Computing in SCIENCE & ENGINEERING

www.computer.org/cise

Vol. 10, No. 1
January/February 2008

SDSS SCIENCE ARCHIVE

The Sloan Digital Sky Survey

Drinking from the Fire Hose

The Sloan Digital Sky Survey Science Archive represents a thousand-fold increase in the total amount of data that astronomers have collected to date. The pioneering instrumentation technology that made this possible is matched by groundbreaking tools that let anyone in the world access terabytes of SDSS data online.

The Sloan Digital Sky Survey (SDSS) is a multi-institution, internationally funded project to map out half of the northern sky in unprecedented detail with a dedicated 2.5-m telescope and special-purpose instruments.[1] The project saw "first light" in 2000 and completed its observations in July 2006. The successor to the first project—SDSS-II—has since been funded by an expanded consortium and just issued its first data release in June 2007; it's due to complete in late 2008.

The SDSS is by far the most ambitious sky survey project to date, and one of the biggest challenges it poses is the sheer size of the resulting data archive. Prior to the SDSS, the total number of galaxies for which astronomers had detailed digital data was on the order of 200,000—after SDSS-I, this number jumped to more than 200 million. A thousand-fold increase in volume presents an additional challenge in terms of distributing the resulting data to a wide community of users. From a user's viewpoint, it's like drinking from a fire hose: unless you know what you're doing, you'll be inundated with data.

This article gives an overview of the project; the rest of the articles in this issue focus on data distribution to astronomers and the general public.

A Huge Science Archive

The digital data archive—officially called the SDSS Science Archive—contains the distillation of the calibrated scientific data from both the SDSS-I and -II surveys. The raw image data is expected to be roughly 15 Tbytes in size, and the archive's catalog data is expected to be roughly 5 to 6 Tbytes by the time SDSS-II finishes in late 2008. SDSS-II actually contains two additional subsurveys—SEGUE (Sloan Extension for Galactic Understanding and Exploration) and the Supernova Survey—in addition to the continuation of the original data (called the Legacy Survey in SDSS-II). Both the raw and catalog data sets are publicly distributed for online access, as mandated by the US National Science Foundation (see the "Relevant URLs" sidebar for the Web site and other useful resources).

Starting with the Early Data Release (EDR) in 2001, the SDSS has had six public data releases to date for SDSS-I and one data release for SDSS-II, as summarized in Table 1.[2–7] The master archive is hosted at Fermilab, but the table lists worldwide

1521-9615/08/$25.00 © 2008 IEEE
Copublished by the IEEE CS and the AIP

Ani R. Thakar
Johns Hopkins University

THIS ARTICLE HAS BEEN PEER-REVIEWED.

Table 1. Sloan Digital Sky Survey public data releases.

Release	Date	Size (catalogs)	Images (millions)	Spectra (thousands)	Distribution/mirrors
EDR[2]	June 2001	200 Gbytes	14	54	Johns Hopkins Univ. (JHU), San Diego Supercomputing Center (SDSC), UK, Japan
DR1[6]	June 2003	1 Tbyte	53	186	JHU, SDSC, Canadian Astronomical Data Centre (CADC), Univ. of Pittsburgh, UK, Germany, Japan, India
DR2[5]	Mar. 2004	2 Tbytes	88	330	JHU, Univ. of Pittsburgh, SDSC, Germany
DR3[3]	Sept. 2004	3 Tbytes	141	478	JHU, UK, India
DR4[4]	June 2005	4 Tbytes	180	608	JHU, Germany, Hungary, Brazil
DR5[6]	June 2006	5 Tbytes	215	738	JHU, India, Russia, Hungary, Australia
DR6[7]	June 2007	6 Tbytes	287 M	1.27 M	JHU, India, Hungary

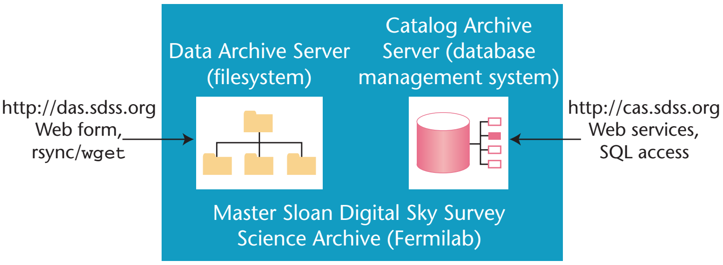

Figure 1. Accessing Sloan Digital Sky Survey data. Of the two portals, the Data Archive Server (DAS) provides access to raw image files and other binary data products, whereas the Catalog Archive Server (CAS) provides access to the catalog data in a database management system.

mirror sites; the SDSS data itself is described in detail elsewhere in the literature.[2–4]

The SDSS Science Archive data set comes in two different formats: raw (binary) image data and distilled and calibrated science parameters extracted by SDSS pipelines[8,9] into a catalog data archive (see Figure 1). The former is of interest mainly to professional astronomers, whereas the catalog data is meant to serve everyone, including professional astronomers, amateur astronomers, and interested members of the lay public.

Let's look more closely here at the SDSS data itself and the processing pipeline the data passes through before it goes into the SDSS Science Archive's databases.

SDSS Data Overview

The SDSS camera collects raw data in the form of Flexible Image Transport System (FITS) files from a 2.5-m telescope located at the Apache Point Observatory (APO) in New Mexico; APO staff then send this data via FedEx to the master SDSS archive site at Fermilab in Illinois after each observing run. The SDSS observing team at APO conducts observing runs during the 20 or so most photometric nights per year. These runs collect spectroscopic data on selected targets when observing conditions aren't optimal for imaging.

The SDSS has two data archives: the operational database (OpDB) and the science database (the SDSS Science Archive itself). Raw data is first stuffed into the OpDB and then "resolved," calibrated, and exported to the SDSS Science Archive via pipelines. FITS data is converted to ASCII comma-separated value (CSV) format before being ingested into the CAS database tables, for example.

Best, Target, and Runs Data Sets

By far, the largest data component is the imaging data for hundreds of millions of astronomical objects. Data taken at the telescope is processed in several complex pipelines[2,8,9] that calibrate and test the data for quality assurance. After calibration, this data passes through a target selection pipeline that selects targets for spectroscopic follow-up. Roughly 1 percent of the imaging objects are selected for spectroscopy based on the following broad criteria:[2]

- galaxies brighter than a certain limit;
- luminous red galaxies (LRGs) based on colors and brightness out to a certain redshift (distance from the Earth), and
- quasars based on their SDSS colors and detection at radio frequencies in other surveys.

Data can also be reprocessed because of bugs or enhancements to the pipeline code.

Relevant URLs

- The main Sloan Digital Sky Survey Web site (www.sdss.org) gives a good overview of the project as well as an entry point to all other SDSS services. Append "DR*n*" to this URL, where *n* is a release number, and you can get details about the data for a particular data release (for example, DR6).
- The main entry point into the SDSS is the Catalog Archive Server (CAS), which provides Web access to the SDSS Science Archive databases (http://skyserver.sdss.org or http://cas.sdss.org).
- The SDSS Data Archive Server (DAS; http://das.sdss.org) lets you download Flexible Image Transport System (FITS) image files and other raw data products.
- CasJobs (http://cas.sdss.org/casjobs/) is a more advanced Web portal to the CAS, and provides users with their own workbench database, called MyDB.
- The Open SkyQuery (www.openskyquery.net) is a distributed query and cross-matching service that federates SDSS and other National Virtual Observatory (NVO) archives' data.
- The entry point to virtual observatory services for SDSS and other data is at www.voservices.org.

Table 2. List of data products available in the Sloan Digital Sky Survey Science Archive.*

Data product	Format	Size	Comments	Server
Object catalog	DBMS	6.5 Tbytes	Parameters of > 10^8 objects	CAS
Redshift catalog	DBMS	0.5 Tbytes	Parameters of > 10^6 objects	CAS
Object catalog	FITS	2.5 Tbytes	Parameters of > 10^8 objects	DAS
Redshift catalog	FITS	300 Gbytes	Parameters of > 10^6 objects	DAS
Atlas images	FITS	4.0 Tbytes	Five color cutouts of > 10^8 objects	DAS
Spectra	FITS	300 Gbytes	One-dimensional form	DAS
Derived (value-added) catalogs	CSV, DBMS	100 Gbytes	Clusters, QSO absorption lines, quasar catalog	DAS, CAS
Binned images, corrected frames	FITS	6.0 Tbytes	Processed image data per field	DAS

** DBMS: database management system; FITS: Flexible Image Transport System; QSO: quasi-stellar object; CAS: Catalog Archive Server; DAS: Data Archive Server; and CSV: comma-separated value.*

As such, more than one "version of the sky" as seen by the SDSS is available and stored in the OpDB:

- The *Runs* sky version is the first run through the photometric pipeline and is often useful for SDSS collaboration members to study survey performance. Available only internally, it isn't part of the public data releases.
- The *Target* version of the sky is a snapshot of the processed data on which the target selection pipeline was run. Preserving it is essential for properly analyzing spectroscopy data, especially for cosmological studies.
- The *Best* sky version is the latest and greatest processing of the data, the one the collaboration thinks best represents the final product.

Each public SDSS data release contains the Best and Target data sets.

SEGUE and Supernova Data Sets

SDSS-II adds two more data sets to those usually released with each public data release: the SEGUE and Supernova Survey data sets. SEGUE data is typically available at both the DAS and CAS portals, whereas Supernova data at the time of this writing is available only from a DAS-like interface. In the future, it will also be available from the CAS.

SDSS Data Products

The SDSS Science Archive contains several distinct data products, including photometric and spectral catalogs, a redshift catalog, and images, spectra, and maps. Its data model is optimized for fast access, with other data products indirectly accessible. Table 2 gives a complete list of data products available through the SDSS Science Archive. The raw data is saved in a tape vault at Fermilab.

The SDSS project and the Science Archive it has produced represent a leap forward for astronomy: this project has essentially increased the number of astronomical objects with high-quality digital data by a factor of a thousand in a few short years. The rest of this special issue describes how people connected with the project have dealt with the unique and daunting challenges that the SDSS presents.

Acknowledgments

Funding for the SDSS and SDSS-II was provided by the Alfred P. Sloan Foundation, the US National Science Foundation, the US Department of Energy, NASA, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions, which include the American Museum of Natural History, the Astrophysical Institute Potsdam, the University of Basel, the University of Cambridge, Case Western Reserve University, the University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, the Ohio State University, the University of Pittsburgh, the University of Portsmouth, Princeton University, the US Naval Observatory, and the University of Washington.

References

1. A. Szalay et al., "Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey," *Proc. ACM SIGMOD*, ACM Press, 2000, pp. 451–462.
2. C. Stoughton et al., "Sloan Digital Sky Survey: Early Data Release," *The Astronomical J.*, vol. 123, no. 1, 2002, pp. 485–548.
3. K. Abazajian et al., "The Third Data Release of the Sloan Digital Sky Survey," *The Astronomical J.*, vol. 129, no. 3, 2005, pp. 1755–1759.
4. J. Adelman-McCarthy et al., "The Fourth Data Release of the Sloan Digital Sky Survey," *The Astrophysical J. Supplement Series*, vol. 162, no. 1, 2006, pp. 38–48.
5. K. Abazajian et al., "The Second Data Release of the Sloan Digital Sky," *The Astronomical J.*, vol. 128, no. 1, 2004, pp. 502-512.
6. J. Adelman-McCarthy et al., "The Fifth Data Release of the Sloan Digital Sky Survey", *The Astronomical J. Supplement Series*, vol. 172, no. 2, 2007, pp. 634-644.
7. J. Adelman-McCarthy et al., "The Sixth Data Release of the Sloan Digital Sky Survey," to be published in *The Astrophysical J. Supplement Series*, 2007.
8. R. Lupton et al., "The SDSS Imaging Pipelines," *Proc. SPIE*, vol. 4836, J.A. Tyson and S. Wolff, eds., SPIE, 2002, pp. 350–356
9. M. Subbarao et al., "The Sloan Digital Sky Survey 1-Dimensional Spectroscopic Pipeline," *Proc. SPIE*, vol. 4847, J.-L. Starck and F.D. Murtagh, eds., SPIE, 2002, pp. 452–460.

Ani R. Thakar *is a research scientist at the Johns Hopkins University. His research interests include science with large databases and interacting galaxies. Thakar has a PhD in astronomy from the Ohio State University. Contact him at thakar@jhu.edu.*

CiSE's editors *extend their special thanks to Sebastian Jester at the Max Planck Institute for Astronomy for managing the peer-review process on the SDSS articles in this issue.*

In Tribute

The Sloan Digital Sky Survey Catalog Archive Server wouldn't have been such a success story without the tremendous and fundamental contributions of Jim Gray. Since his original involvement with the SDSS data in 2000, Jim contributed his unparalleled expertise in the database field, his inexhaustible enthusiasm and can-do spirit, and a great deal of his precious time and hard work to the cause of bringing the SDSS data to the masses.

The SkyServer was a labor of love for Jim. Several years ago, he said, "the average Wal-Mart manager has a better database than the average astronomer," and he made it his mission to change that fact. Jim was also one of the first to recognize the value of the SDSS data to computer scientists and database designers.

Over the years, Jim not only convinced Microsoft SQL Server engineers to use the SDSS as a test data set on which to hone their products, but also invited interested groups in academia and industry to port the large SDSS data set to other platforms.

Jim's untimely disappearance in early 2007 was a great blow not only to the SDSS but to the worldwide scientific community that he influenced and worked closely with in his tireless quest to promote the use of databases in science. One of Jim's favorite aphorisms was, "You have only two things two fear: failure and success." Thanks in very large measure to him, today we're victims of our own success. —*Ani R. Thakar*

ALSO Building Better Search Engines, p. 7
Why Fortran? p. 68
Designing a Cluster, p. 72
July/August 2007
Computing in Science & Engineering
Computing Climate Change