Astro2020 APC White Paper

Community Science and Data-Intensive Astronomy Support at the US National Optical Astronomy Observatory

Type of Activity:
State of the Profession Consideration, Infrastructure Activity, Ground Based Project

Principal Author: Adam S. Bolton
National Optical Astronomy Observatory
950 N. Cherry Ave Tucson, AZ, 85719
bolton@noao.edu, +1 520-318-8130

Co-authors (all NOAO): Mark Dickinson, Kenneth Hinkle, Stéphanie Juneau, Chien-Hsiu Lee, Thomas Matheson, Sean McManus, Catherine Merrill, Robert Nikutta, Dara Norman, Knut Olsen, Stephen Ridgway, Abhijit Saha, Verne Smith, Monika Soraisam, Letizia Stanghellini, and Francisco Valdes

Abstract: Research opportunity in modern astronomy is defined by both access to observing facilities and access to data. The Community Science and Data Center (CSDC) of NSF's National Optical Astronomy Observatory (NOAO) supports the broad US astronomical community through an integrated approach to both of these major modes of access. CSDC's strategic goals are (1) to maximize community science output from the data sets and facilities of today, and (2) to prepare the community for science with the data sets and facilities of tomorrow. In this white paper, we describe the CSDC mission and program, and recommend that the Astro2020 Decadal Survey endorse a strong program of data services as a critical function of a modern National Observatory.

Key Issue and Overview of Impact on the Field

Historically, research opportunity in optical/infrared (OIR) astronomy has been defined by *access to observing facilities* (i.e., telescopes and instruments). In the present day, opportunity is equally defined by *access to data*. This transition is illustrated by the Sloan Digital Sky Survey (SDSS)[1], which has become one of the highest-impact astronomy projects in history by releasing high-quality data products from large homogeneous surveys through powerful and accessible interfaces. It is illustrated by the number of publications based on *Hubble Space Telescope* archival data having surpassed the number of publications from *HST* primary observing programs.[2] It is illustrated by the fact that the largest single NSF investment in OIR astronomy for the 2020s—the Large Synoptic Survey Telescope, LSST—will be available to the community entirely through a combination of archival data access and real-time data-stream subscription. And it is illustrated by the critical importance of high-quality data products and data archives to realize the potential of US community participation in the Giant Magellan Telescope (GMT) and the Thirty Meter Telescope (TMT).

We propose that the Astro2020 Decadal Survey recognize that *access to observing facilities* and *access to data* are essential and interrelated factors for broad US community science participation and leadership in the 2020s, and that a strong program of data services is a critical function of a modern National Observatory.

Strategic Plan

The NSF'S US National Optical Astronomy Observatory (NOAO), managed and operated by the Association of Universities for Research in Astronomy (AURA), has consistently supported broad-based US community science through peer-reviewed access to observing facilities at Kitt Peak National Observatory (KPNO) and Cerro Tololo Inter-American Observatory (CTIO), and to the US share of observing time at the Gemini Observatory. NOAO has also empowered data-intensive astronomy through the IRAF project, through pioneering initiatives to archive OIR telescope data ("Save the Bits"), through the operation of a Survey Program that prioritizes acquisition of data sets with significant legacy value, and through its role as a founding institution for LSST. Finally, NOAO currently represents the interests of the US Gemini user community through the US National Gemini Office.

The NOAO Strategic Plan of 2015/16[3] identified a confluence of factors necessitating evolution in the role of the National Observatory heading into the 2020s:

- The arrival of forefront survey instruments at NOAO telescopes: the Dark Energy Camera (DECam) at the 4m Blanco Telescope at CTIO and the Dark Energy Spectroscopic Instrument (DESI) at the 4m Mayall Telescope at KPNO
- The growth of a petabyte-scale NOAO imaging archive covering nearly the entire sky

[1] https://www.sdss.org
[2] https://hubblesite.org/contents/news-releases/2011/news-2011-40.html
[3] https://www.noao.edu/dir/strategic_plan/NOAO_Strategic_Plan2016.pdf

- The delivery of major catalogs and other high-level science products from the Dark Energy Survey (DES), the DESI Project, and other large surveys at NOAO facilities
- The rise of ubiquitous time-domain astronomy
- The increasing prominence of data science and new computing modalities (e.g., commercial cloud) throughout the economy and society
- The need to support broad-based community participation in data-intensive astronomy with LSST
- The role of the National Observatory in enabling US community participation in GMT and TMT
- The need to be agile and responsive in OIR system optimization and coordination

In view of these factors, and in coordination with NSF, NOAO consolidated and modernized its broad range of OIR-system and data-science program activities within a new Community Science and Data Center (CSDC), with the strategic goals of

1. Maximizing community science output from the data sets and facilities of today, and
2. Preparing the community for science with LSST and GMT+TMT tomorrow.

The current CSDC Program encompasses the following major activities, all of which are closely coordinated with one another, and all of which address the two strategic goals above:

- **NOAO Data Lab:** Online science platform for research with large public astronomical catalogs and survey data sets
- **Time Domain Services:** Infrastructure and collaborative networks to enable time-domain astronomy, including the Arizona-NOAO Temporal Analysis and Response to Events System (ANTARES), a flexible public "event broker" for the LSST era
- **Data Management Operations:** Core capabilities for transport, ingestion, processing, PI access, long-term stewardship, and archival research access for all data from CTIO and KPNO telescopes
- **Telescope Time Allocation Committee:** Peer reviewed open access to all NSF-funded observing time throughout the OIR system
- **US National Gemini Office:** Interface and advocate for the US astronomical community to the Gemini Observatory
- **Community Science Development:** Including an LSST Community Science Center Working Group, the development of potential US partnerships in international facilities such as the Maunakea Spectroscopic Explorer, and incubation and development of the US Extremely Large Telescope Program

The success of the CSDC Program depends crucially on *(a) a research-active and service-oriented scientific staff that is fully engaged with the scientific community and drives the Program accordingly, (b) a professional software engineering workforce to work in collaboration with scientists to deliver scalable, high-quality data systems and services for astronomy research, and (c) dedicated and professional administrative and project-management staff.*

Organization, Partnerships, and Current Status

The following subsections provide a more detailed description of current CSDC Program activities. Many of these activities leverage partnerships with other organizations (e.g., Gemini, LSST Project, LSST Corporation, Las Cumbres Observatory, NCSA, STScI, University of Arizona).

The most significant anticipated future organizational development for CSDC is the reorganization of NOAO, Gemini, and LSST Operations into a single National Center for Optical/IR Astronomy (NCOA), sponsored by NSF and managed by AURA. CSDC will continue as a top-level Program within NCOA, and the new organizational framework will allow for much greater coordination with other NCOA Programs on common needs in data management, science operations, and OIR system optimization.

Of particular note, NCOA provides the organizational scale to take on a long-term stewardship and archival science-support role for the data from LSST and other major OIR astronomy projects and surveys.

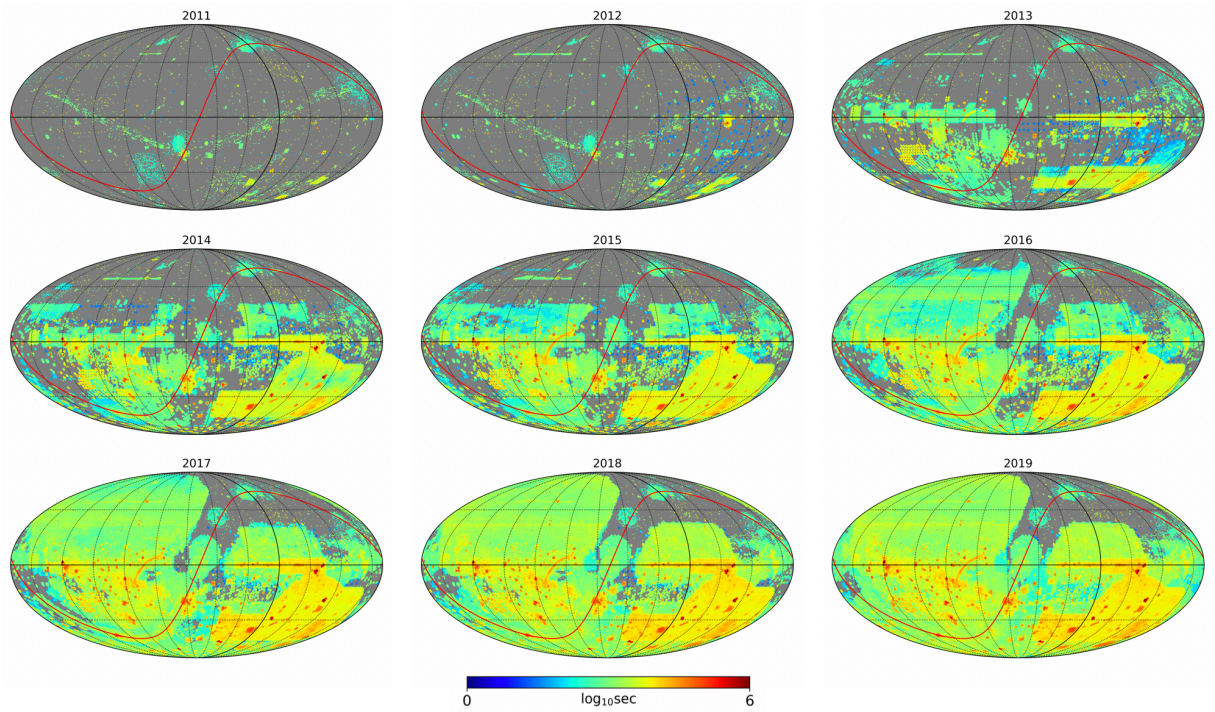

Figure 1: Sky map of integrated exposure time of NOAO archival data holdings (evolution from 2011 to 2019.)

NOAO Data Lab

NOAO's archival data holdings have undergone a phase transition over the past decade, as illustrated in Figure 1. At the start of the decade, most parts of the sky had never been observed by NOAO telescopes or recorded in NOAO data archives. By 2019, most areas of sky are now covered by NOAO's data archives, in some cases over multiple epochs, and/or with

depth and image quality not matched by any other publicly accessible wide-field imaging. To enable the community to fully exploit this resource for science today, and to develop broad based familiarity with the methodologies that will be required for science with LSST data in the future, the Data Lab[4] was launched in 2015 as a cornerstone of the CSDC Program. Public Data Lab services were first released in June 2017.

Catalog-oriented research capabilities: Many science use cases in data-intensive astronomy require sample selection and analysis based on uniform catalogs from large surveys. To support these research applications, Data Lab currently hosts multiple large survey catalogs. Prominent Data Lab-hosted catalogs from surveys with NOAO telescopes include the Dark Energy Survey (DES), the Dark Energy Camera Legacy Survey (DECaLS), the Mayall z-Band Legacy Survey (MzLS), the Survey of the Magellanic Stellar History (SMASH), and the Dark Energy Camera Plane Survey (DECaPS). Data Lab also produces, curates, and serves an "NOAO Source Catalog" (NSC) that provides a uniformly generated object catalog across all survey image data. High-value external reference catalogs from Gaia, SDSS, USNO, and other sources are also hosted within Data Lab. As of May 2019, Data Lab hosts approximately 150 billion rows of catalog data from 17 distinct surveys. These catalogs are accessible through multiple interfaces including a web form, a Python query client, and a Virtual Observatory (VO) standard Table Access Protocol service. Data Lab also provides a cross-match service for users to match their own catalogs against large Data Lab catalog holdings.

Connecting catalogs to pixels: Data Lab hosts an image cutout service based on the VO "simple image access" protocol, allowing users to connect their catalog-based selections to tailored cutouts of the original source images. This efficiently enables science use cases requiring inspection and/or reanalysis of pixel-level imaging survey data. Data Lab is currently developing corresponding spectrum access services, which will be applied to the full public SDSS data sets, as well as to public releases from the DESI survey.

Jupyter notebook services: Data Lab hosts a Jupyter notebook server which allows scientists to run browser-based Python analysis and visualization scripts that leverage the full Data Lab service infrastructure and data holdings. A large library of "example notebooks" created by Data Lab staff provides users with training on using Data Lab services to implement a diverse range of analyses. These example notebooks can be modified and expanded by users to enable their own analyses.

Other services: In addition to the services above, the Data Lab features an interactive "discovery tool" (based on Aladin Lite) with which users can discover available survey data sets, a virtual user storage system (based on the VOSpace protocol) for staging of inputs and results, a "myDB" service for working with personal catalogs within Data Lab, a table browser for exploring the database schema of survey data sets, and a downloadable client library for accessing Data Lab services directly from a user's local machine.

The rapid and efficient deployment of Data Lab has been enabled by leveraging open-source software and common standards wherever possible (e.g., Jupyter, Aladin, Astropy, Pandas, VO standards). Data Lab infrastructure will soon be migrated to a "containerized" framework

[4] See https://datalab.noao.edu and APC white paper by K. Olsen et al.

running under Kubernetes orchestration to enable portability to other environments (e.g., national-level computing centers, or commercial cloud.)

Data Lab staff regularly present hands-on tutorials and demonstrations both in Tucson and at external locations to train astronomers at all career stages in the new data-intensive research methodologies that Data Lab offers.

Currently, the two most significant strategic development priorities for Data Lab are:

- Standardization and technology alignment with LSST and other astronomy data centers, to maximize portability and reusability of infrastructure, commonality of user experience, and interoperability of data sets.
- Support for spectroscopic survey science in the LSST era, motivated by the goal of hosting future public data releases from the DESI survey, as well as enabling maximum scientific interoperability between SDSS and other major data sets.

Time Domain Services

The US community is on the verge of an explosion in time-domain astronomy. LSST is expected to generate of-order 10 million alerts per night, with each alert indicating that an observed object has brightened, dimmed, or moved relative to previous epochs. This "alert stream" promises diverse opportunities in transient and variable science, but only if it is connected to software infrastructure that will allow individual science users and teams to implement their own unique filters for the rarest and most interesting alerts, and to run them in real time.

The Arizona-NOAO Temporal Analysis and Response to Events System (ANTARES)[5], currently developed and operated within CSDC, is NOAO's response to the identified need for a general-purpose community "alert broker" for the LSST era. ANTARES was initiated as a collaboration between NOAO astronomers and University of Arizona computer scientists. The project has been supported by NSF through the INSPIRE program (CISE AST-1344024, PI: R. Snodgrass) as well as through base and supplemental funding to NOAO. The ANTARES project has expanded to encompass collaboration with STScI, NCSA, and Northern Arizona University.

ANTARES is currently running as a live service for filtering of alerts from the public surveys of the Zwicky Transient Facility (ZTF), the most significant precursor to the LSST alert stream. Scaling tests are currently being planned and executed to determine the hardware requirements for running ANTARES as a public service at the scale of LSST. Additional time-domain surveys are being added into the live system, to enable multi-survey filtering (LIGO/Virgo, ASAS-SN, ATLAS).

Within the ANTARES system, users can

- Subscribe to any of a number of pre-defined filtered streams (e.g., "high amplitude", "high SNR", "extragalactic", "in M31")
- Define their own filters to be run in real time, referencing spatial alert history and catalog cross-matches in addition to the parameters of the alerts themselves

[5] See https://antares.noao.edu and APC white paper by T. Matheson et al.

- Develop and test filters using a Python notebook environment within the NOAO Data Lab, and subsequently upload filters into the ANTARES system
- Upload a "watchlist" of objects of interest, to be notified in the event of variability at the position of any watched object
- Receive alerts from their own filters via a web portal, a Slack channel, or a direct programmatic API connection to a Kafka stream
- Query the past history of alerts ingested by ANTARES

ANTARES is implemented using industry-standard open-source technologies (e.g., Kubernetes for container orchestration, Kafka for alert-stream processing), and has been engineered for deployability in multiple environments based on future requirements and opportunities (e.g., on-premises, academic HPC center, commercial cloud).

For the LSST era, CSDC sees ANTARES as a "software instrument" that will make the raw data from the LSST alert stream accessible to all astronomers. The general-purpose nature of ANTARES is an essential design feature: ANTARES is not meant to implement a particular science filter, but rather to be a platform upon which many filters can be implemented. In some cases, individual users will define specific, highly tailored filters, and will receive the output directly. In other cases, ANTARES will be used as a "pre-filter" for dedicated downstream alert processing systems operated by large collaborations focused on specific science cases. To position ANTARES to play this role in the future, CSDC staff are engaging with the LSST Project's "community broker" selection process in the present.

Connecting the broader time-domain ecosystem: ANTARES will be only one part of the time-domain ecosystem in the LSST era. Upstream, LSST and other alert-producing surveys supply the raw material for discovery. Downstream, a network of software systems and telescopes will work together to enable real-time follow-up observations. CSDC staff are collaborating with staff from Gemini, CTIO/SOAR, and Las Cumbres Observatory to develop the Astronomical Event Observation Network (AEON): a federated framework for observational time-domain astronomy.[6]

Within AEON, "target and observation manager" (TOM) software systems can subscribe to events filtered and annotated by ANTARES, and prioritize targets for follow-up based on customized experimental program design and real-time scientist decisions. TOM systems will then forward observing requests to dynamic telescope scheduling systems that balance target-of-opportunity, cadenced, and non-time-critical observations. The resulting data will be reduced and archived automatically. All major components of the AEON system are currently being developed, tested, and integrated, with early science operations expected in 2020.

Data Management Operations

CSDC's Data Management Operations (DMO) group provides core astronomical data-management services for all open-access telescopes at KPNO and CTIO, including:

- Development and operation of a Science Data Archive (SDA) with both interactive and programmatic interfaces

[6] See https://lco.global/aeon/ and APC white paper by B. Miller et al.

- Data capture from 18 different instruments across Blanco, WIYN, Mayall, SOAR, and SMARTS telescopes
- Reliable and validated data transmission and redundant backup
- Regularization and validation of data and metadata for ingest into the SDA
- Operation of the DECam Community Pipeline to provide reduced images to DECam PIs
- Automated implementation of all proprietary periods and PI access mechanisms
- Long-term stewardship of NOAO's petabyte-scale astronomical data holdings
- Support for archival research in collaboration with Data Lab

Driven by DECam, NOAO's archival data holdings have transitioned from terabyte-scale to petabyte-scale over the last decade (reaching over 17 million files totaling over 4 PB uncompressed as of 2019). All data are currently hosted in compressed form by a 2-petabyte GPFS-based storage cluster at NOAO headquarters in Tucson, with double-redundant backups distributed between the University of Arizona, KPNO, and La Serena, Chile. Commercial cloud-based archiving alternatives are evaluated annually but remain cost-prohibitive given the combination of petabyte-scale holdings and a constrained NOAO budget envelope.

DMO systems provide reliable archive ingestion of heterogeneous data from many different telescopes and instruments. Each telescope + instrument combination is characterized by a "personality" that encodes the unique translation and remediation steps needed to ingest its data into the standardized SDA system, which is generic with regard to telescope/instrument particulars. Routine ingestion is highly automated and scalable with minimal human intervention, and new personalities can be added straightforwardly.

DMO is currently deploying a modernized SDA interface based on the open-source Django web programming framework, replacing a legacy system which has become difficult to maintain due to unsupported dependencies and monolithic design, and reducing the number of lines of archive code by a factor greater than 10. The new system will provide programmatic interfaces (APIs) in addition to interactive web interfaces, allowing other software systems to access the SDA directly, and enabling greater integration with Data Lab and the time allocation process (described in the next subsection).

All DMO development activities have to be approached as "fixing the airplane in flight", since daily transport, ingest, and distribution of data are external requirements imposed by ongoing telescope operations. The SDA typically serves data to between 2,000 and 3,000 unique users per month, with an average monthly download volume of 26 terabytes.

DMO also operates the DECam Community Pipeline (CP) and Legacy Survey Calibration Pipelines (LSCP). The DECam CP provides science-quality reduced images to DECam PIs within a week, as well as specialized support for real-time programs (24 hour turnaround), target-of-opportunity asteroid follow-ups, and custom coadds. In collaboration with staff at NCSA, CSDC is currently evaluating options for real-time processing and bulk reprocessing of DECam images.

Time Allocation Committee Program

Open-access observing time on federal and non-federal ground-based optical and infrared telescopes is allocated by CSDC through a peer-reviewed process. This includes time on telescopes at CTIO and KPNO, the US share of Gemini time, open-access time available through

NSF TSIP and MSIP funding, and time available through international time exchanges. Proposals are solicited twice per year for observing time during an "A-semester" (Feb-Jul) and "B-semester" (Aug-Jan).[7] CSDC maintains, updates, and publishes information on the telescope-plus-instrument combinations that are available for open access during a given semester. Three different categories of proposal are described below.

Regular Proposals: Regular observing proposals are reviewed by a Telescope Allocation Committee (TAC), which is convened by CSDC and consists of eight science-based panels: three Extragalactic, three Galactic, one Solar System, and one NASA-Exoplanets (associated with the NASA-NSF Exoplanet Observational Research program "NN-EXPLORE".) The TAC panels meet over a one-week period twice a year in Tucson to rank-order the proposals on the basis of scientific justification and technical/experimental feasibility. The rankings of each panel are merged by a Merging TAC, which provides final recommendations to the NOAO Director (or to the NCOA Director in the future). Once approved by the Director, the ranked proposals are forwarded to their respective observatories for scheduling (or to the International TAC for Gemini.) There are typically 350–400 regular proposals submitted to CSDC each semester.

Survey Proposals: CSDC also issues a Call for Survey Proposals on KPNO and CTIO facilities approximately once per year, and convenes a dedicated Survey panel to review these proposals. Survey programs are characterized by larger time allocations, longer execution terms (up to three years), definition of unique samples, and delivery of datasets with significant archival research value. Memoranda of Understanding are negotiated with the proposing teams of all approved Survey programs to ensure the timely completion of the project and the public availability of data products or other community benefits. Historically, up to 20% of time on KPNO and CTIO telescopes has been made available for allocation to surveys.

Gemini Large and Long Programs: In addition to the Regular and Survey TACs, CSDC works in collaboration with the Gemini Observatory and Gemini partners to support a Gemini Large and Long Program Time Allocation Committee (LP TAC) panel that reviews proposals for Gemini programs that either (i) request larger amounts of time than normally requested or (ii) span a longer length of time (up to six semesters). Participating Gemini partners in 2019-2020 are the US and Canada, and these participants offer up to 20% of their Gemini telescope time to be placed in this large-program pool.

The CSDC TAC program is undertaking developmental activities to improve and modernize the time allocation process. Technical efforts are focusing on modernization of the TAC proposal submission and information-processing system, and include CSDC-DMO staff in collaboration with TAC program staff. The TAC process itself is being evaluated by TAC program staff in collaboration with the CSDC Director's Office, with a focus on new approaches to quantify and minimize unconscious bias in proposal evaluation, and on necessary evolution in policies and procedures to accommodate an increase in time-domain proposals. On the science front, the emphasis and membership of TAC panels is likely to evolve in order to accommodate an increasing fraction of exoplanet and time-domain proposals.

[7] The most recent Call for Proposals is available at http://ast.noao.edu/observing/call-for-proposals-2019b

US National Gemini Office

The US is the majority partner in the Gemini Observatory's twin 8-meter telescopes, with ~150 NSF-funded nights per year per telescope. The US National Gemini Office (US NGO), organized within CSDC, supports the US Gemini user community through both *operations* and *advocacy*. For a full history of the US NGO, see Hinkle et al. (2018).[8]

US NGO staff support US Gemini users in the various phases of science program planning and execution, from proposal preparation to data analysis, including the following activities:

- Triaging all initial Helpdesk requests for the US partner (i.e., tier 1) and supporting specific tier 2 queries
- Liaising with Gemini Instrument Scientists for problem-solving activities
- Supporting the NOAO/CSDC TAC for Gemini-related issues, and for the interface to the Gemini merging TAC (the International TAC, or "ITAC")
- Providing post-observing support in various ways, including informing the US community of new Gemini-related software or data-reduction procedures
- Keeping the US community informed about Gemini opportunities and new observing capabilities through the NOAO Newsletter, the NOAO "Currents" e-news, and with direct e-mail messages
- Holding yearly workshops at the AAS winter meetings on specific topics that are of interest to the US Gemini user community

All these activities are informed and supported though the online US NGO Portal[9], which is updated frequently. A joint working group between CSDC and Gemini is currently addressing the question of how these user-support activities can be optimized in the reorganization to NCOA.

The advocacy role of the US NGO includes a number of additional activities that support the interests of the US Gemini community:

- Representing the US community at Gemini Operations Working Group meetings
- Representing the US community in the Gemini ITAC process
- Negotiating an annual list of US NGO support tasks of priority to the US community and beneficial to Gemini
- Representing the US community in Gemini science meetings by participating in the SOC and LOC for such meetings (e.g., suggesting US invited speakers and meeting topics)
- Supporting and liaising with US community representatives on the Gemini Observatory User's Committee and Gemini Science and Technology Advisory Committee (GSTAC)
- Serving as an ex-officio member of the User's Committee for Gemini Observatory

Through these activities, the US NGO staff is responsible for ensuring that the needs, goals, and aspirations of the US Gemini community are understood and taken into account as Gemini plans current and future operations and development.

[8] K. Hinkle et al. 2018, Proc. SPIE , 10704, 107042M; also https://arxiv.org/abs/1806.10213
[9] http://ast.noao.edu/csdc/usngo

Community Science Development

With NSF support and guidance, NOAO/CSDC recently convened an "LSST Community Science Center" Working Group, which delivered the following five highest-priority recommendations for support to enable broad-based community science with LSST:

1. Improve communication with the community about details of LSST planning & operations
2. Support the community in the paradigm shift to big-data science through provision of tools, services, and training
3. Help to build and support an LSST follow-up network, especially for time-domain science
4. Provide planning and advocacy for LSST follow-up facilities
5. Advocate for "archival" research grants and computational grants, especially grants that target under-served parts of the community (e.g., smaller universities)

CSDC staff activities in response to all these recommendations are ongoing. Recommendations 2 and 3 affirm CSDC's ongoing commitment to Data Lab, ANTARES, and AEON. In response to recommendation 4 (and other aligned recommendations), CSDC staff are promoting US community engagement with the Maunakea Spectroscopic Explorer (MSE) project.

CSDC staff are also leading the development of the NOAO/NCOA role in the US Extremely Large Telescope Program (see APC white paper by Wolff et al.) The overarching strategic goal is to enable broad-based US participation and leadership in the science of GMT and TMT, on par with the major NSF investment that is needed. The NOAO/NCOA role is envisioned to provide community organization and advocacy, time allocation and key science project coordination, instrument- and pipeline-scientist support, integrated archiving and data systems, dynamic queue scheduling, and a unified US community interface to GMT and TMT operations. The development of this role will build on current competencies across the entire NOAO/CSDC program, while also leveraging competencies of Gemini and LSST within NCOA.

Schedule

N/A

Cost Estimates

The planned CSDC budget for FY20 (as currently projected within the NCOA cost model) is approximately $6.5M. This represents approximately 70% direct payroll, 15% direct non-payroll, and 15% NCOA indirect costs. This budget is funded through approximately 75% NSF base funding, 20% NSF supplemental funding, and 5% grants and other sources.

The total planned CSDC staffing in FY20 is approximately 32 FTE total. This FTE total comprises roughly 50% scientific, 40% engineering and technical, and 10% administrative personnel. CSDC scientific staffing includes an integrated total of approximately 6 FTE research time allocation.