Gemini Science Archive

Phase II Proposal

Version 03

9 August 2002

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

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

Table Of Contents

Created:February 14, 2002
Modified:August 9, 2002

Gemini Science Archive

Chapter 1 *Introduction*

This document is a proposal for the implementation and operation of the Gemini Science Archive (GSA). The requirements and design of the GSA are described in detail in [2], [5], [6], and [7].

The implementation of the GSA described in the *Gemini Science Archive Conceptual Design Document* [5] is an extension of the existing Canadian Astronomy Data Centre (CADC) archiving infrastructure. If implemented as described, the GSA will be a unique and ground breaking archive at the forefront of astronomy data archiving. This design of the GSA will enable the archive science described in the *GSA Operational Concept Definition Document* [6], allow the GSA to participate in the current international Virtual Observatories (VO) initiatives, and support astronomical data mining activities which are far beyond any existing telescope archive.

The CADC archiving infrastructure has been under development since 1990, and is unique in its support for archives of data from both ground-based and space-based observatories. The major archives currently hosted by the CADC include those of the Hubble Space Telescope (HST), the James Clerk Maxwell Telescope (JCMT), the Far Ultraviolet Spectroscopic Explorer (FUSE) satellite, the Canadian Galactic Plane Survey (CGPS), and the Canada France Hawaii Telescope (CFHT). Hosting the archives of both ground-based telescopes and space-based telescopes has allowed the CADC to see the potential of astronomical archives in space based telescopes, and to see the changes needed to allow ground based telescopes to share this potential. The CADC has a long history of innovative developments in archiving, including on the fly recalibration, previews, web based interfaces, associated data sets, and derived data products. The CADC intends to continue these innovations in archiving in both the development and operation of the GSA.

Continued development of the GSA, in combination with the development of other archives hosted by the CADC will allow a total development benefit which is far greater than the cost born by any single archive. With continuing support from Gemini, GSA has the potential to be a pioneering force in astronomical archives for many years.

1. References

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

[2] *Gemini to Gemini Science Archive Interface Control Documents*, gsa_ICD/01, Norman Hill, Séverin Gaudet, Daniel Durand, David Schade, David Bohlender, NRC, Felipe Barrientos, Felipe Richardson CONICYT, Kim Gillies, Inger Jørgensen, Gemini

[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 Archive Conceptual Design Document*, gsa_CDD/02,Norman Hill, Séverin Gaudet, Daniel Durand, David Schade, David Bohlender, NRC, Felipe Barrientos, Felipe Richardson CONICYT, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/CDD

[6] *GSA Operational Concept Definition Document*, gsa_OCDD/04, Norman Hill, Séverin Gaudet, Daniel Durand, David Schade, David Bohlender, NRC, Felipe Barrientos, Felipe Richardson CONICYT, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/OCDD/OCDD_initial.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] *Phase II Basic Capabilities Detailed Plan*, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/Proposal/revised/phaseII_basic_detail.pdf

[9] *Phase II Advanced Capabilities Detailed Plan*, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/Proposal/revised/phaseII_advanced_detail.pdf

[10] *Phase II budget*, http://www.hia.nrc.ca/pub/Gemini_HIA/GSA/Proposal/revised/phaseII_budget_detail.pdf

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

[12] *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

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

2. Document Overview

This chapter is an introduction to the GSA Phase II proposal documents. Chapter 2 is a summary of the proposal. Chapter 3 describes the details of the GSA development plan. Chapter 4 describes the details of the operations plan for the GSA.

Created:February 14, 2002
Modified:August 9, 2002

Gemini Science Archive

Chapter 2

Summary

This document summarizes the NRC proposal for developing and operating the Gemini Science Archive.

1. Proposal Overview

This document is an NRC proposal for the implementation and operation of the GSA. NRC recognizes the contribution made by CONICYT to Phase I of the GSA development, and if at some future time Chile establishes an astronomy archiving infrastructure or initiates archive research projects, NRC is open to future collaborations with CONICYT.

This proposal assumes that Gemini Facility instruments will be producing data starting in early 2002, and that development of the GSA will begin in December 2002. The development of the basic capabilities of the archive should be completed approximately one year after the start date. The total cost will not be affected by any reasonable adjustments in the starting date of the development phase of the project.

The implementation plan described in Chapter 3 will result in the creation of both the basic GSA capabilities and the advanced capabilities, with the exception of the correlated object tables (see Chapter 3 for details), as described in [7]. The advanced capabilities are optional and could be implemented over an extended period as a part of the GSA operations budget.

This proposal assumes that operation of the GSA will begin in April 2003. Early operation of the archive will be achieved by extending the prototype developed as part of the conceptual design study workscope to provide a basic raw data archiving system, and incorporating new functionality as it is developed. The plan for operating the GSA is described in Chapter 4.

2. GSA Overview

The document entitled “The Scientific Case for a Gemini Data Archive” [12] presented a vision for a facility that would significantly enhance the scientific output of the Gemini Observatories. This would be achieved by creating a state-of-the-art archive of astrophysical data, giving scientists the power to exploit Gemini data in several new and different ways. First, the data could be used in new areas of study that were not anticipated at the time of their acquisition. Second, scientists could apply newly-developed and more powerful analysis techniques to archival datasets. Third, and most importantly, scientists could exploit the collective value of the entire archive of Gemini data thus having access to an ensemble of datasets that no single Principal Investigator would ever have access to. In the age of the Virtual Observatory the Gemini Science Archive will make it possible to join Gemini datasets with other data from a vast array of instruments sampling many wavelengths. None of these research opportunities could be realized in the absence of a Gemini Science Archive.

The purpose of this summary is to illustrate that the vision of the Gemini Science Archive that was presented in the original “Scientific Case” document was preserved intact throughout the entire Phase I process. The Conceptual Design produced by that process is capable of delivering the scientific power contained in the original vision to users of the Gemini Science Archive. The present proposal from the Canadian Astronomy Data Centre to implement that design in this Phase II proposal document represents a realistic plan to make that vision — of the archive as a facility to maximize the scientific output from the Gemini telescopes — a reality.

2.1 The Phase I Process

The Phase I Workscope document [4] was the result of a process that began with the “Scientific Case for a Gemini Data Archive” [12]. The “Scientific Case” document approached the archive problem from the point of view of a user. What benefits to researchers would follow from the creation of a Gemini Science Archive? What would users want to do with the archive? What was the history of archival research? Ultimately, that document argued that the benefits were very large, not very expensive, and that they should be pursued. That document also laid the groundwork for the Phase I work in a number of ways. It made clear that the job of an archive of astrophysical data was not simply to store pixels. Pixel data have no meaning in themselves. It is only in the context of descriptive information (metadata) about those pixels that the pixel data can be interpreted. The range of necessary metadata is large. It includes telescope and instrument configuration information, observation logs, weather information, and metadata describing any processing that is performed. Calibration data must be available and associated with the observations. The importance of processing pipelines was made clear.

The “Scientific Case” document discussed the importance of the integration of the archive facility into the Gemini operations environment so that its interests (and indirectly the interest of archive users) would be considered at the time that operational decisions were being made. That document also anticipated the need to maintain the capabilities to enable inter-archive access that will be critical in the Virtual Observatory era and it observed that a science archive is an evolving facility that requires resources for ongoing development throughout its operational life.

At its April 1998 meeting the Gemini Science Committee (GSC) reviewed the science cases for a Gemini archive and made 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 resolution was endorsed by the Gemini Board in April 1998. Progress towards the Phase I Workscope continued with a Workshop in Hilo in September 1998 that produced many of the Workscope requirements (see [11]).

The Phase I Work Scope defined a path from the vision expressed in the “Scientific Case” document to a Conceptual Design that could deliver on the promise of that vision.

The Operational Concept Definition Document (OCDD) [6] was comprised of the science cases that drive the design, the operational scenarios describing how an astronomer would interact with the archive, and the user requirements that follow from the science cases via the operational scenarios. This document laid a formal foundation for a number of features that needed to be enabled with the GSA if its users were able to achieve their science. A complete set of descriptors and metadata to describe those descriptors is needed. Users need to have access to a detailed description of the instrument, facility, and environment at the time a particular dataset was created. The need for “advanced” features such as unification of astronomical catalogues to allow joint queries across Gemini and other archives, the importance of data processing as an intrinsic part of the archive, and the linkage of publications to data, all came directly and naturally out of the scenarios that were envisioned for users to interact with the system. It is noteworthy that very reasonable user scenarios produced requirements on the archive that no existing archive can satisfy.

The Functional and Performance Requirements (FPRD) [7] document took the OCDD as input and translated the operational requirements from that document into performance requirements on the Gemini Science Archive. The software requirements defined in this document include requirements on the interface, the input data rates that will need to be accommodated, and the data rates that are required for user access to the archive. It also defines the relationships between the GSA systems and the Gemini Observatory systems. This document defines the requirements of a scientific data archive at a level of detail that had never been done before and represents a step forward in the development of scientific data archiving.

The Interface Control Documents (ICD) [2] describe how the Gemini Science Archive and its staff interacts with the Gemini Observatories and their staff. One of the major design problems in the interface is the method of transferring metadata from their point of collection to the archive and then on to the user. Without reliable metadata the archive of pixel data cannot be scientifically exploited. For example, many applications require that the weather conditions were photometric at the time of observation and it is the metadata that provides the answer to that question, among many others. The proposed solution is a “metadata database” that serves as the collection point for metadata that has been “harvested” from various Gemini Observatory systems. That database is then transferred to the archive using a replication server that automatically copies changes in the database near the telescope to the copy of the metadata database at the archive. This provides an elegant solution to the transfer problem and allows effortless updates and error correction.

The ICD describes important roles for Gemini staff in producing data processing recipes. Staff scientists will have the highest level of understanding of the Gemini instruments and their operations and should be the ones to develop the recipes that will be used to process the data from those instruments. These recipes can then be incorporated into the GSA environment and can process data for archival users. It would be a major advance for ground-based archives to deliver science-quality data products to archive users.

2.2 The result of the Phase I process

The ultimate product of the Phase I process is the Conceptual Design Document [5]. Each document from the Operational Concepts Definition Document through the Interface Control Document was a step toward translating the ideas presented in the “Scientific Cases” document into a design for the archive that could be implemented.

The Phase I process was a lengthy and formal one. There were some delays. However, the process has produced a design that satisfies all of the basic requirements for a science archive, with the additional goals of easy maintenance and extendibility and modularity. The design integrates the Gemini Science Archive into the CADC environment and reuses many components of existing systems in order to reduce cost and development time.

The Conceptual Design satisfies all of the basic requirements for a scientific archive:

- The Data Ingest subsystem adds new datasets and meta-data to the GSA catalogues, and responds automatically to updates and corrections to the meta-data stores by Gemini.
- The Bulk Data Storage subsystem stores and tracks the files of pixel data and the Media Creation subsystem handles creation of removable media for users.
- The User Interface subsystem handles communications between the GSA and its users.
- The Catalogue subsystem manages all types of catalogues in the GSA including the basic catalogues of datasets which the user interacts with to find out what data are available.
- The Data Retrieval subsystem manages requests for data including the enforcement of proprietary data protection, data processing if required, management of an FTP directory for user access, and logging of activity.
- The Data Processing subsystem manages all classes of data processing tasks.

The basic requirements are that users can easily find out what datasets are available and retrieve the ones that they select.

All of the subsystems of the GSA have features that are truly innovative relative to the current state-of-the-art:

- The Data Ingest subsystem initiates “data processing discovery agents” that look for ways that the newly-arrived datasets can be merged with existing datasets via joint processing to create new data supersets and thus add value to the archive content.
- The Bulk Data Storage subsystem handles the primary copies of pixel data but also keeps track of where each dataset can be found on backup media and manages the migration of bulk data from one generation of media to the next.
- The User Interface allows the user to view or retrieve not only raw pixel data but also calibration files, metadata (for example observing logs), and processed data and allows the user to access previews of interesting datasets. It will also provide the links to Gemini documentation about the telescopes and instruments.
- The Catalogue subsystem enables querying on a wide range of catalogue content including advanced data products like source and object catalogues and a variety of metadata.
- The Data Processing subsystem applies data processing recipes (many of which will be supplied by Gemini) to inputs that are identified by the “data processing discovery agents”. This subsystem generates previews, generates image descriptors from the archive content and updates descriptors from NED and SIMBAD (and provides Virtual Observatory linkage in the future).

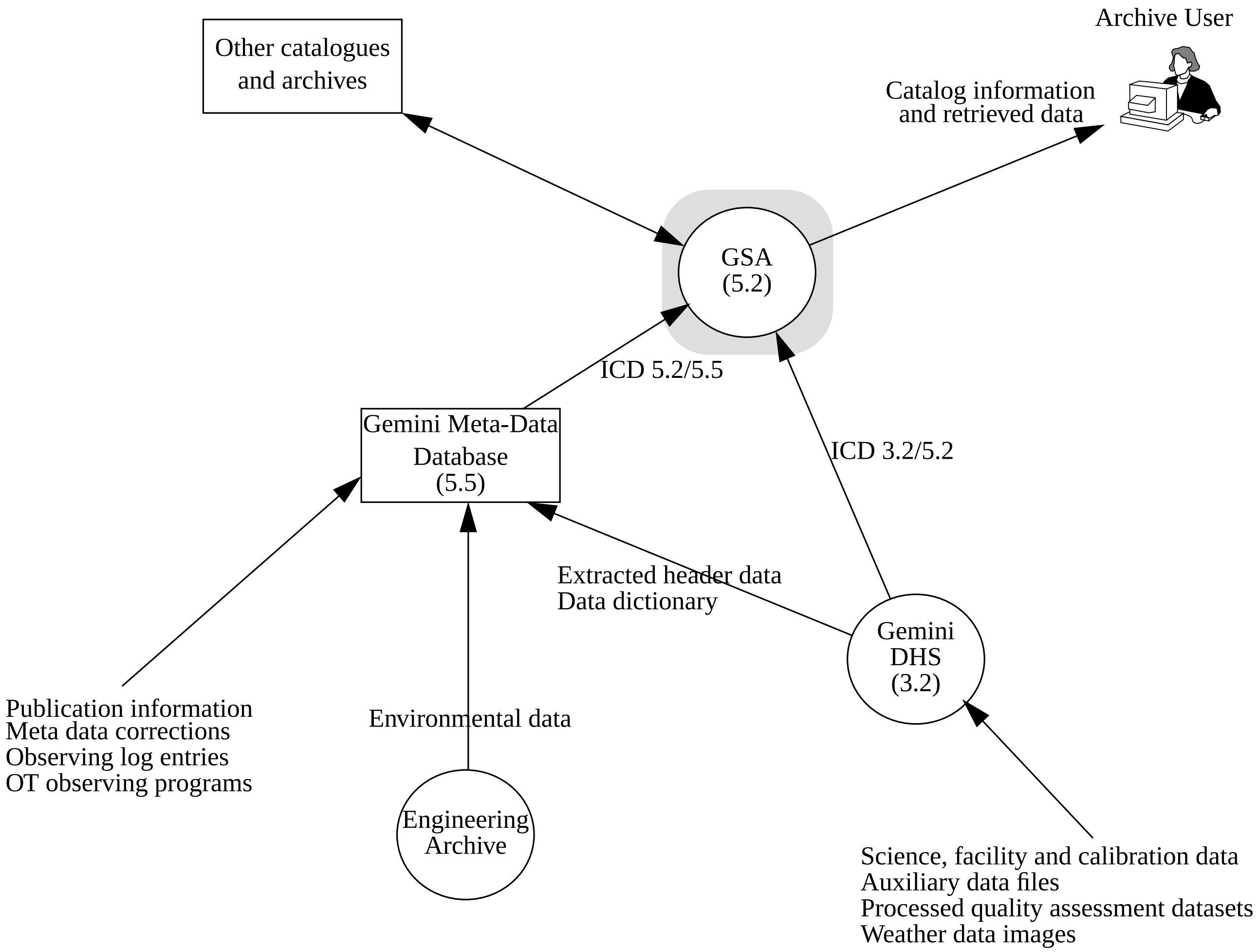
The Conceptual Design for the Gemini Science Archive represents, in itself, a major advance in the art of handling scientific data.

An important point to note is that all of the innovative features presented in the Conceptual Design document were developed as a direct response to the requirements that were developed in the “Scientific Cases” document and in the previous documents that were produce during the Phase I development. This how the design process is intended to work and it has been a complete success. There are “no frills” in our design and there are no “bells and whistles” except for those bells and whistles that scientific users need to ring and blow in order to produce the science that they want to do with the Gemini datasets.

The Phase II proposal is the culmination of a long process of deciding whether the GSA was scientifically important, feasible, and affordable, and deciding what it needed to be able to do in order to be effective. The Conceptual Design is a specification for a system that satisfies all of the requirements for a twenty-first century archive of astronomical data. It sets a new standard for the archiving of scientific data.

An overview of GSA design is shown in Figure 1 on page 7. The relationships of the GSA subsystems are shown in Figure 2 on page 8.

FIGURE 1. Overview of the Interfaces to the GSA.

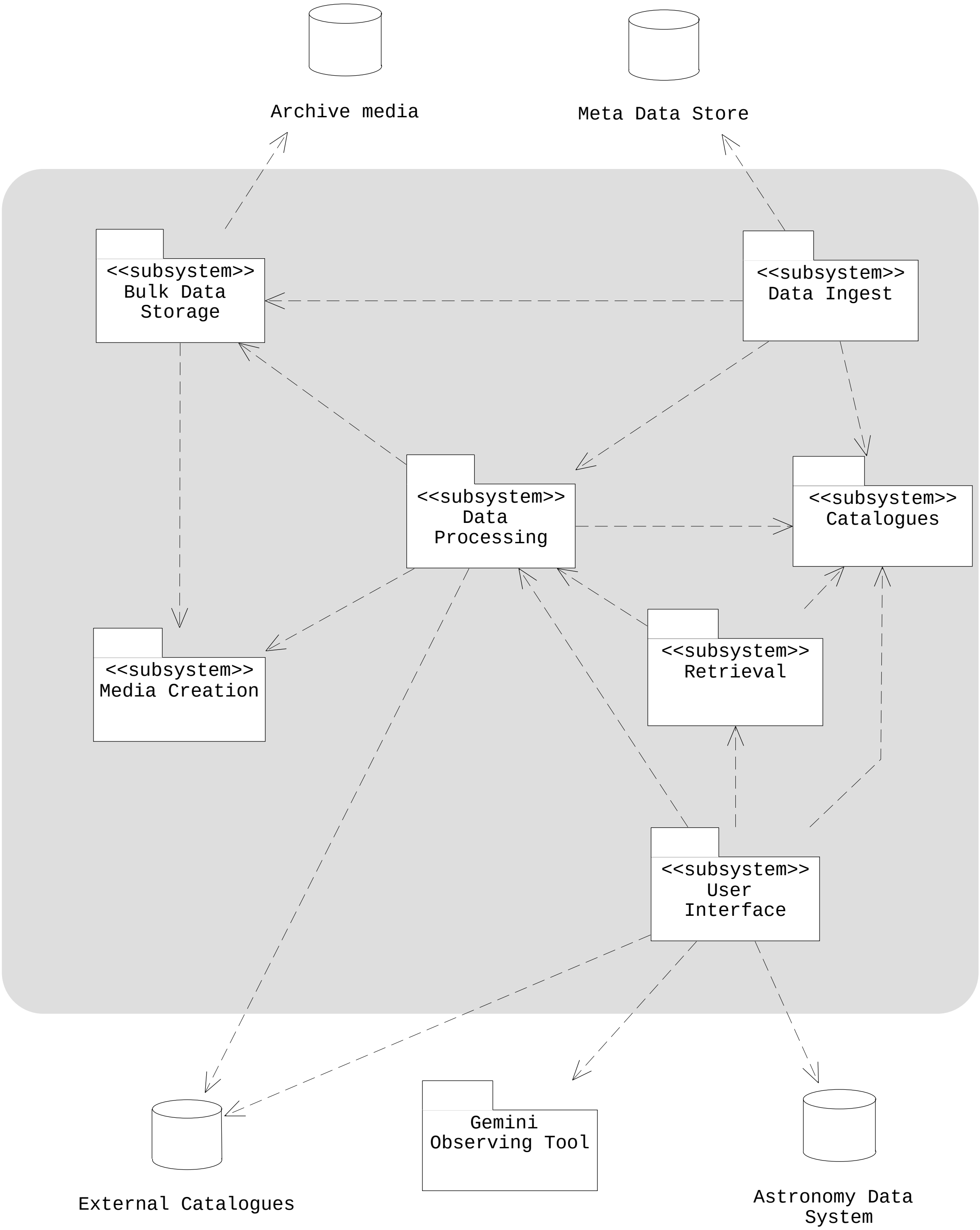


3. CADC and GSA Management Structure

3.1 GSA Phase II Decision Making Process at CADC

The Gemini Science Archive (GSA) is one of a number of development projects that CADC will be dealing with in the next two years. In addition to these new projects, we have a continuing

FIGURE 2. Subsystems of the GSA


Archive media
Meta Data Store
<<subsystem>> Bulk Data Storage
<<subsystem>> Data Ingest
<<subsystem>> Data Processing
<<subsystem>> Catalogues
<<subsystem>> Media Creation
<<subsystem>> Retrieval
<<subsystem>> User Interface
External Catalogues
Gemini Observing Tool
Astronomy Data System

responsibility for operations that must be addressed for all of the CADC archives. These projects compete for the attention of staff.

There are two components to the issue of resource allocation raised by the Conceptual Design Review Committee Report. The first is long-term allocation and is properly addressed by hiring sufficient staff to do the work that is planned. In the case of the Gemini Science Archive there will be a contract in place that specifies the level of staff effort that will be expended on the project and a schedule for expending that effort. CADC will hire sufficient staff so that the agreed-upon level of effort for GSA will be available.

3.2 CADC Organization

The organization of the personnel in the CADC is shown in Figure 3 on page 10. The personnel currently envisioned as being involved with the GSA would be:

- David Bohlender - GSA Project Scientist. In addition to his GSA commitments, David is the JCMT archive Project Scientist.
- Séverin Gaudet - Project manager.
- Norman Hill - Senior programmer.
- Geoffrey Melnychuk or New Hire Junior programmer.

In addition to the personnel shown in the organization chart, we expect that Jennifer Dunn will do the required modifications to the DHS Data Server.

4. Cost breakdown

A summary of the costs associated with proposal are shown in Figure 4 on page 11. These costs include both the development of the GSA, the operation of the prototype archive, and ongoing operation of the archive to March 31 2007. The development costs are broken down into basic capabilities which will be implemented ASAP and advanced capabilities which will be implemented at a later time. The details of the cost breakdown are described in [10].

The dates in the cost breakdown are NRC fiscal years, run from April 1 to March 31.

5. Summary

In summary, we propose to begin development of the GSA in December 2002, and complete development of the basic capabilities approximately one year later. The cost associated with developing the basic archive is $173,168 USD.

If the development of the advanced capabilities begins immediately after the development of the basic capabilities is complete, then the basic capabilities, with the exception of the cross correlated object tables, will be completed by late March 2004. The cost associated with developing the advanced capabilities, with the exception of the cross correlated object tables, is $97,013 USD.

Operation of the basic archive will begin April 1, 2003, with the operational costs being $236,944 USD in fiscal year 2003-2004, and approximately $170,000 USD in per year in subsequent years, in 2002 dollars.

FIGURE 3. CADC Organization Chart

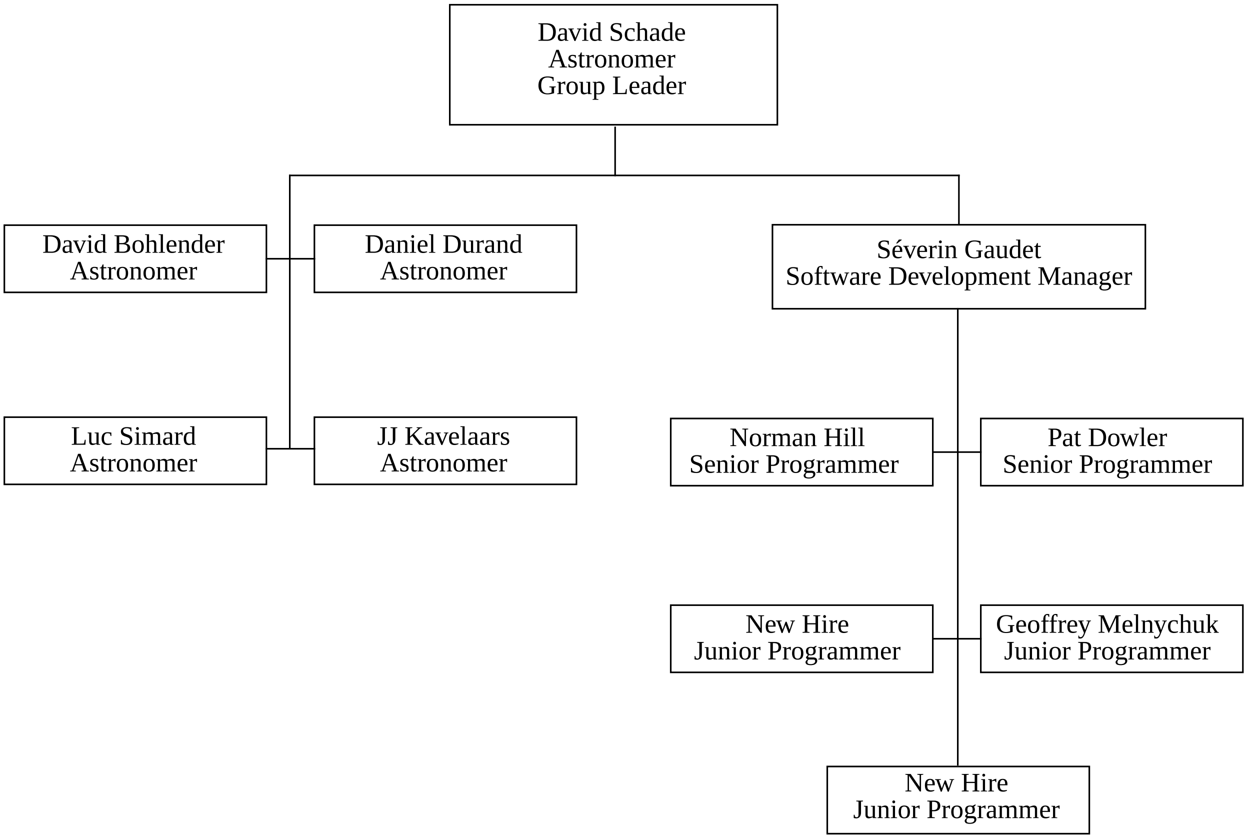


FIGURE 4. GSA cost summary

Gemini Science Archive
Phase II Proposal

Summary
($US)

	2002-2003	2003-2004	2004-2005	2005-2006	2006-2007	2007-2008
Basic Capabilities						
Development[1,2]						
Detailed design and implementation	$ 60,528	$ 76,308				
Hardware and COTS[3] Software	$ 20,332	$ -				
Travel	$ 12,000	$ 4,000				
Total	**$ 92,860**	**$ 80,308**				
Phase 1 Development Cap	$ 92,860	$ 57,140				
Difference	$ -	$ (23,168)				
Operations[1,4]						
Staff[5]	$ -	$ 151,163	$ 122,525	$ 122,525	$ 122,525	$ 122,525
Hardware and COTS[3] Software[6]	$ -	$ 79,781	$ 41,909	$ 40,562	$ 39,349	$ 38,257
Travel		$ 6,000	$ 6,000	$ 6,000	$ 6,000	$ 6,000
Total	**$ -**	**$ 236,944**	**$ 170,434**	**$ 169,087**	**$ 167,874**	**$ 166,782**
Phase 1 Operations Cap		$ 175,000	$ 175,000	$ 175,000	$ 175,000	$ 175,000
Difference	$ -	$ (61,944)	$ 4,566	$ 5,913	$ 7,126	$ 8,218
Advanced Capabilities						
Development						
Detailed design and implementation		$ 97,013				
Total		**$ 97,013**				
Operations						
Staff	$ -	$ -	$ 34,075	$ 34,075	$ 34,075	$ 34,075
Hardware and COTS[3] Software	$ -	$ -	$ 16,708	$ 9,133	$ 9,133	$ 9,133
Total	$ -	$ -	$ 50,783	$ 43,208	$ 43,208	$ 43,208

[1] The start dates of the development and operations phases of the project are independent of each other.

[2] The development phase of the project is scheduled to start December 2, 2002.

[3] COTS - Commercial Off-The-Shelf

[4] The operations phase of the project is scheduled to start April 1, 2003.

[5] Staff costs for operations in 2003-2004 are increased due to handling of the backlog of data from January 1, 2002. Operator costs are expected to be .4 FTE (normally .2 FTE), and project scientist costs are expected to be .75 FTE (normally .5 FTE).

[6] Hardware costs for operations in 2003-2004 are increased due to the initial purchasing of equipment, and the need to purchase storage for both year 2003-2004, and for the backlog of data collected after January 1, 2002.

Created:August 2, 2001
Modified:August 9, 2002

Gemini Science Archive

Chapter 3 Development Description

An overview of the implementation plan for the basic GSA capabilities is given in the Gantt chart shown in Figure 5 on page 14. A detailed version of the Gantt chart can be found in [8]. The tasks to be performed are summarized in Section 2. on page 15, and are described in detail in the *Gemini Science Archive Conceptual Design Document* [5].

Each of the tasks in the project plan will be designed and implemented separately, allowing the functionality developed for each task to become a part of the operating archive as it becomes available.

A Gantt chart showing a plan for implementing the advanced capabilities is available in [9]. This proposal does not include implementation of the advanced capabilities. However, some or all of the advanced capabilities could be moved to the operations budget, where they could be implemented as time permits. The estimates of the cost to develop the data processing recipes required for the advanced capabilities assume that Gemini develop many of the needed recipes for data quality analysis and for use by PIs, and that Gemini will provide these recipes to the GSA.

1. Personnel Allocation

The personnel needed to implement the GSA will be drawn from both existing CADC staff, and from new hires enabled by the revenues from the GSA contract. The tasks to be performed by each individual are shown in [8]. The individuals listed in the development plan are:

PS1 — A senior programmer with database expertise will work full time on the project for most of the first 3 months of the development, after which there will be reduced involvement. (Nominally Norman Hill)

PS2 — A senior web application developer with user interface expertise will contribute 13 days of effort to develop the GSA GUI. (Nominally David Bohlender or Daniel Durand.)

PS3 — A senior programmer with Gemini DHS expertise will contribute 15 days of effort near the beginning of the project to modify the DHS data server to be compatible with the GSA. (Nominally Jennifer Dunn.)

PJ1 — A junior programmer will work full time on the project for the entire duration of the development. (Nominally Geoffrey Melnychuk, or new staff hired with GSA revenue, or a combination of the two.)

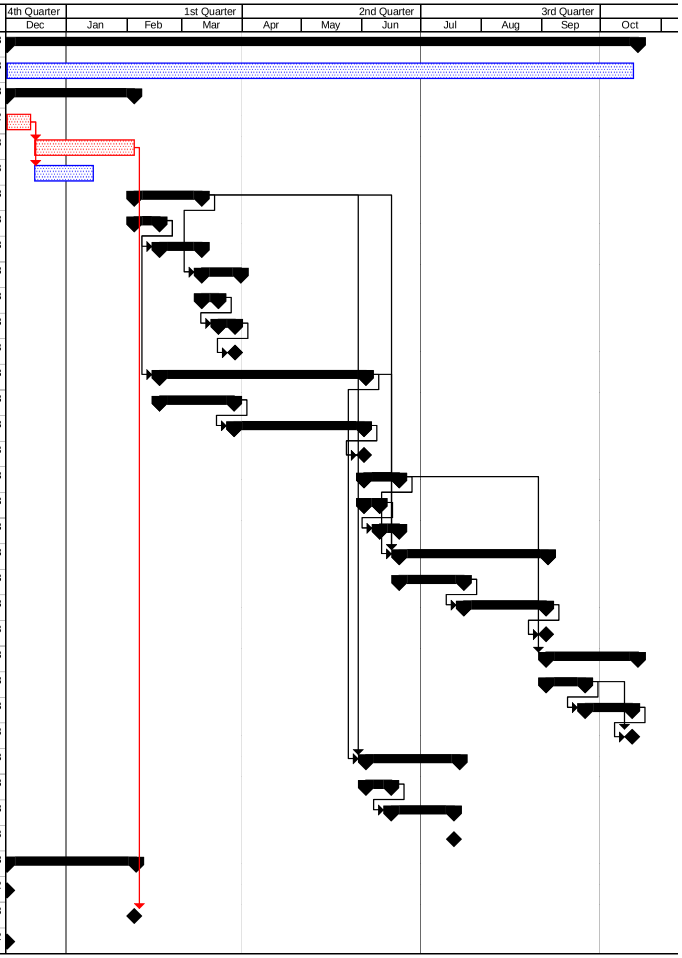
SS1 — A senior scientist will contribute 9 days of effort. (Nominally David Bohlender) This level of effort is for the development of the GSA itself. The development of the Meta Data Database will undoubtedly increase this level of effort.

PM1 — A project manager will contribute 47 days of effort, spread over the duration of the development phase. (Nominally Severin Gaudet.)

FIGURE 5. GSA basic functionality implementation plan

Gemini Science Archive
Phase II Proposal
Development Plan

ID	WBS	Task Name	Work	Cost	Resource Initi	Start	Finish
1	1	**GSA basic capabilities**	**338 d**	**$114,030**		**Mon 12/2/02**	**Mon 10/20/03**
2	1.1	Project Management	54 d	$24,570	PM	Mon 12/2/02	Fri 10/17/03
3	1.2	**Meta Data Database Creation**	**55 d**	**$19,663**		**Mon 12/2/02**	**Tue 2/4/03**
4	1.2.1	Meta Data Database design	10 d	$3,575	PS1	Mon 12/2/02	Fri 12/13/02
5	1.2.2	Meta Data Database implementation	30 d	$10,725	PS1	Mon 12/16/02	Tue 2/4/03
6	1.2.3	Updates to DHS Data server	15 d	$5,363	PS3	Mon 12/16/02	Tue 1/14/03
7	1.3	**Catalogues**	**24.5 d**	**$8,759**		**Wed 2/5/03**	**Tue 3/11/03**
8	1.3.1	**Catalogue design**	**9 d**	**$3,218**		**Wed 2/5/03**	**Mon 2/17/03**
12	1.3.2	**Catalogue Implementation**	**15.5 d**	**$5,541**		**Tue 2/18/03**	**Tue 3/11/03**
16	1.4	**Graphical user interface**	**13 d**	**$4,648**		**Tue 3/11/03**	**Mon 3/31/03**
17	1.4.1	**GUI Interface design**	**6.5 d**	**$2,324**		**Tue 3/11/03**	**Wed 3/19/03**
30	1.4.2	**GUI Interface implementation**	**6.5 d**	**$2,324**		**Thu 3/20/03**	**Fri 3/28/03**
43	1.4.3	GUI release milestone	0 d	$0		Fri 3/28/03	Mon 3/31/03
44	1.5	**Data ingest system**	**71.25 d**	**$19,914**		**Tue 2/18/03**	**Tue 6/3/03**
45	1.5.1	**Data ingest system design**	**28 d**	**$7,826**		**Tue 2/18/03**	**Thu 3/27/03**
54	1.5.2	**Data ingest subsytem implementation**	**43.25 d**	**$12,088**		**Fri 3/28/03**	**Mon 6/2/03**
63	1.5.3	Data Ingest System release milestone	0 d	$0		Mon 6/2/03	Tue 6/3/03
64	1.6	**Bulk data storage system**	**14 d**	**$3,913**		**Mon 6/2/03**	**Fri 6/20/03**
65	1.6.1	**Bulk Data storage system design**	**6 d**	**$1,677**		**Mon 6/2/03**	**Tue 6/10/03**
68	1.6.2	**Bulk Data storage system implementation**	**8 d**	**$2,236**		**Tue 6/10/03**	**Fri 6/20/03**
71	1.7	**Data retrieval system**	**50.25 d**	**$14,279**		**Fri 6/20/03**	**Thu 9/4/03**
72	1.7.1	**Data retrieval system design**	**22 d**	**$6,383**		**Fri 6/20/03**	**Wed 7/23/03**
79	1.7.2	**Data retrieval system implementation**	**28.25 d**	**$7,896**		**Wed 7/23/03**	**Wed 9/3/03**
86	1.7.3	Data Retrieval release milestone	0 d	$0		Wed 9/3/03	Thu 9/4/03
87	1.8	**Batch user interface**	**31 d**	**$8,665**		**Wed 9/3/03**	**Mon 10/20/03**
88	1.8.1	**Batch user interface design**	**14 d**	**$3,913**		**Wed 9/3/03**	**Tue 9/23/03**
92	1.8.2	**Batch user interface implementation**	**17 d**	**$4,752**		**Tue 9/23/03**	**Fri 10/17/03**
96	1.8.3	Batch User Interface milestone	0 d	$0		Fri 10/17/03	Mon 10/20/03
97	1.9	**Preview generation**	**25 d**	**$9,620**		**Tue 6/3/03**	**Mon 7/21/03**
98	1.9.1	**Preview generation design**	**7 d**	**$2,698**		**Tue 6/3/03**	**Mon 6/16/03**
101	1.9.2	**Preview generation implementation**	**18 d**	**$6,923**		**Mon 6/16/03**	**Fri 7/18/03**
105	1.9.3	Preview generation operational	0 d	$0		Fri 7/18/03	Mon 7/21/03
106	2	**External Dependancies**	**0 d**	**$0**		**Mon 12/2/02**	**Wed 2/5/03**
107	2.1	Archive Media Creation	0 d	$0		Mon 12/2/02	Mon 12/2/02
108	2.2	Meta Database replication	0 d	$0		Wed 2/5/03	Wed 2/5/03
109	2.3	Data Processing recipe Creation	0 d	$0		Mon 12/2/02	Mon 12/2/02



2. Development Plan Detail

The tasks listed in Figure 5 correspond to elements of the GSA design, as described in [5]. Each of the tasks are summarized in the following sections. The development of the GSA has been divided into basic and advanced capabilities, which were originally defined in [4]. In general, basic capabilities are capabilities required to allow archive users to locate the data in the GSA which are of interest to them, to retrieve that data to their home computers, and to make productive use of the data.

The advanced capabilities of the GSA are capabilities which will extend the GSA beyond traditional astronomical archives, allowing users to perform more sophisticated queries, making the retrieved data more usable to archive users, and extracting science information from raw data and storing it in catalogues.

2.1 Basic capabilities

This is the development of the software necessary to support the basic archive capabilities described in [6].

2.1.1 Meta-database (WBS 1.2)

Gemini will collect a variety of meta-data describing the state of the observatory before, during, and after an observation. This meta-data will be stored in a meta-database, and the meta-database will be provided to the GSA. The meta-database is described in detail in [2].

After the review of this project it was decided to move the design and implementation of the Meta database into the GSA development scope. Because the effort needed for designing and implementing the meta data database, and the division of responsibilities between CADC and Gemini have not been determined, it is not possible to do a costing for this task at this time.

Because the funds available for the GSA development are capped, development of the meta-database will require other tasks to be removed from the schedule. We have removed time from development of the data processing system to accommodate development of the meta-data database.

2.1.2 Catalogues (WBS 1.3)

This task is the design and implementation of the catalogue tables, and corresponding maintenance software necessary to support the basic capabilities of the GSA. The catalogues are described in Chapter 4 of [5]. This task does not include the catalogue tables which are only necessary to support advanced capabilities. The catalogue tables which are included in this task are:

- The shared tables.
- The data superset catalogue tables. These tables contain meta data describing all datasets and associated data supersets in the GSA.
- The science catalogue tables. These tables contain meta data describing all scientifically interesting datasets and associated data supersets in the GSA.

The catalogue tables needed to support the advanced capabilities will be implemented as part of the *Advanced system* task (see Section 2.2 on page 17).

2.1.3 Graphical user interface (WBS 1.4)

This task is the design and implementation of the GUI components necessary to support the basic capabilities of the GSA. The GUI is described in Chapter 3 of [5]. This task does not include the batch interface, or any GUI components which are only necessary to support advanced capabilities. The GUI components which are included in this task are:

- The GSA opening screen. This web page will be the entry point for GSA users.
- The science catalogue query page. This web page will allow archive users to query the science catalogue tables.
- The data superset catalogue query page. This web page will allow archive users to query the data superset catalogue tables.
- The data superset result list page. This web page will be used to display lists of data supersets returned in response to queries to the database.
- The data superset detailed information page. This web page will be used to display detailed information about a single data superset.
- The preview display page. This web page will be used to select and display a preview image for a data superset.
- The observing program list page. This web page will be used to display a list of observing programs associated with a data superset or proposal.
- The environmental data display web page. This web page will display environmental data collected over a time interval.
- The observing log display web page. This web page will display the observing log data collected over a time interval.
- The data retrieval web page. This web page will allow archive users to select data to be retrieved.
- The proposal Query web page. This web page will allow archive users to query for proposals which match desired criteria.
- The proposal list web page. This web page will display a list of Gemini proposals.

The GUI components which support advanced capabilities will be implemented as part of the *Advanced capabilities* task (see Section 2.2 on page 17). The batch interface will be implemented as part of the Batch user interface task (see Section 2.1.7 on page 17).

2.1.4 Data ingest system (WBS 1.5)

The data ingest system task is the implementation of the GSA side of the Gemini to GSA ICD [2]. The data ingest system is described in Chapter 6 of [5]. This task implements the systems responsible for receiving meta data, meta-data updates, and bulk data from Gemini, and incorporating the data into the GSA.

2.1.5 Bulk data storage system (WBS 1.6)

The implementation of the GSA will generally use existing CADC infrastructure to store and track bulk data. This task will include the design and implementation of two new applications as described in Chapter 8 of [5]. These applications are:

- An application to find files in the bulk data storage area and copy them to local disk.
- An application to monitor and control migration from older archive media to new archive media.

2.1.6 Data retrieval system (WBS 1.7)

The data retrieval system tasks are updates to the existing CADC data retrieval system. These updates are described in Chapter 7 of [5]. The purposes of the updates are to:

- Allow access to proprietary data to authorized Gemini staff.
- Allow access to proprietary data to persons knowing the password for the proposal.
- Incorporate the retrieval processing system into the new CADC data processing infrastructure.

2.1.7 Batch user interface (WBS 1.8)

This task is the implementation of a batch interface to the basic capabilities of the GSA. The batch interface is described in Chapter 3 of [5]. This task does not include any interface components necessary to support the advanced capabilities. The components necessary to support the advanced capabilities will be implemented as part of the *Advanced system* task (see Section 2.2 on page 17). The components of the batch user interface to be implemented as part of this task are:

- A science table query form. This form will allow archive users to query the GSA science table.
- A data superset request form. This form will allow archive users to request data supersets from the archive.
- A query processor to accept queries and data requests, and to perform the required actions.

2.1.8 Preview generation (WBS 1.9)

This task is the implementation of a data processing components described in Chapter 5 of [5] which are necessary to support preview generation. This task includes incorporation of recipes needed to support the basic capabilities of the GSA. Development of recipes needed to support the advanced capabilities of the GSA will be implemented as part of the *Advanced system* task (see Section 2.2 on page 17). The sub-tasks performed as part of this task are:

- Designing and implementing data processing recipes to support the basic capabilities of the archive (preview generation). The costing of this tasks is based on the assumptions that 1) the basic data processing will have been developed by Gemini for data quality assessment and as part of the standard processing software provided for the Gemini instruments, and 2) that only the instruments GMOS North (GMOS-South is expected to be nearly identical and may require little additional effort), NIRI, T-ReCS, and Michelle will be supported from development funds.

  Any recipes developed for additional instruments will be taken from the operations budget, which allows for an average of two new instruments each year.
- Incorporating the data processing recipes into the CADC data processing infrastructure.

2.2 Advanced System

These are the development tasks necessary to implement the advanced capabilities of the GSA described in [6]. A Gantt chart showing the schedule for implementing the advanced capabilities is shown in Figure 6 on page 18.

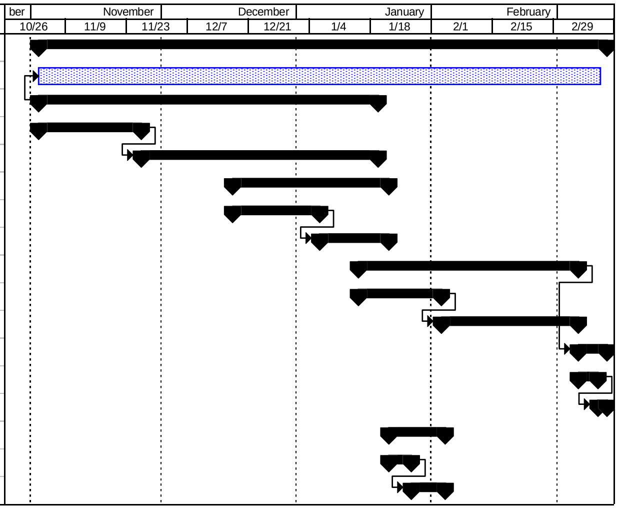
2.2.1 Data Processing (WBS 1.2)

This task is the implementation of the advanced data processing recipes, and the corresponding discovery agents. These recipes would provide archive users on-the-fly recalibration, and the

FIGURE 6. GSA advanced functionality implementation plan

Gemini Science Archive
Phase II Proposal
Advanced Capabilities Development Plan

ID	WBS	Task Name	Work	Cost	Resource Initi	Start	Finish
1	1	GSA advanced capabilities	193.5 d	$75,447		Mon 11/3/03	Fri 3/12/04
2	1.1	Project management	21.25 d	$9,669	PM	Mon 11/3/03	Wed 3/10/04
3	1.2	Data processing	80 d	$29,224		Mon 11/3/03	Mon 1/19/04
4	1.2.1	Design	26 d	$10,842		Mon 11/3/03	Wed 11/26/03
9	1.2.2	Implementation	54 d	$18,382		Wed 11/26/03	Mon 1/19/04
15	1.3	Observation tables	33.25 d	$12,862		Wed 12/17/03	Thu 1/22/04
16	1.3.1	Design	12 d	$4,615		Wed 12/17/03	Tue 1/6/04
21	1.3.2	Implementation	21.25 d	$8,247		Tue 1/6/04	Thu 1/22/04
27	1.4	Source tables	50 d	$20,475		Thu 1/15/04	Fri 3/5/04
28	1.4.1	Design	20 d	$8,450		Thu 1/15/04	Tue 2/3/04
35	1.4.2	Implementation	30 d	$12,025		Tue 2/3/04	Fri 3/5/04
43	1.5	Object tables	0 d	$0		Fri 3/5/04	Fri 3/12/04
44	1.5.1	Design	0 d	$0		Fri 3/5/04	Wed 3/10/04
48	1.5.2	Implementation	0 d	$0		Wed 3/10/04	Fri 3/12/04
52	1.6	Graphical user interface	9 d	$3,218		Thu 1/22/04	Wed 2/4/04
53	1.6.1	GUI design	3.5 d	$1,251		Thu 1/22/04	Tue 1/27/04
56	1.6.2	GUI implementation	5.5 d	$1,966		Tue 1/27/04	Wed 2/4/04



processing required for associated data super sets. The recipes and discovery agents are described in Chapter 5 of [5].

2.2.2 Observation tables (WBS 1.3)

The observation tables are archive independent tables which will contain entries for each scientifically interesting data superset in the GSA. When supported by other archives at the CADC, this table will allow cross archive searches, and is a first step towards participation in a more general Virtual Observatory. This task will include:

- The design and implementation of the observation tables as described in Chapter 4 of [5].
- The design and implementation of GSA specific data processing recipes needed to maintain the Gemini data supersets in the observation tables. These recipes are described in Chapter 4 of [5].
- The design and implementation of a GUI to provide access to the observation tables for archive users. The observation table GUI is described in Chapter 3 of [5].

2.2.3 Source Tables (WBS 1.4)

This task involves the design and implementation of archive independent source tables. These tables will contain information about sources extracted from both Gemini observations, and from observations stored in other archives hosted by the CADC. This task will include:

- The design and implementation of the observation tables as described in Chapter 4 of [5].
- The design and implementation of Gemini specific data processing recipes to needed to maintain the data in the source tables. These recipes are described in Chapter 4 of [5].
- Design and implementation of a GUI to provide access to the source tables for archive users. The source table GUI is described in Chapter 3 of [5]

2.2.4 Object tables (WBS 1.5)

This task involves the design and implementation of archive independent object tables. The object tables are similar to the source tables described in Section 2.2.3 on page 19, but the information from the source tables will be correlated and merged to produce a set of attributes describing astronomical objects. This correlation and merging should cross project, instrument, and wavelength boundaries. This correlation task is a challenging research problem which cannot be scheduled with any degree of accuracy, and so it has been left as on ongoing project which will be funded from the GSA operations budget, and from the budgets of the other archives hosted at the CADC.

2.2.5 Graphical user interface (WBS 1.6)

This task involves implementing the advanced user interface functionality which are not directly related to any of the other advanced capabilities.

3. Dependencies on Gemini

The GSA Data Ingest system will implement the GSA side of the Gemini Telescopes to Gemini Science archive interfaces described in [2]. Before this task can begin, Gemini must provide the following deliverables:

- Documentation describing the Observing program XML documents stored in the Meta-Data observing program table.

The Data Ingest system will be completed and could become operational by January 2003. Before the Data Ingest system can become operational, Gemini must complete the following tasks:

- The Meta-Data database must be created and replicated to the GSA.
- The Gemini DHS Data Server must be modified to extract header information from FITS files and populate the Meta-Data store, and the modified Gemini DHS Data Server must be installed and operational at the Gemini Telescopes.
- The environmental data tables in the Meta-Data store must be populated as a part of normal telescope operations.
- The Electronic Observing log tables in the Meta-Data store must be populated as a part of normal telescope operations.
- The observing program tables in the Meta-Data store must be populated as a part of normal telescope operations.

These tasks will have to be completed at both Gemini telescopes.

4. CADC Infrastructure

The GSA will be implemented as an extension to the existing CADC infrastructure. This will allow the GSA to be implemented with less effort than if the GSA was implemented without re-using the CADC infrastructure, thus reducing the overall cost to Gemini. It should be noted that the GSA will not be a stand-alone archive, since it will be an extension of the existing CADC infrastructure.

The following are the GSA implementations tasks which will benefit from the existing CADC infrastructure:

- The design of the GSA catalogues will be based on prototypes developed at the CADC.
- The GSA GUI will be based on the GUIs used for other CADC archives.
- The GSA Data Processing system will be partially based on prototypes developed for other archives at the CADC, and will use a data processing infrastructure that has been developed by the CADC. The data processing infrastructure will soon be operational for other archives at the CADC.
- The Data Ingest system will be based on designs and concepts used by other archives at the CADC.
- The Data Retrieval System is partially based on the existing CADC data retrieval infrastructure, but will be substantially modified to make use of the CADC’s new data processing infrastructure.
- The Bulk Data Storage will use the CADC’s existing storage infrastructure. New applications will be added to ease the media migration tasks, and to help integrate the Bulk Data Storage system with the new data processing infrastructure.
- Media creation will be done by the CADC’s existing media creation software.
- All of the advanced capabilities are based on concepts developed for other archives at the CADC, although none are currently implemented.

Created:August 2, 2001
Modified:August 9, 2002

Gemini Science Archive

Chapter 4

Operation Detail

The operation of the GSA will begin in April 2003, using the prototypes developed for the conceptual design study workscope. The initial archive will be a raw data archive, providing basic search and retrieval capabilities to archive users. These capabilities will be expanded as the development of the GSA proceeds, and by the end of 2003, the GSA will be a fully functional archive.

After the development phase of the archive is complete, maintenance of Gemini specific software will become necessary, and new functionality will continue to be added to the GSA as time permits. The new functionalities will be implemented under the GSA operations budget.

To allow the archive to begin operation in April 2003, the necessary hardware must be ordered in early February 2003, and will be installed before April 2003.

1. Dependencies on Gemini

To allow operation of the GSA to begin in April 2003, Gemini must deliver archive media to the GSA containing data collected by the Gemini Facility instruments up to that date, and deliveries of archive media must continue as new data are collected.

2. GSA Operations schedule

Some effort will be required to make the GSA prototype operational and available to the public. the steps towards an operational GSA are shown in Figure 7 on page 22. The dates in the figure assume the contract will start April 1 2003. After the prototype is operational, the components developed for the GSA will be integrated into the operational system as soon as practical.

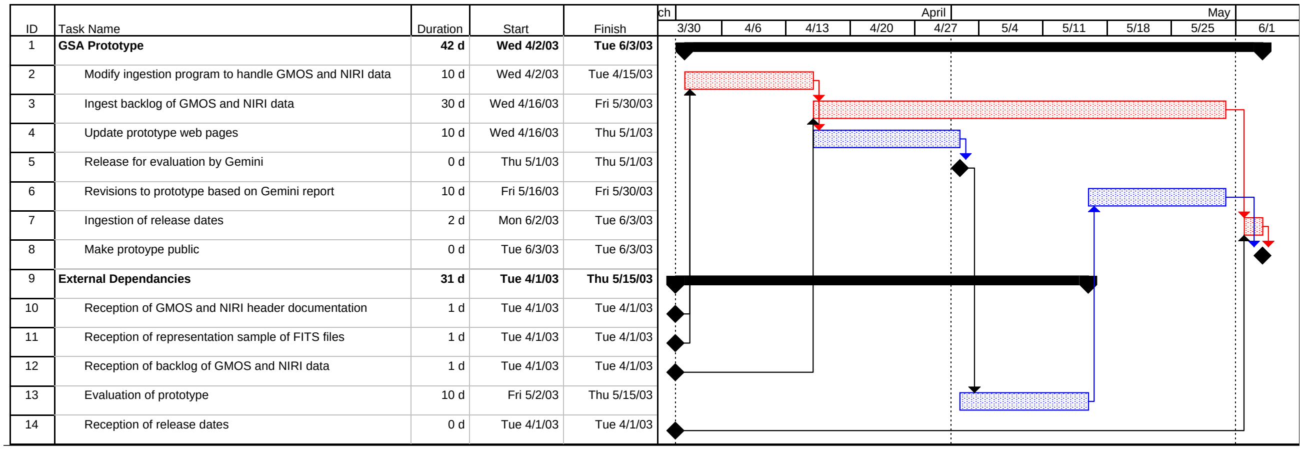
3. Special Projects

The operations budget proposed in this document is sufficient to support the basic operation of the GSA, and the continued development of the GSA to support the evolving Gemini instrument suite. In addition to the budgeted operations, there may be unforeseen special projects which cannot be supported under the GSA operations budget. These special projects will be dealt with on a case by case basis as they arise. Some examples of these special projects are:

- Changes in Gemini operations, or development of new instruments which dramatically increase the data volume or data complexity.
- Gemini initiated special projects which may affect the archive. An example of this type of project is the Hubble Deep Field project.
- GSA initiated proposals to extend the capabilities of the GSA. The GSA may request separate funding from Gemini, or may attempt to fund these projects from other sources.

FIGURE 7. GSA operations plan

Gemini Science Archive
Operations Plan



ID	Task Name	Duration	Start	Finish
1	**GSA Prototype**	**42 d**	**Wed 4/2/03**	**Tue 6/3/03**
2	Modify ingestion program to handle GMOS and NIRI data	10 d	Wed 4/2/03	Tue 4/15/03
3	Ingest backlog of GMOS and NIRI data	30 d	Wed 4/16/03	Fri 5/30/03
4	Update prototype web pages	10 d	Wed 4/16/03	Thu 5/1/03
5	Release for evaluation by Gemini	0 d	Thu 5/1/03	Thu 5/1/03
6	Revisions to prototype based on Gemini report	10 d	Fri 5/16/03	Fri 5/30/03
7	Ingestion of release dates	2 d	Mon 6/2/03	Tue 6/3/03
8	Make protoype public	0 d	Tue 6/3/03	Tue 6/3/03
9	**External Dependancies**	**31 d**	**Tue 4/1/03**	**Thu 5/15/03**
10	Reception of GMOS and NIRI header documentation	1 d	Tue 4/1/03	Tue 4/1/03
11	Reception of representation sample of FITS files	1 d	Tue 4/1/03	Tue 4/1/03
12	Reception of backlog of GMOS and NIRI data	1 d	Tue 4/1/03	Tue 4/1/03
13	Evaluation of prototype	10 d	Fri 5/2/03	Thu 5/15/03
14	Reception of release dates	0 d	Tue 4/1/03	Tue 4/1/03

4. Operations Staff

The duties of the Gemini operations staff are based on two factors: the activities required to support the Gemini specific components of the CADC archive systems, and a “fair share” of the required support for the shared CADC infrastructure. The “fair share” is based on the assumption that the GSA will be one of the major archives supported by the CADC. Some of the duties are essentially open ended, and will be performed on an “as time permits” basis, in particular, the enhancements to the data products available from the GSA.

This budget allows for the changes that will be necessary to support the continued evolution of the Gemini Telescopes and the introduction of an average of two new instruments each year. The budget does not allow for major changes to the interface between the Gemini telescopes and the GSA.

In addition to the listed duties, the NRC overhead will cover the following duties:

- Training.
- Personnel management.
- Generating reports for NRC.
- CADC and NRC management.

The duties of the personnel identified under the GSA operations budget are:

4.1 Project Manager

- Generating project reports for Gemini.
- Coordinating GSA activities with other CADC projects.
- Purchasing.
- Management level interaction with Gemini.

- GSA task planning.
- GSA budget management.

4.2 Scientist

Note that although the GSA project scientist has primary responsibility for the GSA science activities, some of the activities may be delegated to other HIA astronomers in order to match skill sets to tasks.

- Following the development of the Gemini telescopes and instruments. This would involve attaining a PI level of understanding of the data generated by the telescope and instruments.
- Representing the GSA to Gemini. This includes presenting the archive perspective on Gemini operations and ongoing development.
- Representing the GSA and Gemini in CADC meetings, and in planning future CADC projects.
- Generating project reports for Gemini.
- Interacting with Gemini scientists.
- Interacting with users of the GSA.
- GSA user community out-reach.
- Testing of the GSA from an archive user’s perspective.
- GSA user interface evaluation, and enhancement recommendations.
- Collaborating in the initial definition, and ongoing evolution of the meta database.
- Representing the GSA in the larger astronomical archiving community, and specifically in the ongoing International Virtual Observatory efforts.
- Extending the quality and types of reduced data products available from the GSA.
- Supporting the activities of other GSA staff.

4.3 Software Engineer

- Modifying the GSA to support new and upgraded Gemini instrumentation.
- Modifying the GSA to support any changes to the Gemini operations model.
- Fixing bugs in the GSA software.
- Enhancements to GSA user interfaces.
- Extending the quality and types of reduced data products available from the GSA.
- Supporting technological upgrades to the GSA.
- Design of new functionality or modifications to existing functionality of the GSA.
- Implementation of new functionality or modifications to existing functionality of the GSA.
- A fair share of the maintenance of the shared CADC infrastructure.

4.4 System Administrator

- New hardware integration, and in particular addition of new storage devices.
- Maintenance of existing hardware, including failure repair.
- Installing and upgrading system software.

- System monitoring.
- Trouble shooting.
- System performance and tuning.
- Backup.
- A fair share of the maintenance of the shared CADC infrastructure.

4.5 Database administrator

- Upgrades to database hardware and software.
- Database performance and tuning.
- Day to day maintenance of the databases.
- System Monitoring.
- Trouble shooting.
- Backups.
- A fair share of the maintenance of the shared CADC infrastructure.

4.6 Operator

- Reception of data from Gemini.
- Integration of new data into the GSA.
- Migration of GSA data new storage technologies.
- Monitoring GSA operations.
- GSA user support, including creating removable media for archive users.
- Some interaction with Gemini (e.g. modifying Gemini staff permissions to access proprietary data).

5. GSA Safety and Security

The current plan is that Gemini will make two copies of the distribution media. One copy will be retained at Gemini, and the other sent to the CADC. The CADC will copy the data from the distribution media to magnetic disk. This will result in three copies of all raw data, at two separate sites.

The CADC storage area for off-line backups is separated from the CADC operations room by a fire door, and does not contain any likely sources of combustion. The CADC operations room is protected by a fire suppression system and so it is unlikely that a fire in the operations room would result in damage to the GSA backup media.

To prevent unauthorized access to proprietary data, all copies of GSA data will be stored in a physically secure area, and the CADC computer systems will be physically and electronically secure.

To limit the seriousness of miscalculating the shelf-life of future archive storage media, CADC will retain one previous generation of archive media in addition to the newest generation of archive media.

Gemini will retain ownership of all media containing GSA archive data including next generation media created by CADC and media containing data products.

Off site backup of the CADC software and environment allows re-creation of the operating environment.

Regular backups will be made of all GSA derived products and catalogues. In the event of disaster, GSA derived data products and catalogues can be re-created from the raw data, although the complete restoration may take several months.

The fate of the GSA in the event of termination or non-renewal of the contact to operate the GSA will be negotiated between Gemini and HIA, and will be stated in the GSA operating contract.