Key Issue and Overview of Impact on the Field:

The previous three decades have seen the development, commissioning, and operation of a variety of large (6.5+ meter) Optical/Infrared (OIR) telescopes on the ground with varying degrees of access to the U.S. community. Some facilities (e.g. Gemini, Keck) have associated public data archives, others offer a sampling of their data to the public, and some offer no public access to their data at all. Although some archives host higher-level science products, these tend to be small releases with a specific science focus, or special positions on the sky. Many of the public data archives for large telescopes available to the United States OIR communities are merely repositories for raw data and few, if any, offer the full functionality of science archives from space missions.

The current state of the ground-based, large-aperture OIR data archive landscape is a direct consequence of two key factors: lack of investment and lack of will. On the investment side, budget constraints force observatories and instrument development teams to deprecate or eliminate data reduction pipelines, archives, and their associated infrastructure. On the will-side, the large-aperture OIR community tends to consider their data as a custom, single-use observations: little attention is given to the future science utility of the data and to the need to uniformly calibrate it. Decades of invaluable data produced by large facilities are available, in principle. In practice, using such data requires an intimate and highly specialized knowledge of the instruments and of the way the data was acquired. In some cases, the lack of adequate calibrations simply render the data unusable. This state of affairs is in sharp contrast with space-based facilities. For example, approximately one half of papers utilizing science from the Hubble Space Telescope today are doing so from archival data.

The community's need to quickly access science-ready data is becoming increasingly important, for example, as the fields of Time Domain and Multi-Messenger Astronomy (TDA and MMA, respectively) continue to emerge. Such is the science potential of MMA that the National Science Foundation (NSF) has identified it as one of its "Ten Big Ideas" (https://www.nsf.gov/news/special_reports/big_ideas/nsf2026.jsp), and over 40 Astro2020 Science white papers have been submitted on the topic of TDA/MMA. Large OIR facilities are called upon to provide critical information (e.g. redshift, host galaxy type, environment) on transient events after being localized by smaller facilities on the ground or in space. Coordination of observations in time and location is often essential, but is hampered if the data arrives too late (e.g. the target has dimmed so much as to be un-observable), and in raw form. The TDA/MMA cases place strong demands on archiving and data infrastructure. For over 20 years, NASA has funded successful thematic and mission archives that have had remarkable scientific impact on astronomy.

These archives are underpinned by robust architectures, but they are not intended to respond in near-real time to cases like TDA/MMA. The archiving community has begun to evaluate the technologies needed to support what might be called "next-generation" archives. We recommend adequate investment to ensure that these technologies can be implemented and sustained. The APC white paper by Miller et al. (2019) similarly argues the need for a coherent infrastructure across ground-based facilities.

Strategic Plan:

To best realize the scientific potential of the data from large telescopes in the ground OIR portfolio and to properly integrate with robust offerings from space missions, the community and the federal agencies that fund it must invest in a comprehensive data services initiative. The envisioned initiative requires that data arriving from the telescope have reliable and complete metadata. It requires that data must have uniform calibrations and associations to the science targets. It requires near real-time processing and ingestion into archives, particularly to aid rapid-response transient and multi-messenger astronomy, and it requires those archives to be easily linked to others on the ground or in space.

We will use the W. M. Keck Observatory and the Keck Observatory Archive (KOA) as case studies for evaluating the investments needed, but note that the recommended infrastructure changes be applied, if possible, to the entire large OIR landscape. KOA archives all data acquired by all Observatory instruments that have been commissioned since the Observatory began operations in 1994. The raw data prepared for ingestion into the archive at the Observatory are transmitted to the archive at Caltech/NExScI by the afternoon following the observations. Where automated pipelines are available, KOA creates browse quality reduced products that are served with the raw data. The raw data are written as FITS files, with the metadata written as keyword-value pairs in a flat block of text. Keywords written in this format are well suited for representation in a relational database such as Oracle or PostGresql, through which queries f formulated in a user interface are performed. Data are downloaded for analysis at the users desktop.

Science goals in the 2020s will require a substantially different architecture. For example, localization and identification of TDA/MMA sources and subsequent analysis mandate near real-time ingestion of data and creation of science products. Moreover, the data should be amenable to fast discovery, access and analysis through science platforms - open-source software hubs that support all tasks needed to access, integrate and analyze data, already under development in astronomy. Desai et al.

(2019, APC white paper) argue that these science platforms are necessary to support the complex analyses to support science goals described in science white papers.

Furthermore, modern data sets are becoming too complex to represent in a FITS/relational database model (Thomas et al. 2015). A simple example from WMKO is the position and orientation of the slits used in multi-object spectroscopy. Work on new formats is underway - JWST will store hierarchical metadata in the ASDF format (Greenfield, Droettboom, and Bray 2015) - but broad take-up by the community will require time and investment. Such metadata are better organized in document-based "no-sql" databases than in relational databases.

Thus we are led to a model underpinned by data formats and representations that respond to modern data, with archive services that support fast ingestion and creation of science products, all of which can be integrated with multiple data sets in science platforms. Fast ingestion also implies tighter integration between telescope and archive.

A detailed description of implementation of a data services initiative at Keck or any other large OIR facility is beyond the scope of this paper, but we outline the essential components here. The goal of a modern data-oriented infrastructure is to enable the community to obtain their data in a usable form and enable high-quality, reproducible science. Meeting this goal requires a number of key elements:

Consistent Calibration: Un- or poorly- calibrated data has little scientific utility, particularly to those who are not already experts on the instrument that produced the data. For those observatories like Keck that are ‘classically’ scheduled, there is little to no requirement that individual observers calibrate their data to a uniform standard for future usability. This contrasts with queue scheduled observatories like Gemini that often have the requisite calibrations for scheduled observations performed by staff. As such, different observatories will have to raise the bar to different levels to produce future-usable data for the community. In the case of classical facilities, this will require significant work to establish calibration monitoring and scheduling software that can dynamically assess the science observations to see if that have met an established calibration threshold. Ideally these systems will have minimal impact on the observations themselves. Observer buy-in will be essential to success, so the systems should be designed to be simple, reliable and effective. The calibration ’problem’ cannot be solved with smart software alone: classically scheduled observatories will need to work hand in hand with their user communities to establish the requisite new policies that ensure that data is consistently and uniformly calibrated.

Complete Metadata: The degree to which essential metadata is written to each file varies significantly from observatory to observatory and, within observatories, from instrument to instrument. Historically, information about sky conditions (weather, seeing, clouds) has been relegated to observing logs, not to the data files, deprecating future utility by the community. For instruments like multi-object spectrographs, the information used to derive slitmasks or fiber placements is often separate from the data itself, and often lost to the community entirely, hampering precise reduction of the observations, and removing essential information such as astrometric mapping of slits on metal to objects on the sky. Additionally, any information that defines a sequence of observations should be written to each file. An example might be an IR observation requiring an ABBA sequence, or a sky frame taken after a target observation. Another example might be an observation as part of a tiling on a region of sky. Each individual file should be able to inform both an observer and the reduction pipeline what the provenance and context of the data is. It is to be noted that complete metadata can only be produced if the observing infrastructure is overhauled so that observations are almost entirely automated (scripted). Keck is already investing heavily in this direction, while focusing on preserving the flexibility and real-time decisions that have been the observatory's main strength so far.

Facility Pipelines: Although general reduction facilities have existed for decades (e.g. IRAF), most instruments on large OIR telescopes have data reduction pipelines (DRPs) tailored to specific instruments. Unfortunately, many of these pipelines are written by individuals or teams of observers with little focus on community adoption and are often focus towards a specific scientific goal. Although the efforts of these teams should be applauded (as often these are the only pipelines available for some instruments), a modern data-centered approach to pipelines should be adopted. To begin, and with community input, observatories should adopt pipeline frameworks that set the requirements for scope of use, freedom of codebase, documentation, support, and archive interoperability. Ideally, these frameworks would be developed in conjunction with the calibration and metadata policies outlined above, as they all flow down to the pipelines. Keck has started the development of a unified, modern framework for data reduction pipelines, similar to the Gemini DRAGONS framework capable of supporting individual instrument pipelines as modules of the general infrastructure. The code is open source and open developed: the main contributors are Keck staff and instrument developers, but members of the community are already making significant contributions. Our strategic model envisions a tiered approach for facility pipelines: some of the existing pipelines will become legacy and will have little or no support (tier 3); new pipelines will be developed as modules for the framework, but will be able to run

independently while the framework is being fully implemented (tier 2); finally, all the tier 2 pipelines will be ingested into the framework and run on a unified platform (tier 1). One of the main requirements for tier 2 pipelines is that they provide a fully automated science-quality execution mode, which will be run with KOA to provide reduced data to all our observers. The observatory will also provide a full-fledged help desk for users of tier 1 and 2 pipelines: this is the key to rapidly produce scientific data and increase productivity.

Rapid Archival Ingestion: As part of the new approach to a data-centered observing infrastructure, we also envision a much faster and tighter interaction between instruments and the archive, promoting our archive from a passive repository to the primary data access facility both at night and after the observations are concluded. The first component of this approach is the implementation of real-time ingestion, whereby FITS files produced by the instruments are processed through our ingestion pipeline and made available to observers within minutes. A quick-loop pipeline will also run on the data as they are being taken, allowing for quality assessment and helping observer optimize their observing strategy. As soon as complete datasets are produced, a science-quality pipeline would also run, producing the best version of publishable data that pipelines can produce. Turning KOA into the main data access method will allow large, dispersed team to collaborate more effectively and will enable rapid turn-around TDA/MMA observation. Rapid ingestion of science-ready data into archives can significantly broaden the scope of access for data to the community. NASA's Key Strategic Mission Support programs (KSMS), for example, have as a requirement that the data be delivered to the KOA in a reasonable time and in a format suitable for use by the community. Rapid ingestion would significantly speed this process, and allow KSMS teams to focus more on the analysis of the data, not its reduction. Other programs may have zero proprietary period, and rapid ingestion brings the data to the full community quickly. Finally, rapid ingestion may facilitate dynamic observing: human and machine algorithmic analysis of one night's data could significantly alter the plan for the next night, maximizing science return.

Discoverable data: In a survey of users by the Infrared Science Archive, 39% of respondents indicated that they had difficulty finding data. This unacceptable state of affairs argues that the US must continue to take a leading role in evolving standards for interoperability and data discovery as the data landscape changes, through its membership in the International Virtual Observatory Alliance (IVOA). Fabbiano et al (2019; science white papers) stressed this point too In response to the need for

improved data discovery, NASA thematic archives are systematically implementing these protocols, and exposing them through web interfaces and through command-line interface. The latter specifically includes deploying Python interfaces consistent with those defined by *astroquery* project. These interfaces will be integrated into science platforms to enable discovery of data. This endeavor is part of a world-wide effort to provide comprehensive data discovery services. Thus, KOA, though not formally part of NAVO, is participating in this enterprise and deploying these interfaces too. We encourage all large US OIR facilities with public archives to do the same.

Science platforms integrated with archives:
As archival data sets become larger, more numerous, and more complex, scientific analysis is greatly empowered by high-level "science platforms" integrated with lower-level data archives. As an example, the NOAO Data Lab (datalab.noao.edu) serves catalogs from the Dark Energy Survey (DES), Dark Energy Camera Legacy Survey (DECaLS) and other major survey projects through multiple database query interfaces, and provides crossmatch services, image cutout services, virtual user storage, and a Jupyter notebook environment to facilitate discovery, visualization, and analysis "close to the data". Data Lab leverages VO standard protocols wherever possible to maximize discoverability and interoperability, and is integrated with NOAO's main Science Data Archive. Similar services are being deployed for LSST data (the LSST Science Platform), and for JWST data analysis at the Space Telescope Science Institute. Ideally, high-level science platform services should be adaptable and deployable at different scales and in different environments according to need. This approach is facilitated by modern virtualization and cloud-computing frameworks.

Schedule & Organization:
Realizing such an initiative will require focused federal agency investment in the form of specific funding opportunities associated with new instrumentation and facilities. Furthermore, targeted funding should be made available to bring the last quarter century of data as close to in-line with new data as is feasible. Beyond resources, the community itself must embrace the philosophy of the future science their data can realize, even at the potential small cost cost of some on-sky time for proper calibrations. The community must likewise be willing to participate in expanding the depth and breadth of publically available data reduction and analysis software, and reward that participation accordingly.

The Role of the Federal Agencies:
Federal agencies, the NSF and NASA in particular, will play a critical role in facilitating and driving implementation of a US OIR data services initiative. Naturally, targeted

funding will be important, if not essential. For example, the NSF could augment the key instrument building funding mechanisms (MRI and MSIP) to explicitly require pipelines that adhere to the pipeline framework for the observatory where the instrument is to be deployed, and make the data services component of an instrument a critical component of the budget and schedule that cannot be fully descoped. NSF could also mandate that data produced by the pipelines be compliant with national and international standards to be easily discoverable in archives. NASA, through its investments in Keck and other facilities, along with its archival research program ADAP (which does support research on data from archives like the KOA) could expand these programs to fund infrastructure and pipeline development. Sustained investment by the agencies in archives coordinated with the observatories will yield significant returns for the entire community. Finally, NSF and NASA can continue to support data archiving and data-product production as key competencies of NSF- and NASA-funded Centers such as NOAO, NRAO, Gemini, LSST, STScI, and IPAC.

The role of the observatories:

Naturally, a significant portion of implementing a data services initiative will fall on the observatories. Significant changes will need to be made on software infrastructure, including interfaces to instrumentation and the telescopes. For partnerships like Keck+KOA, significant, sustained coordination is required. All these elements require long-term resource allocation and prioritization to be successful. Observatories will also need to work with the agencies and with each other on data reduction pipeline framework and archive standardization to optimize science return to the community. In many cases, observatories will need to modify their cooperative agreements to formally include data services. In other cases, existing observatory data-management practices will require significant upgrades and continuous improvement to maintain performance in a rapidly evolving scientific and technological landscape. While data services represent either a significant increase or new resource allocation entirely, the return for observatories is significant: they will better serve the US community, and represent a better investment for future instrumentation that is at least partially federally funded. Given the science of the next decade and beyond, new instrumentation and better data are required, so the need for an observatory to be a better partner with the community is further underscored.

The role of the community:

Finally, community buy-in and participation is critical. For classically scheduled observatories like Keck, the current norm for observing differs from that which is envisioned for the future under a data services initiative. The primary difference is

calibrations. In an ideal implementation, calibrations will be performed before or after the night's observing, and in an automated way "behind the scenes", but some instrumentation or observing modes will require consistent calibrations that occur during twilight (e.g. sky flats or flux and telluric standard stars). While the time needed for these calibrations is small, they are not zero, and the community will need to work with observatories to set calibration policies that benefit both the immediate and archival use of the data. How observations are planned for execution will also change, but here the impact is limited to learning new systems. Nevertheless, the observing community will need to buy-in on the need for change. Even more critical is the need for community support for software, specifically data reduction pipelines, not just for future instrumentation, but for data from current and past instrumentation to leverage two-plus decades of raw data from large OIR telescopes. Observatories may set frameworks, and instrument teams may develop pipelines, but much of the pipeline development or enhancement will come from the community, as will the analysis tools for the post-processed data. The community, the agencies, and observatories need to create and adopt new mechanisms and policies to support and reward open software development, as altruism alone will not bring the software to match the science needs of the next decade.

Cost Estimates:

Costing of a national OIR data services initiative will depend on individual observatories, their communities, and their relationship with federal agencies and private funding sources. As such, we cannot provide a cost estimate in this white paper. Nevertheless, the initiative will certainly require significant investments, particularly in the early stages of bringing observatories up to a common level. We recommend that federal agencies augment their existing instrument building and archival exploration funding lines, or create new programs targeted at data services infrastructure. Whatever the final cost, which should be small compared to a new observatory or space mission, the return to the entire U.S. astronomical community would be significant: entire new sources of data from the world's most powerful facilities in a readily usable form, in many cases stretching back decades.

Given the considerations presented here and in other white papers, we recommend that the agencies invest in data services initiatives and the necessary infrastructure across all large ground-based OIR facilities to ensure a uniform level of service for the astronomical community.

References

Desai, V. et al. 2019. “A Science Platform Network to Facilitate Astrophysics in the 2020s.” Astro2020 Activities, Projects, or State of the Profession Consideration White Paper.

Fabbiano, G. et al. 2019. “Increasing the Discovery Space in Astrophysics - A Collation of Six Submitted White Papers.” Astro2020 Science White Papers. https://arxiv.org/abs/1903.06634

Greenfield, P., Droettboom, M. and Bray, E. 2015. “ASDF: A new data format for astronomy.” Astronomy and Computing, 12, 133-145, 240251. https://doi.org/10.1016/j.ascom.2015.06.004

Miller, B. W. et al. 2019. “Infrastructure and Strategies for Time Domain and MMA and Follow-Up.” Astro2020 Activities, Projects, or State of the Profession Consideration White Paper.

Thomas, B. et al. 2015. “Learning from FITS: Limitations in use in modern astronomical research.” Astronomy and Computing, 12, 133-145. https://doi.org/10.1016/j.ascom.2015.01.009