Preserving the Legacy of Ground-Based Data Sets

Submitted by: Michael Blanton,Joel Brownstein, Nancy Chanover, Brian Cherinka, Juna Kollmeier, Karen Masters, Constance Rockosi, José Sánchez-Gallego, Jen Sobeck, Ani Thakar, Benjamin Weaver, Anne-Marie Weijmans

In response to the request from the National Optical Astronomy Observatory (NOAO) for input on the 2020 Decadal Survey of Astronomy and Astrophysics, and the areas in which NOAO can provide critical resources, we are writing to advocate for a national center for archiving and distribution of ground-based data sets. Our discussion here will be centered on the use case for the Sloan Digital Sky Survey (SDSS) long-term archiving needs and those of similar projects. To summarize: without a long-term archiving center for ground-based data that can handle large coherent data sets with sophisticated access methods, the community's access to SDSS data and the guarantee of its data integrity into the future will become at risk. With the right set of strategic choices, NOAO or its reconfiguration as a National Center for Optical-Infrared Astronomy (NCOA) could serve as the natural host for such a center and play a critical role in preserving the legacy of ground-based data into the distant future.

The SDSS is a large astronomical survey which began taking observations in 1998. Its operations are managed by the Astrophysical Research Consortium (ARC), and it is currently supported by the Sloan Foundation, the U.S. Department of Energy, and a large international consortium of over 60 institutions. In the past, it has been supported also by NASA and by the U.S. National Science Foundation. Currently, the institutional members provide the majority of the funding for operating the observatories and executing the project. The project is now in its fourth phase (SDSS-IV), which is due to end in 2020. Planning for a fifth phase is underway.

Between 1998 and 2009, SDSS used the 2.5 meter Sloan Foundation telescope at Apache Point Observatory (APO) to execute the widest field digital imaging program up to that time. During this time, the imager also executed a large time-domain survey (the SDSS Supernova Survey). Since 1999, SDSS has been executing wide field optical and infrared spectroscopy at APO. These observations included the largest redshift surveys to date: the Legacy Survey, the Baryon Oscillation Spectroscopic Survey (BOSS), and currently the extended Baryon Oscillation Spectroscopic Survey (eBOSS). The Mapping Nearby Galaxies at APO (MaNGA) program uses fiber-based integral field units to study the internal structures of nearby galaxies, and has created the largest existing sample of spatially resolved spectroscopy. The SDSS Extension for Galactic Understanding and Exploration programs (SEGUE-1 and SEGUE-2) mapped the stellar halo of our Galaxy with optical spectroscopy. The Apache Point Observatory Galactic Evolution Experiment programs (APOGEE-1 and APOGEE-2) use high signal-to-noise ratio infrared spectroscopy to study the Galaxy. In 2017, APOGEE-2 has begun a new program to map the Galaxy from the southern hemisphere, using the du Pont Telescope at Las Campanas Observatory.

From these data, SDSS has produced 15 public data releases, starting with the Early Data Release in 2001 and most recently Data Release 14 in July 2017. These data releases experience heavy use from the public and the astronomical community. The public web sites experience half a billion hits per year currently (2.5 billion hits to date) from over half a million unique users, and are used by thousands of students in classes. The “power-user” tools, also publicly available but primarily used by professional astronomers, have over 5000 unique users per year. SDSS data has been used in over 7,000 refereed journal articles; the majority of these (> 80%) are from astronomers using the public data sets.

There is no mechanism to continue serving this data to the astronomy community or the public indefinitely. Developing the data releases and maintaining the data distribution systems is a costly endeavor, amounting to of order 10% of the total project cost. Providing continued access to the data sets through the existing interfaces---or at all---has always been contingent on the existence of the next phase of SDSS maintaining the data releases from the previous phases. ARC does not have the central resources available to maintain the data sets otherwise.

The SDSS is not unique in this respect. Other U.S. facilities will produce large data sets that require a long-term archive. Some of these, such as the Dark Energy Camera (DECam) and the Dark Energy Spectroscopic Instrument (DESI), may be able to find a home in large centers existing for the use of particle physics (such as NERSC). However, the experience with SDSS suggests that relying on the DOE national laboratories continues to work only as long as the science program remains of interest to current particle physics priorities. This may not be true over the long term for DECam and DESI. Additionally, other ground-based programs using U.S.-related public and private facilities (the Zwicky Transient Facility, Las Cumbres Observatory, the Prime Focus Spectrograph, and others) will ultimately face the same reality as SDSS: when the project itself is no longer funded, they will either end the distribution of their data, find a way to continue funding the data distribution facility alone, or shoehorn themselves into some other facility’s data distribution system.

Obviously, the Large Synoptic Survey Telescope (LSST) will produce the largest new data sets. It too will ultimately face this same problem. We note that the effort of maintenance is driven partly by data size, but also by data complexity. Therefore some of the spectroscopic programs (integral field spectroscopy, for example) may require substantially more effort than the data size alone implies.

The astronomers running all of these projects will be committed to keeping their data available to maximize their impact. However, as matters currently stand, they will all find different and potentially incompatible ways to address the long-term archiving issues they face. This situation will make for a more heterogeneous system for users, that is also in aggregate more expensive to operate, and harder to maintain into the future. If a central effort can be organized, it could be cheaper, easier to maintain, and easier to interact with.

A central long-term distribution system for ground-based data is therefore called for; however, it is not an easy task. There are three levels of such a system: the archiving of the data, providing access points for the data, and maintaining programmatic and user interface tools. For the first two levels, NOAO has experience archiving and distributing many disparate, heterogeneous types of data. For the third level, incorporating and maintaining the interface tools developed by individual projects could be a challenge; for example, even within SDSS, there are two different platforms (one based on Microsoft tools and one based on open source tools) used for these tools, both heavily used but for different purposes. Therefore, the ambitions for maintaining interface tools need to be clearly defined and include enough standardization to make the system maintainable, without being so restrictive as to make it impossible to deploy the tools developed by the individual projects.

The SDSS has begun an effort, funded by the Sloan Foundation, that will start exploring the possibility of handing off its public data distribution systems to NOAO and/or to the Mikulski Archive for Space Telescopes (MAST). SDSS is re-engineering its data distribution systems, and consulting with NOAO and MAST to assure it is doing so in a way that brings the software closer to being transferable to those institutions or to others depending on their prospects for long-term maintenance of the archiving and distribution systems. This project will inform us on how to proceed in the longer term with ensuring access to SDSS data and other data sets for the future.

For the sake of legacy access to SDSS and to other data sets, a priority for the 2020s needs to be the establishment of a long-term ground based archiving center for public data sets produced by national and private facilities. This facility would lower costs, provide more efficient execution of science by the community, and ensure legacy access to astronomical data over the very long term. NOAO (or NCOA) would serve as a natural host for such a facility, and is in a position to solve these pressing problems at a national level. We write to encourage NOAO to build such a center into its strategic planning for the next decade.