VOL. 104 | NO. 4
APRIL 2023

Eos

SCIENCE NEWS BY AGU

Engineering with Nature

Starry, Starry Nights

Gaming-Based Virtual Field Trips

DATABASE UPDATES

Scientists reassess dated data in the time of the Cloud.

AGU
ADVANCING EARTH AND SPACE SCIENCE

DELUGES OF DATA Are Changing Astronomical SCIENCE

BY KATHERINE KORNEI

Astronomers today are more likely than ever to access data from an archive rather than travel to a telescope—a shift that's democratizing science.

In this composite image of the Tarantula Nebula, the blue and purple patches represent X-ray data from the Chandra X-ray Observatory, and the red and orange gas clouds, which look like roiling fire, represent infrared data from the James Webb Space Telescope. Credit: X-ray: NASA/CXC/Penn State Univ./L. Townsley et al.; IR: NASA/ESA/CSA/STScI/JWST ERO Production Team

Radio telescopes in Chile—part of the Atacama Large Millimeter/submillimeter Array (ALMA)—point toward the night sky. Credit: ESO/C. Malin, CC BY 4.0 (bit.ly/ccby4-0)

For scientists who study the cosmos, hard-to-grasp numbers are par for the course. But the sheer quantity of data flowing from modern research telescopes, to say nothing of the promised deluges of upcoming astronomical surveys, is astounding even astronomers. That embarrassment of riches has necessitated some serious data wrangling by myself and my colleagues, and it's changing astronomical science forever.

Gone are the days of the lone astronomer holding court at the telescope. Modern astronomy is most decidedly a team sport, with collaborations often spanning multiple institutions and particularly large scientific endeavors regularly producing papers with more than a hundred coauthors. And rather than looking through an eyepiece, like astronomers of yore, researchers today collect an enormous array of observations across the electromagnetic spectrum, from X-rays to radio waves, using sophisticated digital detectors. In recent years, scientists have also probed the universe using gravitational waves—an advance made possible by exquisitely sensitive instrumentation.

With research-grade telescopes peppered across all seven continents—and also in space—there's no shortage of astronomical data. And thanks to advances in detector technology, cosmic data are being collected more rapidly, and at a higher density, than ever before. The challenge now is storing and organizing all of those data and making sure they're accessible and useful to a wide variety of scientists around the world.

Bringing the Data Home

Only a few decades ago, just about everyone engaged in professional observational astronomy would have traveled to a telescope to collect their own data. That's what Chuck Steidel, an astronomer at the California Institute of Technology, remembers doing as a graduate student in the 1980s. Between 1984 and 1989, he made four trips by himself to Chile.

Steidel's destination was Las Campanas Observatory, where he used a telescope to observe "quasi-stellar objects," intensely bright and distant astronomical bodies believed to be powered by supermassive black holes. To transfer the astronomical data that he collected back to his home institution for analysis, Steidel recorded them onto dinner plate–sized magnetic storage tapes known as 9-track tapes.

Each observing run generated a lot of tapes to haul back to the United States, said Steidel. "A weeklong observing run would take about 24 of these, or two boxes, weighing about 40 pounds each." The load was too bulky to bring with him on an airplane, however, so Steidel had to ship the tapes back to the United States via boat, a process that took several weeks.

Around the time Steidel began advising graduate students of his own in the mid-1990s, technology had marched on, and

magnetic cassette tapes were in use for data storage. The palm-sized disks held far more data than 9-track tapes, and they weren't nearly as cumbersome to transport. It was suddenly possible to carry telescope data home immediately after an observing run, said Alice Shapley, an astronomer at the University of California, Los Angeles who joined Steidel's group as a graduate student in the late 1990s.

By the late 2000s, when I was a graduate student in astronomy working with Shapley, digital video discs (DVDs) were the preferred medium for transporting astronomical data. I remember leaving Hawaii's W. M. Keck Observatory one morning bleary from a lack of sleep but content to have my observations literally in hand on thin disks that I could slip into my carry-on luggage.

My experiences in graduate school differed from those of my adviser and her adviser in more than just the ways in which we transported our data, however. Steidel obtained all of the data for his thesis by traveling alone to a telescope. Shapley also collected much of her thesis data herself, but she supplemented her observations with data provided by other members of her adviser's research group. I, on the other hand, gathered a significant portion of my data from astronomical archives.

Data for Everyone

The concept of a data repository for astronomical observations is relatively new. It was just over 2 decades ago that the Sloan Digital Sky Survey (SDSS) started amassing data from a modest-sized telescope in southern New Mexico and making those observations available in the form of a catalog, said Ani Thakar, a computational astronomer at Johns Hopkins University in Baltimore, and a catalog archive scientist with SDSS. "Before SDSS made [its] data public to the world, there was nothing like it," he said.

During its first phase of operations, from 2000 to 2005, SDSS increased the number of known galaxies from 200,000 to 200 million. "It ushered in the era of big data in astronomy," said Thakar. SDSS is still going strong today; it recently celebrated its eighteenth data release, and the archive now includes observations of nearly half a billion unique objects. From developing high-quality processing pipelines to building server-side analysis tools, the goal has always been to streamline data storage and access and provide high-quality observations that are useful to the scientific community, said Thakar.

Many more astronomical archives exist today. The Mikulski Archive for Space Telescopes (MAST), managed by the Space Telescope Science Institute in Baltimore, is one of the largest. MAST contains images, spectra, and other forms of observations from more than 20 telescopes and space missions. Some of those data—amounting to several petabytes in all—were gathered by individual scientists observing specific celestial objects; others were obtained as part of systematic sky surveys.

> "IT'S A WAY TO DISCOVER DATA SETS."

The point of amassing all of those data in a searchable archive is to help ensure that they're useful to the larger scientific community in perpetuity, according to David Rodriguez, an astronomical data scientist at the Space Telescope Science Institute and a classmate of mine from graduate school. "We collect and archive all of that information and make it available to everyone," he said.

No longer are observations gathered by a researcher the sole purview of that researcher and their collaborators forever—instead, they're often archived and released to the public after some predetermined proprietary period (typically 12 months). That democratic access to data is changing astronomical science.

Astronomers once stored telescope data on 9-track tapes. Credit: Hannes Grobe, Wikimedia, CC BY-SA 2.5 (bit.ly/ccbysa2-5)

The ability to pluck existing data from an archive can be a godsend for researchers working on a timeline. I know that firsthand—I was able to access Hubble Space Telescope data, which were critical to both my master's and doctoral theses, from archives rather than having to write applications to use the telescope, which is heavily oversubscribed. (In the most recent round of proposals for so-called General Observer programs with the Hubble Space Telescope, astronomers asked for more than 5 times the amount of telescope time available.)

Particularly for early-career scientists seeking to finish a dissertation or establish themselves in a research track, applying for telescope time is a stressful experience fraught with uncertainty. Having access to archival data means that it's no longer necessary to travel to a telescope, a potentially expensive and time-consuming endeavor. (However, some telescopes, like those at the W. M. Keck Observatory in Hawaii, can be remotely accessed.)

The resources needed to apply for, and collect, telescopic observations in the traditional way can be substantial. It's therefore not surprising that researchers based in countries with a lower gross domestic product per capita tend to produce a larger fraction of publications based on archival data

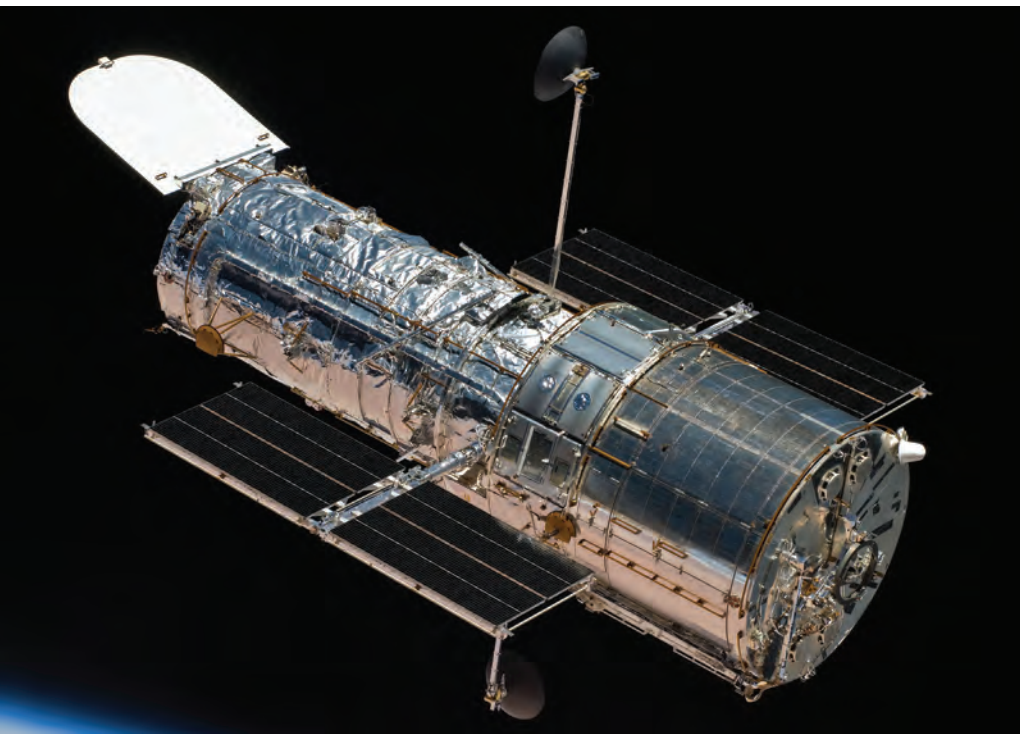
than researchers living in more affluent countries.

Astronomical archives clearly provide more equitable access to data, but they're valuable for another fundamental reason, too: They open up new research avenues. The very act of digging through a data repository often turns up unexpected observations that might have been taken years ago and that a researcher didn't know existed, said Rodriguez. "It's a way to discover data sets." Those data could prove useful for current or future research projects or even spur entirely new investigations, he said.

Organized and Accessible

A key tenet of any archive is that its data are well organized and accessible. That's where Rodriguez plays a key role: He helps standardize all of the metadata—for instance, the date of the observation, the name of the object being observed, and its sky coordinates—associated with astronomical observations in MAST. "I work toward consolidating the various types of metadata we have across all missions," said Rodriguez. The goal is to ensure that data from different telescopes and space missions can be easily and uniformly queried in the MAST database, he explained.

"THE LEGACY OF AN OBSERVATORY DEPENDS ON HOW MUCH PEOPLE USE THE ARCHIVED DATA."



The Hubble Space Telescope has been observing the universe since 1990. Credit: NASA, Public Domain

Ample data show that archival observations are being put to use. Hundreds of scientific papers are published each year using data from MAST, and that number has increased by more than a factor of 2 since the early 2000s.

A separate archive devoted to just one astronomical observatory—the Atacama Large Millimeter/submillimeter Array (ALMA), an ensemble of radio telescopes in the Atacama Desert of Chile—has seen similar successes. Data from ALMA are funneled into the ALMA Science Archive for public access after a 12-month proprietary period.

Adele Plunkett, an astronomer working with the ALMA Science Archive, said that it's easy to access the observations, which number in the tens of millions of files and total more than a petabyte. "You don't even need to create an account. You can just go to our website, and you can start browsing and downloading the data," she said.

Plunkett and her colleagues have shown that roughly 3 times more data are downloaded by users each month than are taken in anew from ALMA. That's evidence that users are accessing substantial amounts of archival data, said Plunkett. "Many people are able to access the same projects and therefore can maximize the utility of observations."

And scientists are publishing results using those archival data. In 2021, roughly 30% of ALMA-based publications incorporated archival observations, the team found. That's a significant increase from 10% just a decade ago, and it's something to be proud of, said Plunkett. "The legacy of an observatory depends on how much people use the archived data."

Wrangling large quantities of archival data takes not only technical expertise but also an eye toward how people interact with a user interface. Several of Plunkett's colleagues have backgrounds in user experience. "We think a lot about the design of the archive and the usability of it," she said. The team often takes a cue from other online platforms that involve searchable databases. "We look at Amazon and Netflix and online retailers," said Plunkett.

Archives of the Future

The next generation of telescopes is currently being developed in tandem with the next generation of data archives. Those facilities have the advantage of coming of age in a world primed for big data, said Rodriguez. One example is the Vera C. Rubin Observatory in Chile, which is slated to collect several tens of petabytes' worth of

This image of a portion of Messier 92, one of the brightest globular clusters in the Milky Way, was created with data captured by the James Web Space Telescope's Near Infrared Camera, or NIRCam. Credit: NASA, ESA, CSA, Alyssa Pagan (STScI)

images of the night sky. "They're starting from modern data technology," said Rodriguez. "They're starting cloud ready."

Beginning in 2024, the Simonyi Survey Telescope at the Vera C. Rubin Observatory will image the entire visible sky about every 3 days and will continue doing so for a decade. That massive undertaking, known as the Legacy Survey of Space and Time (LSST), will not only provide a comprehensive look at billions of stars and galaxies but also reveal how transient objects such as asteroids and supernovas vary in brightness over time, said Leanne Guy, the LSST data management project scientist at the Vera C. Rubin Observatory. "Because we can observe the sky so rapidly, we can see things changing," she said.

The observations of the LSST will essentially produce an evolving picture of the cosmos. "It will be the greatest movie of the night sky," Guy said. Not surprisingly, there will be a whole lot of data involved; the LSST will yield roughly 20 terabytes of raw data every night. Those data—in the form of images obtained at wavelengths ranging from ultraviolet to near infrared—will be transferred from Chile to the SLAC National Accelerator Laboratory in California. From there, they'll be distributed to other data-processing facilities around the world, and the final data products will be made available via Google Cloud Platform.

A set of web applications known as the Rubin Science Platform will allow users to access, view, and analyze LSST data. That's a shift away from the traditional model, in which scientists download data to their computer, said Guy. But that change is necessary, she said, because it allows researchers to efficiently mine petabyte-scale data sets. "It is no longer feasible for scientists to just download a data set to their computer and load it into memory," said Guy.

> "IT WILL BE THE GREATEST MOVIE OF THE NIGHT SKY."

The deluge of astronomical observations now available to anyone with an Internet connection is changing how research is being done and even what's being researched. As scientists embrace the tools of "big data," they're able to dig into far-flung research questions that couldn't have been answered just a few decades ago, like how galaxies are arranged in space.

And graduate students around the world are already writing theses based largely, and sometimes wholly, on archival data; more than 300 astronomy Ph.D. theses have been written to date using SDSS data. Time will tell whether the experience of observing at a telescope will go the way of the dodo. Probably not, but astronomical archives are obviously here to stay.

Welcome to the era of archives.

Author Information

Katherine Kornei (@KatherineKornei), Science Writer

▶ Read the article at bit.ly/Eos-astronomical-data