Search by position, qualified by image descriptors
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
OS 8	Search by proposal
SC 5.1 The Galactic centre	
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors
SC 5.2 High-order bars	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 6.1 Einstein crosses	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 6.2 Luminosity evolution of elliptical galaxies	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 5	Selection of data unbiased by field selection criteria
OS 8	Search by proposal
SC 6.3 Gravitational optics	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 7.1 AGN host galaxy studies	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 5	Selection of data unbiased by field selection criteria
OS 8	Search by proposal

TABLE 2. Science case to operational scenario cross reference

SC 7.2	**Stellar populations in the Milky Way**
OS 5	Selection of data unbiased by field selection criteria
OS 8	Search by proposal
SC 8.1	**Proposal generation**
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 8	Search by proposal
SC 8.2	**Re-processing of archived observing runs**
OS 8	Search by proposal
SC 8.3	**Archival research training**
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 5	Selection of data unbiased by field selection criteria
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
OS 8	Search by proposal
SC 9.1	**Supernova study**
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors
SC 9.2	**Transient target detection**
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors

2. Operational Scenarios for the Gemini Science Archive

This section describes each of the operational scenarios. Each step of the interaction between the archive user and the archive is described in detail. Following the scenario, the key GSA requirements implied by the scenario are listed, with a description of the importance of the requirement to the scenario. Note that only key requirements are listed here, for a complete list of the requirements and the scenarios to which they apply, see Chapter 5.

OS 1 Search by position

In this operational scenario, an astronomer is doing a simple search by position:

1. Determine accurate coordinates for the search.
2. Enter the coordinates into the GSA interface. The coordinates should have an associated radius to allow searching for extended objects, or for targets where the position isn't accu-

rately known. There should be a way of searching for a list of coordinates provided by the user.

3. Retrieve a list of data supersets from the GSA catalogue which match the specified coordinates.
4. Evaluate each data superset's suitability, based on the information stored in the GSA catalogue, including date the data superset was taken, instrument configuration, telescope configuration, observing conditions, and data quality indicators.
5. View previews of selected data supersets in order to evaluate the suitability of the data supersets.
6. Retrieve all suitable public data from the GSA, including calibration data.
7. Optionally retrieve re-calibrated data.

The key requirements associated with this scenario are:

- The archive data must be fully described

  This is necessary to allow the archive user to determine which data supersets are suitable for his/her project.
- Select sets of data supersets based on coordinate

  This allows the astronomer to search for data supersets which have a specified coordinate in the field of view. This is necessary in order to complete this scenario.
- Search based on a list of coordinates

  This allows the astronomer to simultaneously search the archive for multiple targets.
- The GSA must display data superset descriptors to the user

  This allows the astronomer to select specific data supersets which are relevant to the investigation. If this isn't possible, the astronomer will have to retrieve a superset of the desired data, and then eliminate inappropriate data.
- GSA users may preview data for selected data supersets

  This provides the astronomer with the information required to determine the suitability of individual data supersets.
- Automatic re-processing of data should be an option

  This eliminates the need for the astronomer to re-process the data being retrieved. This would be a significant advantage for complex instruments with which the astronomer is not familiar.
- Calibration data must be easily available

  This provides astronomers with the option to re-process the raw data manually, if the automatic re-processing procedures are inappropriate.
- Data requested by users should be available quickly

  In order to support the target of opportunity science case, data requested by users must be available to them reasonably quickly.
- The GSA should be available 24 hours a day

  To support the target of opportunity science case, GSA users must be able to get data at any time of the day.
- Retrieve proprietary data

  In science case SC 9.1, it is important that the astronomer have a possible method for gaining access to proprietary data, since the observations of most interest would invariably be proprietary.

OS 2 Retrieve proprietary data associated with an known proposal

In this scenario, the GSA user is a PI or CoI who wants to retrieve some or all of the data associated with completed observations for a Gemini proposal, including data which is still proprietary.

1. The user enters the proposal id into the GSA user interface, and retrieves a list of data available in the GSA for that proposal.
2. The user selects some or all of the data for retrieval.
3. The user provides proof of his/her right to access the proprietary data.

The key requirements of this scenario are:

- Search for data based on proposal

  This allows the user to select all data supersets associated with a proposal.
- Retrieve proprietary data

  This allows the user to retrieve data before it becomes public.

OS 3 Search by position, qualified by image descriptors

In this operational scenario, an astronomer detects and wishes to investigate some time variable phenomenon. The time variable nature of the phenomenon may known beforehand (for example, in the case of jets and knots associated with recently formed stars), or it may be discovered serendipitously during the collection, processing, or analysis of new data (for example the discovery of emission sources where it is known that no source existed at some previous date). In this scenario, the GSA would be used to locate and retrieve historical data which could be used to investigate the evolution of the phenomenon. In this scenario the astronomer would:

1. Determine accurate coordinates for the phenomenon to be investigated.
2. Specify which instrument or instruments are of interest based on the capabilities of the instrument. It should also be possible to qualify on instrument mode and other data descriptors (for example wavelength, exposure time, etc.)
3. Specify a minimum signal to noise ratio.
4. Indicate that only data supersets where the coordinates of the target of interest is in the field of view should be listed.
5. Retrieve a list of data supersets from the GSA catalogue.
6. Evaluate each data superset's suitability, based on the information stored in the GSA catalogue, including observation date, instrument configuration, telescope configuration, and data quality indicators. This may also include previewing the data.
7. Retrieve all suitable public data from the GSA. Since the astronomer may be retrieving data from several archives and from several different instruments from each archive, providing appropriately re-calibrated data could save considerable work on the part of the astronomer. In this case, re-calibration should **not** include stacking overlapping images, since observation date and time are important parameters.
8. If re-calibrated data are either not available or not appropriate, the astronomer should be able to easily get the raw data and all required calibration data, including PSF models, and instructions describing the appropriate processing procedures for the data.

The key requirements associated with this scenario are:

- The archive data must be fully described

This is necessary to allow selection of subsets of the data supersets based on the values of descriptors or coordinates, and to provide the user with the information required to evaluate the data supersets.

- Select sets of the data supersets in the GSA

  This allows the astronomer to eliminate irrelevant data supersets from the set displayed to the user. This is necessary in order to complete this scenario.

- Select sets of data supersets based on coordinate

  This allows the astronomer to search for data supersets which have a specified coordinate in the field of view. This is necessary in order to complete this scenario.

- The GSA must display data superset descriptors to the user

  This allows the astronomer to select specific data supersets that are relevant to the investigation. If this isn't possible, the astronomer will have to retrieve a superset of the desired data, and then eliminate inappropriate data.

- GSA catalogue needs descriptors derived from the data

  Derived descriptors would include signal to noise ratio, which in this scenario is one of the search parameters. S/N could also be used by the user when evaluating individual data supersets.

- Data quality indicator must be one of the data descriptors

  This allows the user to reject data supersets which don't have the required data quality.

- Support multi archive searches

  This would allow the astronomer to search many archives without the need to be aware of the detail content of each archive.

- A general description of the GSA must be available to users

  This allows astronomers to easily determine if the GSA is likely to contain data relevant to their investigation.

- A description of the instruments must be available

  This allows the astronomer to determine if an instrument and instrument mode could produce data relevant to their investigation. It would also allow astronomers to determine the correct calibration and processing procedures for the instrument.

- GSA users may preview data for selected data supersets

  This provides the astronomer with the information required to determine the suitability of individual data supersets for their investigation.

- Automatic re-processing of data should be an option

  This eliminates the need for the astronomer to re-process the data being retrieved. This would be a significant advantage for complex instruments with which the astronomer is not familiar.

- Users should be able to choose one or more processing recipes

  The option to choose from several processing recipes would allow the user to avoid getting stacked data supersets, which in this scenario are undesirable.

- Information about re-processing recipes should be available

  This allows astronomers to determine which, if any of the automatic re-processing procedures are appropriate for their investigation.

- Calibration data must be easily available

  This provides astronomers with the option to re-process the raw data manually in the event that the automatic re-processing procedures are inappropriate.

OS 4 Search by class of object, qualified by data superset and object descriptors

In this scenario, an astronomer is looking for spectroscopic data supersets of stars, where both the class of star being observed, and the nature of the data are qualified.

1. Optionally restrict the search to one of the megallanic clouds by specifying a region of the sky to search. The region is specified by a centre coordinate (r.a. and dec.) and a search radius (seconds of arc).
2. Select all spectroscopic data supersets of O- and B-type stars brighter than a specified limiting magnitude. Note we're not searching by coordinate or ID but by class of object and object properties, both taken from SIMBAD catalogue. It would also be of interest to sort the data supersets by distance or the reddening parameter E(B-V).
3. Filter this list by restricting it to spectra with resolution, R, > 50,000. This could also be done by restricting the search to specific instruments and modes, however this is less satisfactory.
4. Continue filtering by assuring that a wavelength range or ranges of particular interest are contained within a spectrum. (HROS is to provide continuous spectral coverage between 325-830nm, but there are also DIB beyond 830nm...)
5. Filter the list still further by limiting selected data supersets to those with a S/N greater than, say, 1000.
6. Since we're not interested in time variable phenomena, it would be best to be able to combine automatically all multiple spectra of individual objects and then evaluate the S/N of the combined spectra before eliminating spectra. Hence the reference to "associations" above.
7. Retrieve the processed spectra and proceed with analysis.

This scenario specifically addresses the needs of science cases SC 2.1, and SC 3.1 in Chapter 3, however it could easily be generalized to support several other science cases (see Table 1 on page 37 for a list).

The Key requirements associated with this scenario are:

- The search may be limited to a large region of sky based on a centre coordinate and a radius.

  This is necessary for SC 3.1 in order to limit the results to data supersets of objects in one of the megallanic clouds.
- The archive data must be fully described

  This is necessary to allow selection of subsets of the data supersets based on the values of descriptors, and to allow the user to view the descriptors for each data superset.
- Select sets of the data supersets in the GSA

  This allows the astronomer to search by class of observation (image verses spectral), spectral resolution, instrument, instrument mode, by spectral band, and for single aperture spectra only.
- A description of the instruments must be available

  This allows the astronomer to determine if an instrument and mode could produce data relevant to their investigation. It would also allow astronomers to determine the correct calibration and processing procedures for the instrument.
- Automatic re-processing of data should be an option

  This eliminates the need for the astronomer to re-process the data being retrieved. This would be a significant advantage for complex instruments with which the astronomer is not familiar.
- Users should be able to choose one or more processing recipes

The ability to chose processing recipes would allow the astronomer to retrieve associated datasets wherever possible. In this case, associated datasets would mean averaging overlapping spectra of the same object, and/or merging spectra which cover different wavelengths.

- Information about re-processing recipes should be available

  This allows the astronomer to determine if the automatic re-processing procedures are appropriate for their investigation.

- Select data supersets based on object parameter

  This allows the astronomer to select data supersets which are appropriate for the investigation. In this scenario, it is assumed that the object parameters are extracted from the SIMBAD or NED catalogues, although other catalogues could also be included (EMMS for example).

- Sort the displayed set of data supersets

  This allows the astronomer to sort the displayed set of data supersets based on distance or reddening parameter.

- GSA catalogue needs descriptors derived from the data

  This would allow the astronomer to display and search by signal to noise ratio.

- The catalogue should contain information on associated data

  This would allow the astronomer to locate and retrieve associated data.

- Calibration data must be easily available

  This provides the astronomer with the option to manually re-calibrate the raw data in the event that the automatic re-calibration procedures are inappropriate.

OS 5 Selection of data unbiased by field selection criteria

A researcher wishes to study the structure of the Milky Way galaxy by doing deep star counts in selected fields. Fields that were chosen in any way which would yield a biased estimate of the stellar population should be excluded.

1. Choose deep (exposure time > 1000 seconds) Gemini NIRI fields.
2. Exclude fields which are targeted at star clusters (globular or open).
3. Exclude fields with targets that might be associated with large galactic extinction (i.e. star forming regions).
4. The positions of the resulting set should be plotted in galactic coordinates to allow evaluation of spacial coverage.
5. If the number of images matching the above criteria is large (e.g. >500), reducing the spacial density of image set to a function based on galactic latitude.
6. The selected images should be re-calibrated using the latest calibration data and algorithms.
7. Down-load the re-calibrated image data.

The key requirements for this scenario are:

- Reject data supersets based on field selection criteria

  To avoid selecting a statistically biased set of data supersets, the astronomer must be able to exclude data supersets where the field was intentionally selected to include star clusters, or areas with large galactic extinction. If these data supersets cannot be excluded from the results, the data will be statistically biased.

- It should be possible to plot the fields of view of a set of data supersets

  This allows the astronomer to view the spacial distribution of the images, and provides the information required to decide if the set is evenly distributed spatially.

- Select a set of data supersets based on spatial density

  This allows the astronomer to reduce the number of candidate images by randomly removing some images from the set, from areas where there is a high spacial density of images.

OS 6 Cross archive search

In this scenario, the astronomer is looking at K-band morphology of galaxies with redshifts > .5 and which are bright in the sub-mm wavelengths. In order to do this, K-band images containing candidate galaxies must be retrieved from the GSA.

1. Using a cross archive interface, select images with exposure times greater than 600 seconds, with a K-band filter, and containing galaxies with redshift > .5.
2. Further qualify the set of images by specifying a minimum acceptable spacial resolution.
3. Further qualify the set of images by specifying that at least one of the galaxies contained on the images coincide with a sub-mm source with sub-mm magnitude > *x*.
4. Retrieve a list of the galaxies on the set of images with redshift > .5, and which coincide with a bright sub-mm source. The retrieved list of galaxies should include galaxy coordinates, name (if available), redshift, and sub-mm magnitude. The list should also identify which of the images in the list produced by step 3 contain each galaxy.
5. Down-load any Gemini images from the GSA. The down-loaded data should be re-calibrated, and any possible stacked associations should be created automatically.
6. The astronomer now has the data required to determine the morphology of the galaxies.

The key requirements of this scenario are:

- Support unification of astronomical catalogues

  This scenario cannot be completed by only examining the content of the GSA. Since data must be qualified on sub-mm target descriptors, derived data from other archives must be available.

- Other astronomy archives must participate in unification efforts

  This scenario is only possible if the unified archive contains sub-mm data, including target information derived from the original data. If this kind of information is not available in the unified archive, then participation of the GSA in the unified archive has little value.

- Automatic re-processing of data should be an option

  This eliminates the need for the astronomer to re-process the data being retrieved. This would be a significant advantage for complex instruments with which the astronomer is not familiar, or if data from several different instruments is being used.

- Select data supersets based on object parameter

  This scenario requires the GSA to find image containing galaxies with redshift > .5. This implies searching based on object parameters be an option.

- The GSA should have an object catalogue

  The requirement on an archive of sub-mm data to provide a catalogue of sources implies a reciprocal requirement on the GSA to provide the same service. This object catalogue would permit searches based on object properties as derived from the Gemini data.

OS 7 Multi instrument and multi archive search

In this scenario, an astronomer has just received data from an HST program for which he is PI. The PI observations use WFPC2 to get I-band images of 10 AGNs. The astronomer needs to find HST, EUVE (Extreme Ultraviolet Explorer), and Gemini images and spectra of these same

AGNs. The astronomer also needs to find images from the GSA containing other AGNs with the same redshift as the 10 primary AGNs.

1. Using the coordinates of the 10 AGNs, retrieve public HST direct imaging (WFPC2, NICMOS, or STIS) of any of these objects.
2. Using the coordinates of the 10 AGNs, retrieve public EUVE (Extreme Ultraviolet Explorer) images of any of these objects.
3. Using the coordinates of the 10 AGNs, locate any Gemini spectra of any of the objects listed in the GSA catalogue and present the list to the astronomer.
4. Select relevant public data supersets for retrieval. An option is selected that will cause retrieved data to be re-processed to calibrate and stack overlapping images wherever possible.
5. Identify relevant proprietary data supersets and note the associated PI. This list will be used to contact the PI and attempt to obtain permission to use the proprietary data.
6. The astronomer retrieves the proprietary data authorized by the PI and/or Gemini. The data is processed and delivered in exactly the same way as the public data request.
7. Using the coordinates of the 10 AGNs, locate any images taken with ALTAIR and NIRI. (maybe qualify minimum PSF & S/N?)
8. Items 4, 5, and 6 are repeated for the new list of data supersets. In this case however, the evaluation of the data supersets includes viewing preview images, and evaluating the S/N ratio for each image.
9. The ALTAIR/NIRI images retrieved in 8 should include a model (as a FITS format image) of the PSF. This is needed to de-convolve or model the light profile.
10. Locate other Gemini images of Quasar/AGN in the same redshift range.

The key requirements associated with this scenario are:

- The archive data must be fully described

  This is necessary to allow selection of subsets of the data supersets based on the values of descriptors or coordinates, and to allow the user to view the descriptors for each data superset.
- The GSA must display data superset descriptors to the user

  This allows the astronomer to select specific data supersets that are relevant to the investigation. If this isn't possible, the astronomer will have to retrieve a superset of the desired data, and then eliminate inappropriate data.

  This will also provide the astronomer with contact information for the PI, in the event that the required data is proprietary.
- Select sets of data supersets based on coordinate

  This allows the astronomer to search for specific objects based on their coordinates. This is necessary in order to complete this scenario.
- Support multi archive searches

  This will allow astronomers to search multiple archives simultaneously. This makes the search easier, however it is still possible to get the same data by searching each archive in turn.
- A description of the instruments must be available

  This allows astronomers to determine how the raw data for a given instrument and instrument mode should be calibrated.
- GSA users may preview data for selected data supersets

  This provides the astronomer with the information required to determine the suitability of individual data supersets for their investigation.

- Automatic re-processing of data should be an option

  This eliminates the need for the astronomer to re-process the data being retrieved. This would be a significant advantage for complex instruments with which the astronomer is not familiar.
- Information about re-processing recipes should be available

  This allows the astronomer to determine if the automatic re-processing procedures are appropriate for their investigation.
- Select data supersets based on object parameter

  This allows the astronomer to select data supersets which are appropriate for the investigation. In this scenario, it is assumed that the object parameters are taken from the NED database.
- Calibration data must be easily available

  This allows the astronomer to locate and retrieve the PSF images associated with the data superset.
- Retrieve proprietary data

  If any of the data is proprietary, the user should have the option to attempt to get authorization to retrieve the data.

OS 8 Search by proposal

In this scenario, an instructor teaching an astronomy course wants his/her students to gain experience processing astronomical data.

1. The instructor views proposals that have been executed by the Gemini Telescopes looking for proposals related to the subject matter of the course, and where all of the data is public.
2. The instructor selects a set of proposals.
3. The instructor, or the student selects all data associated with a proposal.
4. The un-calibrated data, calibration data, and re-processed data are retrieved.
5. The student proceeds with the processing of the data.
6. The instructor determines the student’s success by comparing the students results with the re-processed data and any results published by the original PI.
7. There is a discrepancy between the results of one of the students, and the process results supplied by the GSA, which the instructor can’t explain. The instructor sends an e-mail message to the GSA asking for clarification of the processing procedure.

The requirements associated with this scenario are:

- A description of the instruments must be available

  This allows the instructor and students to determine how the raw data for a given instrument and instrument mode should be processed.
- Automatic re-processing of data should be an option

  This allows the instructor to get baseline re-processed data to compare with the results achieved by the students.
- Users should be able to choose one or more processing recipes

  This allows the instructor to get baseline re-processed data to compare with the results achieved by the students.
- Information about re-processing recipes should be available

  This allows the instructor to determine which available re-processing would be best suited to the educational purposes.

- Proposal information must be searchable

  This allows the instructor to locate and examine proposals in order to find appropriate data.
- Search for data based on proposal

  This allows the instructor to select all science data associated with a proposal.

  Being able to “point and click” from the list of proposals would make the archive more usable.
- Calibration data must be easily available

  This allows the instructor or student to retrieve all calibration data associated with science data.
- Publications should be linked to the data

  This allows the instructor to locate publications associated with the proposal. The publications may include results which will assist the instructor in evaluating the performance of the student.
- The GSA staff must provide support to GSA users

  This provides a means for the instructor to get additional information about the automatic data processing.

Gemini Science Archive

Chapter 5

User Requirements

1. Introduction

This chapter describes the user requirements for the GSA. These requirements are from the operational scenarios described in Chapter 4. Table 1 on page 51 and Table 2 on page 54 provide a cross reference between the user requirements and the operational scenarios in Chapter 4 to which they could apply. According to the IEEE standards described in [8], these user requirements must be unambiguous, complete, verifiable, consistent, modifiable, traceable, and usable.

Note that some of the requirements listed here may have to be dropped from the GSA implementation, either to keep the GSA within the available budget, because implementation of the requirement within the GSA is impractical, or because the impact of the requirement on Gemini operations is unacceptable. If requirements do have to be dropped, this document will provide the basis for predicting the impact of the requirement on the functionality of the GSA.

These requirements will be used as the basis of the next phase of the development of the GSA, which will be the Functional and Performance Requirements Document.

TABLE 1. Science requirement to operational scenario cross reference

UR1.1 The archive data must be fully described	
All	
UR1.2 There should be meta data describing the data descriptors	
All	
UR1.3 Select sets of the data supersets in the GSA	
All	
UR1.4 Select sets of data supersets based on coordinate	
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 5	Selection of data unbiased by field selection criteria
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
UR1.5 The GSA must display data superset descriptors to the user	
All	
UR1.6 A general description of the GSA must be available to users	
All	
UR1.7 A description of the instruments must be available	
All	

TABLE 1. Science requirement to operational scenario cross reference

UR1.8 A description of Gemini facilities must be available	
All	
UR1.9 GSA users may select sets of data supersets for retrieval	
All	
UR1.10 Calibration data must be easily available	
All	
UR1.11 Retrieve proprietary data	
All	
UR1.12 Select data supersets based on object parameter	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
UR1.13 Display object parameters	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
UR1.14 Proposal information for public data must be available	
OS 8	Search by proposal
UR1.15 Proposal information must be searchable	
OS 8	Search by proposal
UR1.16 Search for data based on proposal	
OS 2	Retrieve proprietary data associated with an known proposal
OS 8	Search by proposal
UR1.17 Support unification of astronomical catalogues	
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
UR1.18 Reject data supersets based on field selection criteria	
OS 5	Selection of data unbiased by field selection criteria
OS 6	Cross archive search
UR1.19 Data requested by users should be available quickly	
OS 1	Search by position
UR1.20 The GSA should be available 24 hours a day	
OS 1	Search by position
UR1.21 The GSA should have an object catalogue	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 5	Selection of data unbiased by field selection criteria
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
OS 8	Search by proposal

TABLE 1. Science requirement to operational scenario cross reference

UR1.22 The GSA staff must provide support to GSA users	
All	
UR2.1 Support multi archive searches	
OS 3	Search by position, qualified by image descriptors
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
UR2.2 Automatic re-processing of data should be an option	
All	
UR2.3 Automatic re-processing inputs and options should be documented	
All	
UR2.4 Users should be able to choose one or more processing recipes	
All	
UR2.5 Information about re-processing recipes should be available	
All	
UR2.6 GSA users may preview data for selected data supersets	
All	
UR2.7 Preview must be display promptly	
All	
UR2.8 Sort the displayed set of data supersets	
All	
UR2.9 GSA catalogue needs descriptors derived from the data	
OS 4	Search by class of object, qualified by data superset and object descriptors
OS 5	Selection of data unbiased by field selection criteria
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search
OS 8	Search by proposal
UR2.10 The catalogue should contain information on associated data	
OS 4	Search by class of object, qualified by data superset and object descriptors
UR2.11 A single query needs to address any or all instruments	
All	
UR2.12 Access to the GSA catalogue must be interactive	
All	
UR2.13 Publications should be linked to the data	
All	
UR2.14 It should be possible to plot the fields of view of a set of data supersets	
OS 5	Selection of data unbiased by field selection criteria
UR2.15 Select a set of data supersets based on spatial density	
OS 5	Selection of data unbiased by field selection criteria

TABLE 1. Science requirement to operational scenario cross reference

UR2.16 Data quality indicator must be one of the data descriptors	
All	
UR2.17 Search based on a list of coordinates	
OS 1	Search by position
OS 3	Search by position, qualified by image descriptors
OS 7	Multi instrument and multi archive search
UR2.18 Monitor progress of data requests	
All	
UR2.19 View and retrieve the science program	
All	
UR2.20 The content of user requests should be documented	
All	
UR2.21 The GSA should support a batch mode interface	
All	
UR2.22 The GSA should be accessible to users with poor internet connections	
All	
UR3.1 Other astronomy archives must participate in unification efforts	
OS 6	Cross archive search
OS 7	Multi instrument and multi archive search

TABLE 2. Operational scenario to science requirement cross reference

OS 1 Search by position	
UR1.1	The archive data must be fully described
UR1.2	There should be meta data describing the data descriptors
UR1.3	Select sets of the data supersets in the GSA
UR1.4	Select sets of data supersets based on coordinate
UR1.5	The GSA must display data superset descriptors to the user
UR1.6	A general description of the GSA must be available to users
UR1.7	A description of the instruments must be available
UR1.8	A description of Gemini facilities must be available
UR1.9	GSA users may select sets of data supersets for retrieval
UR1.10	Calibration data must be easily available
UR1.11	Retrieve proprietary data
UR1.19	Data requested by users should be available quickly
UR1.20	The GSA should be available 24 hours a day
UR1.22	The GSA staff must provide support to GSA users
UR2.2	Automatic re-processing of data should be an option
UR2.3	Automatic re-processing inputs and options should be documented
UR2.4	Users should be able to choose one or more processing recipes

TABLE 2. Operational scenario to science requirement cross reference

	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets
	UR2.7	Preview must be display promptly
	UR2.8	Sort the displayed set of data supersets
	UR2.11	A single query needs to address any or all instruments
	UR2.12	Access to the GSA catalogue must be interactive
	UR2.13	Publications should be linked to the data
	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.17	Search based on a list of coordinates
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program
	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections
OS 2	**Retrieve proprietary data associated with an known proposal**	
	UR1.1	The archive data must be fully described
	UR1.2	There should be meta data describing the data descriptors
	UR1.3	Select sets of the data supersets in the GSA
	UR1.5	The GSA must display data superset descriptors to the user
	UR1.6	A general description of the GSA must be available to users
	UR1.7	A description of the instruments must be available
	UR1.8	A description of Gemini facilities must be available
	UR1.9	GSA users may select sets of data supersets for retrieval
	UR1.10	Calibration data must be easily available
	UR1.11	Retrieve proprietary data
	UR1.16	Search for data based on proposal
	UR1.22	The GSA staff must provide support to GSA users
	UR2.2	Automatic re-processing of data should be an option
	UR2.3	Automatic re-processing inputs and options should be documented
	UR2.4	Users should be able to choose one or more processing recipes
	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets
	UR2.7	Preview must be display promptly
	UR2.8	Sort the displayed set of data supersets
	UR2.11	A single query needs to address any or all instruments
	UR2.12	Access to the GSA catalogue must be interactive
	UR2.13	Publications should be linked to the data
	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program

TABLE 2. Operational scenario to science requirement cross reference

	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections
OS 3	**Search by position, qualified by image descriptors**	
	UR1.1	The archive data must be fully described
	UR1.2	There should be meta data describing the data descriptors
	UR1.3	Select sets of the data supersets in the GSA
	UR1.4	Select sets of data supersets based on coordinate
	UR1.5	The GSA must display data superset descriptors to the user
	UR1.6	A general description of the GSA must be available to users
	UR1.7	A description of the instruments must be available
	UR1.8	A description of Gemini facilities must be available
	UR1.9	GSA users may select sets of data supersets for retrieval
	UR1.10	Calibration data must be easily available
	UR1.11	Retrieve proprietary data
	UR1.22	The GSA staff must provide support to GSA users
	UR2.1	Support multi archive searches
	UR2.2	Automatic re-processing of data should be an option
	UR2.3	Automatic re-processing inputs and options should be documented
	UR2.4	Users should be able to choose one or more processing recipes
	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets
	UR2.7	Preview must be display promptly
	UR2.8	Sort the displayed set of data supersets
	UR2.11	A single query needs to address any or all instruments
	UR2.12	Access to the GSA catalogue must be interactive
	UR2.13	Publications should be linked to the data
	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.17	Search based on a list of coordinates
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program
	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections
OS 4	**Search by class of object, qualified by data superset and object descriptors**	
	UR1.1	The archive data must be fully described
	UR1.2	There should be meta data describing the data descriptors
	UR1.3	Select sets of the data supersets in the GSA
	UR1.4	Select sets of data supersets based on coordinate
	UR1.5	The GSA must display data superset descriptors to the user
	UR1.6	A general description of the GSA must be available to users
	UR1.7	A description of the instruments must be available

TABLE 2. Operational scenario to science requirement cross reference

	UR1.8	A description of Gemini facilities must be available
	UR1.9	GSA users may select sets of data supersets for retrieval
	UR1.10	Calibration data must be easily available
	UR1.11	Retrieve proprietary data
	UR1.12	Select data supersets based on object parameter
	UR1.13	Display object parameters
	UR1.21	The GSA should have an object catalogue
	UR1.22	The GSA staff must provide support to GSA users
	UR2.2	Automatic re-processing of data should be an option
	UR2.3	Automatic re-processing inputs and options should be documented
	UR2.4	Users should be able to choose one or more processing recipes
	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets
	UR2.7	Preview must be display promptly
	UR2.8	Sort the displayed set of data supersets
	UR2.9	GSA catalogue needs descriptors derived from the data
	UR2.10	The catalogue should contain information on associated data
	UR2.11	A single query needs to address any or all instruments
	UR2.12	Access to the GSA catalogue must be interactive
	UR2.13	Publications should be linked to the data
	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program
	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections
OS 5	**Selection of data unbiased by field selection criteria**	
	UR1.1	The archive data must be fully described
	UR1.2	There should be meta data describing the data descriptors
	UR1.3	Select sets of the data supersets in the GSA
	UR1.4	Select sets of data supersets based on coordinate
	UR1.5	The GSA must display data superset descriptors to the user
	UR1.6	A general description of the GSA must be available to users
	UR1.7	A description of the instruments must be available
	UR1.8	A description of Gemini facilities must be available
	UR1.9	GSA users may select sets of data supersets for retrieval
	UR1.10	Calibration data must be easily available
	UR1.11	Retrieve proprietary data
	UR1.18	Reject data supersets based on field selection criteria
	UR1.21	The GSA should have an object catalogue
	UR1.22	The GSA staff must provide support to GSA users
	UR2.2	Automatic re-processing of data should be an option

TABLE 2. Operational scenario to science requirement cross reference

	UR2.3	Automatic re-processing inputs and options should be documented
	UR2.4	Users should be able to choose one or more processing recipes
	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets
	UR2.7	Preview must be display promptly
	UR2.8	Sort the displayed set of data supersets
	UR2.9	GSA catalogue needs descriptors derived from the data
	UR2.11	A single query needs to address any or all instruments
	UR2.12	Access to the GSA catalogue must be interactive
	UR2.13	Publications should be linked to the data
	UR2.14	It should be possible to plot the fields of view of a set of data supersets
	UR2.15	Select a set of data supersets based on spatial density
	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program
	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections
OS 6	**Cross archive search**	
	UR1.1	The archive data must be fully described
	UR1.2	There should be meta data describing the data descriptors
	UR1.3	Select sets of the data supersets in the GSA
	UR1.4	Select sets of data supersets based on coordinate
	UR1.5	The GSA must display data superset descriptors to the user
	UR1.6	A general description of the GSA must be available to users
	UR1.7	A description of the instruments must be available
	UR1.8	A description of Gemini facilities must be available
	UR1.9	GSA users may select sets of data supersets for retrieval
	UR1.10	Calibration data must be easily available
	UR1.11	Retrieve proprietary data
	UR1.12	Select data supersets based on object parameter
	UR1.13	Display object parameters
	UR1.17	Support unification of astronomical catalogues
	UR1.18	Reject data supersets based on field selection criteria
	UR1.21	The GSA should have an object catalogue
	UR1.22	The GSA staff must provide support to GSA users
	UR2.1	Support multi archive searches
	UR2.2	Automatic re-processing of data should be an option
	UR2.3	Automatic re-processing inputs and options should be documented
	UR2.4	Users should be able to choose one or more processing recipes
	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets

TABLE 2. Operational scenario to science requirement cross reference

UR2.7	Preview must be display promptly
UR2.8	Sort the displayed set of data supersets
UR2.9	GSA catalogue needs descriptors derived from the data
UR2.11	A single query needs to address any or all instruments
UR2.12	Access to the GSA catalogue must be interactive
UR2.13	Publications should be linked to the data
UR2.16	Data quality indicator must be one of the data descriptors
UR2.18	Monitor progress of data requests
UR2.19	View and retrieve the science program
UR2.20	The content of user requests should be documented
UR2.21	The GSA should support a batch mode interface
UR2.22	The GSA should be accessible to users with poor internet connections
UR3.1	Other astronomy archives must participate in unification efforts
OS 7	**Multi instrument and multi archive search**
UR1.1	The archive data must be fully described
UR1.2	There should be meta data describing the data descriptors
UR1.3	Select sets of the data supersets in the GSA
UR1.4	Select sets of data supersets based on coordinate
UR1.5	The GSA must display data superset descriptors to the user
UR1.6	A general description of the GSA must be available to users
UR1.7	A description of the instruments must be available
UR1.8	A description of Gemini facilities must be available
UR1.9	GSA users may select sets of data supersets for retrieval
UR1.10	Calibration data must be easily available
UR1.11	Retrieve proprietary data
UR1.12	Select data supersets based on object parameter
UR1.13	Display object parameters
UR1.17	Support unification of astronomical catalogues
UR1.21	The GSA should have an object catalogue
UR1.22	The GSA staff must provide support to GSA users
UR2.1	Support multi archive searches
UR2.2	Automatic re-processing of data should be an option
UR2.3	Automatic re-processing inputs and options should be documented
UR2.4	Users should be able to choose one or more processing recipes
UR2.5	Information about re-processing recipes should be available
UR2.6	GSA users may preview data for selected data supersets
UR2.7	Preview must be display promptly
UR2.8	Sort the displayed set of data supersets
UR2.9	GSA catalogue needs descriptors derived from the data
UR2.11	A single query needs to address any or all instruments
UR2.12	Access to the GSA catalogue must be interactive
UR2.13	Publications should be linked to the data

TABLE 2. Operational scenario to science requirement cross reference

	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.17	Search based on a list of coordinates
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program
	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections
	UR3.1	Other astronomy archives must participate in unification efforts
OS 8	**Search by proposal**	
	UR1.1	The archive data must be fully described
	UR1.2	There should be meta data describing the data descriptors
	UR1.3	Select sets of the data supersets in the GSA
	UR1.5	The GSA must display data superset descriptors to the user
	UR1.6	A general description of the GSA must be available to users
	UR1.7	A description of the instruments must be available
	UR1.8	A description of Gemini facilities must be available
	UR1.9	GSA users may select sets of data supersets for retrieval
	UR1.10	Calibration data must be easily available
	UR1.11	Retrieve proprietary data
	UR1.14	Proposal information for public data must be available
	UR1.15	Proposal information must be searchable
	UR1.16	Search for data based on proposal
	UR1.21	The GSA should have an object catalogue
	UR1.22	The GSA staff must provide support to GSA users
	UR2.2	Automatic re-processing of data should be an option
	UR2.3	Automatic re-processing inputs and options should be documented
	UR2.4	Users should be able to choose one or more processing recipes
	UR2.5	Information about re-processing recipes should be available
	UR2.6	GSA users may preview data for selected data supersets
	UR2.7	Preview must be display promptly
	UR2.8	Sort the displayed set of data supersets
	UR2.9	GSA catalogue needs descriptors derived from the data
	UR2.11	A single query needs to address any or all instruments
	UR2.12	Access to the GSA catalogue must be interactive
	UR2.13	Publications should be linked to the data
	UR2.16	Data quality indicator must be one of the data descriptors
	UR2.18	Monitor progress of data requests
	UR2.19	View and retrieve the science program
	UR2.20	The content of user requests should be documented
	UR2.21	The GSA should support a batch mode interface
	UR2.22	The GSA should be accessible to users with poor internet connections

2. User Requirements for a Gemini Science Archive

The requirements have been placed into three different subsections. Section 2.1 on page 61 contains all requirements which if not implemented will make at least once science case impossible. Section 2.2 on page 65 contains requirements which will make the GSA more usable. Section 2.3 on page 68 contains requirements which are beyond the scope of the Gemini project and the GSA, but which would benefit the GSA.

The requirements which have implications for Gemini operations have an additional section describing how Gemini operations must support the requirement.

2.1 Functional Requirements

This section defines the functional requirements which the Gemini Science Archive must have in order to support all of the science cases listed in Chapter 3. If some of these requirements are not met, then some of the operational scenarios described in Chapter 4 will not be possible, and therefore some of the science cases in Chapter 3 will not be possible. This section also includes requirements that if not met, would result in some scenarios being so difficult that it is unlikely that a GSA user would attempt them.

UR1.1

Short Description: The archive data must be fully described

Description: All descriptors necessary to qualify and to quantify the data must exist, be accessible, and be accurate. These descriptors include a record of the instrument mode and configuration, complete optical path including any Gemini facilities which might affect the data (e.g. ALTAIR), observatory systems status, environmental conditions (e.g, logs of weather and instrument temperatures), accurate WCS information describing the field of view, wavelength being observed, and proposal information (there should be a link between data and the proposal information described in UR1.14). For spectral data, this should also include spectral resolution.

Requirement on Gemini: This requirement implies that Gemini will gather all of the information listed above, and make the data available to the GSA.

UR1.2

Short Description: There should be meta data describing the data descriptors

Description: The data descriptors should themselves be described, allowing GSA users to determine the exact meaning of the descriptor and what it implies for the data. This is essentially online help for the descriptors.

Requirement on Gemini: This requirement implies that Gemini will make descriptions of data descriptors available to the GSA, and notify the GSA when data descriptors change.

UR1.3

Short Description: Select sets of the data supersets in the GSA

Description: The GSA must supply tools that allow GSA users to select a pertinent set of data supersets stored in the GSA catalogue by specifying a value, a set of values, or a range of values for one or more of the descriptors for a data superset. This selection should work in conjunction with UR1.4, UR1.12, and UR1.17 to allow an intersection of the sets to be created.

UR1.4

Short Description: Select sets of data supersets based on coordinate

Description: The GSA must be able to select a set of available data supersets controlled by whether or not a specified region overlaps the field of view. This is somewhat different than UR1.3 in that the search must use the idea of “overlapping regions”, rather than the simple equality/set of values/range search specified in UR1.3. The region should be specified by a position in ra. and dec., and a radius in seconds of arc. This must work in conjunction with UR1.3, UR1.12, and UR1.17 to allow a intersection of the sets to be created.

Requirement on Gemini: Gemini must provide enough information to accurately determine the field of view for each dataset and frame.

UR1.5 **Short Description:** The GSA must display data superset descriptors to the user

Description: After the user has selected a set of the data supersets stored in the GSA, the GSA must be able to display all descriptors for the set to the user.

UR1.6 **Short Description:** A general description of the GSA must be available to users

Description: In order to determine if the GSA might contain data appropriate to a given investigation, a brief description of the GSA content should be available to users. This should include the wavelength coverage of the archive, and the types of instruments included in the archive. Note that this is somewhat different than the descriptions that might be available for current Gemini Telescope operations in that this description must include historical capabilities of the telescopes which may not be present in the descriptions of the current operational telescopes.

Requirement on Gemini: Gemini will have to make the information available to the GSA, and must provide updates as telescope capabilities change.

UR1.7 **Short Description:** A description of the instruments must be available

Description: A description of each of the instruments used to collect the data in the archive must be available to users. This must include instrument capabilities, limitations, operating modes, and calibration procedures. Note that this is slightly different than the descriptions that might be available for the current telescope operations in that obsolete instruments and historical information about evolving instruments must also be included.

Requirement on Gemini: Gemini will have to make the instrument descriptions available to the GSA, and must provide updates as instruments evolve, and as new instruments are introduced.

UR1.8 **Short Description:** A description of Gemini facilities must be available

Description: In order to allow the user to evaluate and process data, a description of any Gemini facilities that may affect the data (e.g. ALTAIR) must be available. Note that this is slightly different than the descriptions that might be available for the current telescope operations in that obsolete facilities and historical information about evolving facilities must also be included.

Requirement on Gemini: Gemini must supply the descriptions of the facilities.

UR1.9 **Short Description:** GSA users may select sets of data supersets for retrieval

Description: Once a set of the GSA data supersets have been selected as described in UR1.3, UR1.4, UR1.12, and UR1.17, the GSA user must have the ability to either select individual data supersets or the entire set for retrieval. The retrieved data may be either raw data, or the re-processed data described in UR2.2. The retrieval method may be either via the internet, (e.g. FTP), or via physical media (e.g. EXABYTE or CD-ROM). Note that the delivery media supported by the GSA operations must evolve to stay current with media in common use by the astronomical community. This evolution must be supported by both the GSA design, and the GSA operations.

Requirement on Gemini: This implies that Gemini will make raw data available to the GSA.

UR1.10 **Short Description:** Calibration data must be easily available

Description: When retrieving science data, the user should have the ability to retrieve all calibration data associated with the science data.

Requirement on Gemini: This implies that Gemini is supplying enough information about the science and calibration data to allow the GSA to associate science data with appropriate calibration data. It also implies that Gemini is collecting a reasonable set of calibration data for all science data, even in cases where the PI does not require the calibration data.

UR1.11 **Short Description:** Retrieve proprietary data

Description: A GSA user should have a way of attempting to get permission to access proprietary data. The mechanism for getting permission to access proprietary data is TBD, but should require getting permission from the PI, and possibly some entity within the Gemini organization.

Requirement on Gemini: Gemini must have a policy detailing how and when archive users can gain access to proprietary data, and if the implementation of the policy requires the active participation of Gemini, Gemini must provide support for the implementation.

UR1.12 **Short Description:** Select data supersets based on object parameter

Description: It must be possible to select a set of data supersets by specifying a value, range of values, or set of values for parameters describing objects in the field of view. These parameters include object class, limiting magnitude, distance, reddening parameter, and redshift. These object parameters supplied by PIs in the proposal, they may be derived from the data, or they may be extracted from other catalogues (e.g. SIMBAD, NED, EMSS, etc.). Note that the object parameters should reflect the content of the other catalogues at or near the time of the query, which may be different from the catalogue content at the time the data became part of the GSA catalogues. This selection should work in conjunction with UR1.3, UR1.4, and UR1.17 to allow an intersection of the sets to be created.

Requirement on Gemini: If target parameters are extracted from other catalogues, it implies that the WCS coordinates supplied by Gemini are accurate enough to determine which objects are within the field of view. Gemini should ask each PI to supply an object class for the principal target in an image, where the class is chosen from a Gemini supplied list of object classes.

UR1.13 **Short Description:** Display object parameters

Description: It must be possible to view parameters associated with objects in the field of view of an data superset. These parameters are described in UR1.12.

Requirement on Gemini: See UR1.12.

UR1.14 **Short Description:** Proposal information for public data must be available

Description: After data become public, information about Gemini proposals should be incorporated into the GSA database and must be viewable by GSA users. The release date of the proposal information must follow Gemini policy.

Requirement on Gemini: Gemini must have a policy on releasing proposals to the public. Gemini must provide proposal information to the GSA, preferably in electronic form.

UR1.15 **Short Description:** Proposal information must be searchable

Description: It must be possible to find Gemini proposals by searching on science category, keyword, or some other similar mechanism.

Requirement on Gemini: Science category or science keyword must be available in the proposal information collected by Gemini and supplied to the GSA. To ease the operational costs, it must be possible to electronically parse these values from the data supplied to the GSA.

UR1.16 **Short Description:** Search for data based on proposal

Description: It must be possible to locate data associated with a given proposal given the proposal id, and also possibly also with a point and click interface from the proposal information display required in UR1.14.

Requirement on Gemini: Gemini must provide the GSA with enough information to associate every dataset with a proposal.

UR1.17 **Short Description:** Support unification of astronomical catalogues

Description: It should be possible to search for data from the GSA based on the availability or content of data from other catalogues. In other words, it will be possible for a GSA user to select data supersets from the GSA catalogue, where there exists an entry in another catalogue which has some properties. This selection should work in conjunction with UR1.3, UR1.4, and UR1.12 to allow an intersection of the sets to be created.

While this requirement is beyond the scope of the GSA development or operations, it is possible for the GSA to be an important part of a collaborative effort (see also UR3.1) to achieve this goal. As an ongoing task the GSA should encourage, and have resources to participate in efforts to create unified astronomy catalogues.

UR1.18 **Short Description:** Reject data supersets based on field selection criteria

Description: Some science cases driving the GSA require that an unbiased sample of data supersets be selected. In order to do this, it must be possible to reject data supersets based on classes of objects which are intentionally included on a field, and also on classes of objects which are intentionally excluded from a field.

Requirement on Gemini: The field selection criteria must be supplied by the astronomer designing the observing program, collected by Gemini (as part of the phase II proposal preparation?), and provided to the GSA.

Note that while this information is very useful to the GSA, it depends on complete and accurate but difficult to verify information being provided by all of the astronomers designing observing programs. The difficulty in ensuring completeness and correctness may make this requirement impossible to achieve in a useful way.

UR1.19 **Short Description:** Data requested by users should be available quickly

Description: When a user requests data to be delivered over the Internet, the data should available to the user promptly. The elapsed time between the time the user submits a request for raw data and the time the user is notified that the data is available for down-load should be less than 5 minutes + 10 seconds per Mbyte of data requested, 80% of the time. It would be desirable to put a performance requirement on retrieving re-processed data, but the difficulty in quantifying the time it will take to do the re-processing makes this difficult.

UR1.20 **Short Description:** The GSA should be available 24 hours a day

Description: To support researchers in different time zones, and night-time usage by observers at various telescope sites, the GSA catalogue and data retrieval should be available 24 hours a day. The nominal availability of the GSA should be better than 95%.

UR1.21 **Short Description:** The GSA should have an object catalogue

Description: The GSA should have a catalogue of objects derived from the observation data. There should be automatic data processing in place to create the object catalogue from the raw data, and derive as many object descriptors as is practical from the data.

Requirement on Gemini: See UR2.2.

UR1.22 **Short Description:** The GSA staff must provide support to GSA users

Description: While interaction with GSA staff should not be required for normal use of the GSA, GSA users must have a method for obtaining assistance from the GSA staff within a reasonable period of time. Available assistance should include software support for the catalogue interface, automated data processing software, and data retrieval. Assistance should also be available from an astronomer knowledgable about the content of the GSA. Typical questions asked by CADC archive users are:

- "I've lost my password."
- "I requested data a while ago, but haven't received a message notifying me that is complete yet" (usually their e-mail address has changed).
- "The processed data I requested didn't have step *x* applied, why not?"
- "The processed image I requested is blank, why?"
- "I've lost my password again."

Requirement on Gemini: In rare cases where questions cannot be answered by GSA staff, it may be necessary to forward some questions to Gemini staff, or to the staff of the Gemini project office of one of the partner countries.

2.2 Usability Requirements

This section lists requirements which will increase the productivity of a user of the GSA. If these requirements are not met, it will be still be possible to do all of the scenarios in Chapter 4, however the GSA will be more difficult to use, which may cause the GSA to be used less often, reducing the overall benefit derived from the archive.

UR2.1 **Short Description:** Support multi archive searches

Description: The GSA should support simultaneous searches of multiple telescope archives, including both archives stored at the archive centre hosting the GSA, and other archives throughout the world. This differs from UR1.17 in that it only implies searching multiple archives, but not the cross referencing of data from multiple archives. Like UR1.17, this requirement is beyond the scope of the GSA, but could be supported by the same unified catalogue referred to in UR1.17.

UR2.2 **Short Description:** Automatic re-processing of data should be an option

Description: The necessary elements must be in place to enable calibration by an automated processing pipeline. These elements include the existence of reliable calibration material for **all** science data (whether queue-mode or classical) and pipeline processing software. Calibration of science data may be performed at the time of the observation but the capability must exist to re-calibrate all data at any future time to take advantage of increased knowledge of the instruments and improved calibration material.

Requirement on Gemini: This implies that sufficient information is supplied with all science data to allow appropriate calibration files to be identified, and to allow appropriate processing steps to be determined without any intervention from archive users or staff. Creation of these recipes will also require the cooperation of Gemini instrument scientists, although the recipes will probably be very similar to the recipe used by the Gemini data processing system during telescope operations.

UR2.3 **Short Description:** Automatic re-processing inputs and options should be documented

Description: Descriptions of the specific inputs and options passed to a recipe when doing automatic re-processing should be provided to a GSA user requesting re-processed data. The information should be in the form of a log file associated with the processed data.

UR2.4 **Short Description:** Users should be able to choose one or more processing recipes

Description: The archive user should be able to select from a set of possible processing recipes if more than one exists for a given data superset. For example, stacking of images which overlap may be desirable for some purposes but not for others. Users should have the option to select re-calibration with or without stacking.

UR2.5 **Short Description:** Information about re-processing recipes should be available

Description: Descriptions of the re-processing recipes described in UR2.2 and UR2.4 should be available to the archive users. These descriptions should be complete enough to allow a user to determine if the processing is appropriate for their purposes.

Requirement on Gemini: Since many of the processing recipes will be based on those used for Gemini, descriptions of the Gemini processing recipes will have to be supplied to the GSA.

Requirement on Gemini: See UR2.2.

UR2.6 **Short Description:** GSA users may preview data for selected data supersets

Description: Each public data superset in the GSA will have a “preview” available. This preview will allow the user to visually examine the data in order to determine its suitability. The preview image may be of lower quality in order to allow for reasonable storage space, Internet transfer, and display time. Note that the usefulness of this preview depends on automatic re-calibration as described in UR2.2.

UR2.7 **Short Description:** Preview must be display promptly

Description: Preview image should be displayed within 10 seconds 80% of the time, using the typical internet connection between the CADC in Victoria and the Gemini North base facility in Hilo as the baseline.

UR2.8 **Short Description:** Sort the displayed set of data supersets

Description: Users must be able to sort the displayed set of data supersets based on the data descriptors described in UR1.1, or based on the object descriptors described in UR1.12.

UR2.9 **Short Description:** GSA catalogue needs descriptors derived from the data

Description: The catalogue should contain descriptors derived from the raw data, for example: signal to noise ratio, signal to noise ratio of associated data supersets, and spectral coverage. This implies automatic pipeline processing of data to calculate the descriptors. Users should be able to search for and display the descriptors as described in UR1.3 and UR1.5. These derived descriptors are assumed to be subject to the same proprietary period as the raw data.

UR2.10 **Short Description:** The catalogue should contain information on associated data

Description: The GSA will find possible “associations” of data. Associations are groups of datasets which can be merged to form a single data superset. The associations may be based on overlap of image field of view, allowing creation of a single, stacked image with a higher S/N ratio. Associations may also be other scientifically meaningful groupings of datasets identified by GSA astronomers, Gemini astronomers, or GSA users.

Requirement on Gemini: In order to associate images, the WCS information associated with each frame must be accurate enough to determine how various data overlap.

UR2.11 **Short Description:** A single query needs to address any or all instruments

Description: The GSA will be considerably less usable if a user must interact with the archive separately for each instrument in the set of instruments.

UR2.12 **Short Description:** Access to the GSA catalogue must be interactive

Description: All of the scenarios described in Chapter 4 describe an interactive use of the GSA. While a non-interactive (batch mode) interface to the GSA would allow all of the scenarios to be completed, an interactive interface will make the GSA much more user friendly.

UR2.13 **Short Description:** Publications should be linked to the data

Description: Publications based on Gemini data should be linked to the proposals for which the data was collected. The list of publications should be available through the proposal searching interface described in UR1.15. This list should include both publications by the PI, and publications based on archive data.

Requirement on Gemini: This requires that Gemini will acquire data on publications based on data collected at the Gemini Telescopes, including the proposal associated with the published data.

UR2.14 **Short Description:** It should be possible to plot the fields of view of a set of data supersets

Description: It should be possible to plot the fields of view a set of data supersets from the GSA in either right ascension and declination, or in galactic coordinates.

UR2.15 **Short Description:** Select a set of data supersets based on spatial density

Description: Given the field of view position plot described in UR2.14, it should be possible to automatically select images from the set, to create a new set with a more uniform spacial distribution. In other words, the new set should consist of all data supersets from regions of the sky where there are few data supersets, but only a random subsample of the data supersets from regions of the sky where there are many data supersets.

UR2.16 **Short Description:** Data quality indicator must be one of the data descriptors

Description: An indication of the data quality of each dataset and data superset should be one of the descriptors. The data quality indicator should be based on evaluation of the data by Gemini staff, and should provide a coarse indication of the usefulness of the data (e.g. “meets Gemini quality criteria, but doesn’t meet the requirements for the observing program”, or “meets both Gemini and observing program quality requirements”).

Requirement on Gemini: Gemini staff must evaluate each dataset and assign a data quality value.

UR2.17 **Short Description:** Search based on a list of coordinates

Description: It should be possible to search for data supersets where the field of view overlaps any one of a list of possible regions. This search should be identical the search described in UR1.4, except that any number of coordinates can be entered.

UR2.18 **Short Description:** Monitor progress of data requests

Description: It should be possible for users to monitor the progress of data requests. Enough information should be provided to allow a user to determine when the requested data is likely to be available, and if any problems have occurred during the processing of the request.

UR2.19 **Short Description:** View and retrieve the science program

Description: A GSA user should be able to view and retrieve the science program used to create a dataset. The science program name/id should be viewable from the GSA catalogue display described in UR1.5, and it should be possible to retrieve the science programs when retrieving data from the GSA.

Requirement on Gemini: If the GSA is to display the science program, it must have a display tool to do so. Since Gemini already has the Observing Tool (OT) which displays science programs, it would save the GSA effort to either re-use the OT in its entirety, or some of its components in the GSA display. This would require that the OT be able to re-load an science program supplied by the GSA.

The GSA also has a requirement to display all science programs stored in the GSA, and if the OT is used as the science program display, it must either maintain backward compatibility whenever the format of the science programs changes, or there must be utilities to transform old style science programs into the current format.

UR2.20 **Short Description:** The content of user requests should be documented

Description: Data requests should contain information describing how the data supersets in the request were selected. This will allow archive users to remind themselves of the purpose of requests, particularly in cases where a user submits multiple requests.

UR2.21 **Short Description:** The GSA should support a batch mode interface

Description: Some queries to the GSA catalogue may take a long time to process. To make these queries more practical for GSA users, the GSA interface should support a batch mode which does not require the GSA user to remain connected to the GSA while waiting for the query to complete.

UR2.22 **Short Description:** The GSA should be accessible to users with poor internet connections

Description: The GSA should provide support to users who do not have reliable, high speed access to the GSA archive centre. This support may be provided by ensuring that the primary interface to the GSA does not depend on good internet connectivity, or by providing an alternate, secondary interface which does not depend on good internet connectivity.

2.3 External Requirements

These requirements are external to the Gemini systems. While neither GSA nor Gemini can directly implement these requirements, they can and should attempt to encourage their development whenever possible.

UR3.1 **Short Description:** Other astronomy archives must participate in unification efforts

Description: Any participation by the GSA in a astronomical catalogue unification effort (see UR1.17) must be accompanied by participation by other astronomical archives. Without significant participation from other archives, participation of the GSA has little value.

Gemini Science Archive

Chapter 6
Review Comments and Responses

1. Introduction

These are the reviewers comments and our responses for the initial version of the OCDD.

2. Reviewer comments

2.1 Fred Gillett

2.1.1 **Comment:** We must have a single set of definitions of terms between Gemini and the GSA. We must agree on things like “observation”, “dataset”, “science program”, “image quality” and the like. Right now it looks like we are quite far apart in some areas.

Reply: We will change our definitions to match Gemini’s. See comment 2.1.15.

2.1.2 **Comment:** the Initial OCDD should clearly identify those requirements that basic GSA requirements from the work scope and those that are different levels of “advanced requirements” as defined in the work scope. If there are other requirements that you have identified that go beyond the work scope advanced requirements, those should be separately identified as well. The basic requirements are much more straightforward and can be represented by quite simple science scenarios, like long term preservation of data, an observer accessing his proprietary data, a user wanting to know what observations have been made at a particular ra and dec, are they proprietary or not, and being able to retrieve previews and the available data, calibration frames, processed data, etc. The initial OCDD and particularly the initial FPRD have to clearly differentiate between what are the basic requirements and what are different levels of advanced requirements. All your scenarios appear to be focused on advanced capabilities, although in many cases, basic capabilities are embedded with advanced capabilities.

I would suggest that you consider adding a section that outlines and identifies the basic scenarios and how they link to the basic requirements and using different identifiers to discriminate between basic and advanced capabilities.

Reply: The differentiation between different levels of requirements is done in the FPRD.

We have created simplified scenarios.

The linking of scenarios to requirements is done in chapter 5.

2.1.3 **Comment:** We discussed quite extensively in Victoria, what information the GSA will get from Gemini, and I thought we were in general agreement that basically the GSA will get the same data package that a queue user gets, with all the necessary information in fits images, headers and attachments, this includes the necessary identifiers, environmental info, processing, info, descriptions with pointers, etc. The concept presented in Chapter 2 Section 7. is not correct and needs to be changed. We have to establish a simple interface between Gemini and the GSA, and I would think that it could be described by a single ICD identifying deliverables and formats. The GSA will not be interfacing directly to the variety of Gemini subsystems identified in Chapter 2 Section 7.

Reply: We will modify Chapter 2 Section 7. to indicate that all communication is through the Gemini - GSA ICD. The type of communication between Gemini and the GSA shouldn't be restricted at this stage of the development, unless Gemini wants to limit the communications as a new requirement, and this limitation may have an impact on the cost and functionality of the GSA. The exact nature of the communication between the GSA and Gemini will be determined during the development of the Gemini - GSA ICD, which will include substantial input from Gemini staff regarding how the Gemini side of the interface will work.

2.1.4 **Comment:** The Mission statement in section Chapter 2 Section 7. is not appropriate. The statement as taken from "The Scientific Case for a Gemini Data Archive" was not adopted by Gemini. My suggestion for a slightly modified version, basically GSC resolution 12.9 is the following:

The Gemini Science Archive should provide the scientific community with tools for effective online access to all Gemini science data and supporting information in order to promote further scientific exploitation of those data. The Gemini Science Archive should guarantee that the valuable datasets obtained with the Gemini Telescopes are usable by future generations for research and education.

Reply: The text has been modified.

2.1.5 **Comment:** I assume the definition [of dataset in the glossary] is consistent with the Gemini Definitions?

Reply: This definition is from ICD3, which we believe is still current.

2.1.6 **Comment:** [The definition of descriptors in the glossary] should not include weather data as a descriptor. All necessary descriptors are in headers or as extensions.

Reply: Unless Gemini already has a definition for "descriptors", we would argue that the definition should include all information which describes a dataset, including the weather conditions, science program, etc. This also implies that not all descriptors will be included in a dataset's FITS file.

2.1.7 **Comment:** Need definition of FITS extension.

Reply: The definition has been added.

2.1.8 **Comment:** The definition of Gemini Science archive in the glossary, should be reworded as: An archive system intended to *preserve scientific* data collected by the Gemini Telescopes *and make it* available to *and useable by* the general astronomical community, after the proprietary period for the data has expired.

Reply: Ok, unless the phrase "preserve scientific data" implies the GSA will be an independent safe store for Gemini data, including duplicate off site storage. We have assumed that Gemini, and possibly a second GSA archive site would function as redundant stores.

2.1.9 **Comment:** [Image quality definition] Image quality should mean image quality, i.e. characterization of PSF. Data with no scientific value is not an issue for the GSA. Data with limited scientific value should not be an issue for the GSA.

Reply: We have renamed this to “Data quality” and updated the document.

Having low quality data available from the GSA isn’t a problem, as long as the user knows the data isn’t good, and has the option (possibly the default behaviour) to eliminate poor quality data from their results. Having low quality data is useful in those cases where the user is willing to deal with the problems.

2.1.10 **Comment:** [Observation definition] Need a consistent definition.

Reply: We will replace “observation” with “data superset”. The definition of data superset will be:

Data superset — A data superset is either a single dataset, or an association of two or more related datasets. An associated data superset may include any combination of datasets and associated data supersets. An associated data superset may be a virtual concept, with only a database entry describing how members are combined, or it may be a FITS file, which contains the result of some data processing step (e.g. a mosaiced image created from separate datasets). An associated data superset has meta data which describes whole set of data (total exposure time, total field of view, etc.).

2.1.11 **Comment:** [Planned observing definition] planned observing is strongly urged for classical users.

Reply: Ok.

2.1.12 **Comment:** [Principal investigator definition] ... person who has sole access to the *science archive* data associated ...

Reply: Ok.

2.1.13 **Comment:** [Queue observing definition] Replace “queue” with “list” in the definition.

Reply: Ok.

2.1.14 **Comment:** Need a consistent definition for Science program.

Reply: I have added a definition of science programs, paraphrased from the OT documentation.

2.1.15 **Comment:** [Service observing definition] This is similar to ~~Classical~~ *queue* observing, ~~except~~ *in [the] sense* that the execution of a science program is ~~overseen~~ by a member of the observatory staff...

Reply: Ok.

2.1.16 **Comment:** There is one aspect that will have to be cleaned up once we have defined the Gemini data package. That is the Interface between GSA and other Gemini systems (p. 15 and UR1.1):

- The figure on p. 15 has been redrawn, but the text still states that the GSA interfaces to the other Gemini systems individually rather than through a Gemini run packaging of the data.
- UR1.1 mentions that a strong link between the GSA and GEA is needed.

Much of the information that needs to be associated with the data we plan to include in the FITS headers. These issues will be clarified once we discuss in more detail how Gemini will distribute the data. Inger is currently working on the draft description of Gemini data distribution. This should form the basis of the Gemini-GSA ICD.

Reply: You are correct. We have removed the line from UR1.1 which specified a link between the GSA and the engineering archive.

2.1.17 **Comment:** UR1.12: I think it needs to be mentioned that a list of object classes should be made, and that the PI has to assign each observation an object class. Otherwise a search by object class will not be possible. We will not derive S/N and/or limiting magnitude as part of the data quality assessment. Does the GSA plan to derive it? (See also comment 2.1.19.)

Reply: We were thinking in terms of object classes from SIMBAD and NED. Object classes supplied by the PI do have value not duplicated by object classes from SIMBAD and NED, and so we have added a phrase to UR1.2 saying so. This will add a requirement on Gemini in the FPRD to collect this information in the proposal process and to provide to the GSA.

See response to comment 2.1.19 about S/N and limiting magnitude.

2.1.18 **Comment:** UR1.21: Who will derive the object catalogues? It is not (and will not be) part of the Gemini pipeline reductions to derive object catalogues. Will the GSA write software to derive object catalogues and make these catalogues available through the GSA? In that case, we need to make sure that the catalogues are not derived until the data becomes public.

Reply: Deciding where the information will be created is part of the ICD/design process, but we expect that the object catalogues will be derived by the GSA after data becomes public. Gemini will not be expected to provide the software to do this.

2.1.19 **Comment:** UR2.9: The S/N is listed as a descriptor. I have the same questions for the S/N as for the limiting magnitude - who derives it? We will not routinely derive it as part of the data quality assessment. There may be a few cases where we do derive S/N estimates, but there are no plans to put these estimates in the image headers. The relevant method for deriving the S/N also depends on the data. Let me just list a few examples: S/N of the continuum of a spectrum, S/N of an emission line in a spectrum, S/N within an aperture for imaging data, S/N per square arcsec for some extended object.

Reply: Deciding where the information will be generated is part of the ICD/design process, but we expect that all derived parameters will be generated by the GSA after the data becomes public.

The derived descriptors listed in UR2.9 are examples intended to support the point that the GSA must have the infrastructure required to support automatic derivation of descriptors, and we have changed the text to make that clear. The initial decision about which descriptors will be generated will be made during the GSA design, based on the usefulness of descriptors to GSA users and the practicality of automatically generating the descriptors. We also expect the list of generated descriptors will evolve during the lifetime of the GSA in response to evolution of the data supplied by Gemini, evolution of GSA user needs, and evolution of the processing algorithms available to the GSA.

2.1.20 **Comment:** UR2.16: It is my opinion that data that have no scientific value and data that do not meet Gemini quality criteria should not be available in the GSA.

Reply: We have removed those two examples from UR2.16. The discussion of what is or is not available from the GSA falls out from the ICD discussions which is the next step after FPRD.

2.2 Felipe Barrientos

These comments are for the FPRD, however we feel that they are more appropriate to the OCDD document.

2.2.1 **Comment:** In the document you emphasize the GSA must be interactive and therefore it must have a quick response (although some queries could take a long time). I think the GSA should also have a batch capability, probably along the same line of NED. In any case, if this is not possible the minimum would be to be able to query for a set of different positions in the same session (the HST archive at STSCI has that capability now and it is very nice).

Reply: User requirement UR2.17 requires that the GSA support queries for a list of positions. We have added new user requirements for a batch interface, and for making the GSA available to users with poor network connections to the archive site (UR2.21, and UR2.21). We have also modified UR1.9 to indicate that a GSA users can receive data on transportable media such as tape.

2.2.2 **Comment:** Add more information on the retrieving process. The GSA should provide a report on the parameters used to query and later to retrieve the data. Sometimes I have search the CADC and retrieved some data for different objects and I have got confused on which data is for what. Providing a file with this information will help.

Reply: This is a good idea, and not difficult to implement. We have it as a new user requirement (UR2.20).

2.2.3 **Comment:** I wasn't clear what the status of the "seeing" is in the GSA. It could be considered as part of the weather information. It should be available for all the science observations.

Reply: There may be a "seeing" monitor as part of the weather system, which will be available immediately, for all data. A better "seeing" parameter may be derived from the data, but that may not be available until the data is public (that is a Gemini policy issue).