

The Sloan Digital Sky Survey Archive

Brian Yanny Fermilab May 16, 2011



It is essential for the advancement of astrophysics that data be preserved in a readable, understandable format for long periods of time (hundreds of years).

Examples:

1. Babylonian estimate (format: clay tablets of length of the month:

Mean length of Synodic(new moon to next new moon) month:

29;31,50,8,20d [base60] (29.530594 days),
vs. modern value: 29.530590d

[Babylonians knew it to 2 parts in 10 million because they had centuries of records].

Example 2: Precession of the equinoxes (Age of Aquarius, Scorpios are really Libras, etc)
(format: clay tables, papyrus records)

This system of specifying positions is quite stable, but not perfect. Around 150 BC the Greek astronomer Hipparchus carefully compared his own observations of certain stars with observations of the same stars recorded by Timocharis 169 years earlier (and with some even earlier measurements from the Babylonians), and noted a slight but systematic difference in the longitudes. Of course, these were all referenced to the supposedly fixed direction of the line of intersection between the Earth's rotational and orbital planes, but Hipparchus was led to the conclusion that this direction is not perfectly stationary, i.e., that the direction of the Sun at the equinoxes is not constant with respect to the fixed stars, but precesses by about 0.0127 degrees each year. This is a remarkably good estimate, considering the limited quality of the observations that were available to Hipparchus. The accepted modern value for the precession of the equinoxes is 0.01396 degrees per year, which implies that the line of the equinoxes actually rotates completely around 360 degrees over a period of about 26,000 years.

From "Reflections on Relativity" by Kevin Brown

Example 3: Precession of Perihelion of Mercury (format: paper in books and journals)

The general flavor of Le Verrier's work is a complicated analytical fit to the observational Mercury data over about 50 years. This consisted of deriving, "Condition Equations," containing seven types of terms, some of which are periodic and others time-dependent, plus approximate perturbations of these terms to obtain equations that match the observations with minimal residuals. The final product is a series of Tables predicting when Mercury would be observed in the future. A byproduct of the analysis was the discovery that the contributions of the other planets left an unexplained residual shift of the perihelion by 39 arc-seconds/century.

From a monograph by Roger A. Rydin

The answer, which he triumphantly announced in the third of his four November lectures, came out right: 43 arc-seconds per century.⁷⁸

“This discovery was, I believe, by far the strongest emotional experience in Einstein’s scientific life, perhaps in all his life,” Abraham Pais later said. He was so thrilled he had heart palpitations, as if “something had snapped” inside. “I was beside myself with joyous excitement,” he told Ehrenfest. To another physicist he exulted: “The results of

Mercury’s perihelion movement fills me with great satisfaction. How helpful to us is astronomy’s pedantic accuracy, which I used to secretly ridicule!”⁷⁹

Einstein as told to Sommerfeld, from Issacson's
Biography of Einstein

Viewpoint: Right Ascension 12 hours, Decl. 55 degrees, Distance 0
Looking towards: Right Ascension 12 hours, Decl. 55 degrees, Distance 99999



Dubhe

Mizar

Alioth

Merak
Ursa Major

Alkaid

Phad

Canes Venatici

α -2 CVn

Leo Minor

Viewpoint: Right Ascension 12 hours, Decl. 55 degrees, Distance 0
Looking towards: Right Ascension 12 hours, Decl. 55 degrees, Distance 99999
Time since present: 100000 yr



Image from 2005 Sep 12

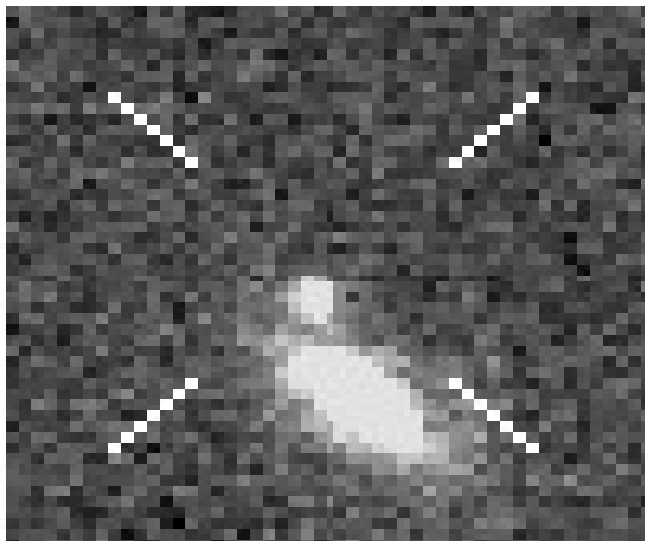
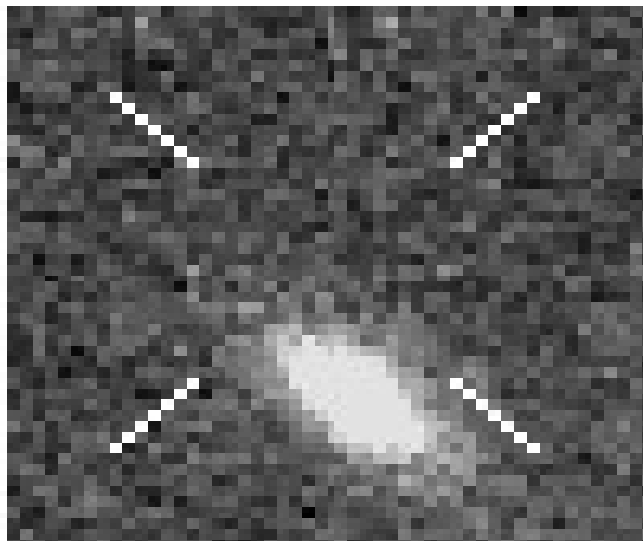
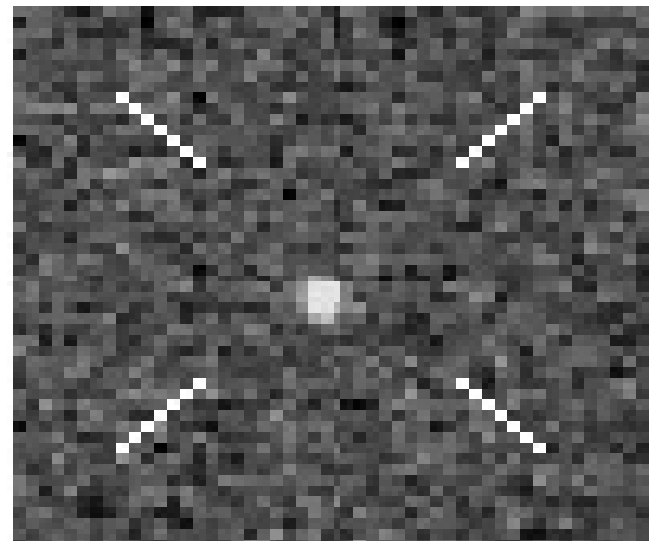


Image from 2004



Difference image,
Showing Supernovae



Example from the SDSS Supernovae survey to measure the accelerating expansion of the universe (format: digital FITS images).

The lasting theoretical interpretation or discovery is often done by someone other than the observer, (sometimes several lifetimes later!).

Thus it is important that the data be in an accessible format.

Observations which are not possible (say X-ray wavelength), need to be matched against optical wavelength observations from an earlier epoch in order to 'finish the project'.

The SDSS data archive:

Archive contents/volume: SDSS DR7 CAS (10TB) and DAS (70 TB)

ENSTORE Tape backup of files and copy of archive format CAS

Raw/unprocessed data

Nightly Log information

(processed) Images, including Object Catalogs

(processed) Spectra, including Object Catalogs

Calibration telescope raw and processed data

Software repository (source code to go from raw to processed data)

Auxiliary calibration and reference catalogs (astrometric, photometric, spectrophotometric, dust catalog)

Documentation on file formats, data processing instructions

Value Added catalogs (rearrangements of the data and further processing of subsets of the data by interested scientists)

Published papers about the data

Older processing(s) of the (same) raw data (not necessary, but good if you have room)

Eric Neilsen (SDSS/FNAL) collected basic documentation, and generated A 'flat file' simple web-based front end to the SDSS DATA Archive (DAS)



SDSS Data Archive Server

This is the SDSS Data Archive Server (DAS) main page. For general information on the SDSS, please consult the [SDSS homepage](#). For more information on using SDSS data, please consult the [SDSS DR7 web site](#).

The SDSS Data Archive Server (DAS) serves files produced and consumed by the SDSS data processing pipelines. See [this overview](#) of the pipelines for a description of the pipelines, references to more detailed information on them, and tables of what files are available in the DAS and where to find them. For the online SQL database and more advanced interactive data exploration tools, see the [Catalog Archive Server \(CAS\)](#).

The DAS itself provides direct access to the [directory tree](#) with the data, [interactive forms](#) that allow users to upload tables of data of interest for exploration of bulk download, and web pages for browsing [data release 7](#) and [older data releases](#).

Interactive download tools

SDSS Run 1889, Rerun 40, Camcol 4 (35S)

Files for the run as a whole

	Directory	File
Astrometric calibration	1889/40/astrom	asTtrans-001889.fit
Photometric calibration	1889/40/nfcalib	fcPCalib-001889-4.fit
Summary QA	QA/1889/40/qa	summary-runQA-1889-40.html
Full QA	QA/1889/40/qa	all-runQA-1889-40.html
Frames QA	QA/1889/40	qaFrames.html
Astrometric calib. QA	QA/1889/40	qaAstrom.html
Photometric calib. QA	QA/1889/40	qaNfcalib.html
PSP QA	1889/40/objcs	pspQA-001889.html

Files for each field


Field	Zoom	Corrected frames	Binned frames	Masks	Others
40		u fpC-001889-u4-0040.fit.gz g fpC-001889-g4-0040.fit.gz r fpC-001889-r4-0040.fit.gz i fpC-001889-i4-0040.fit.gz z fpC-001889-z4-0040.fit.gz	u fpBIN-001889-u4-0040.fit g fpBIN-001889-g4-0040.fit r fpBIN-001889-r4-0040.fit i fpBIN-001889-i4-0040.fit z fpBIN-001889-z4-0040.fit	u fpM-001889-u4-0040.fit g fpM-001889-g4-0040.fit r fpM-001889-r4-0040.fit i fpM-001889-i4-0040.fit z fpM-001889-z4-0040.fit	<p style="text-align: center;">Calibrated object catalogdrObj-001889-4-40-0040.fit</p> <p style="text-align: center;">Field calibration and statisticsdrField-001889-4-40-0040.fit</p> <p style="text-align: center;">Atlas imagesfpAtlas-001889-4-0040.fit</p> <p style="text-align: center;">Preliminary calib. and final PSF fitspsField-001889-4-0040.fit</p> <p style="text-align: center;">Uncalibrated object catalogfpObjc-001889-4-0040.fit</p>
41		u fpC-001889-u4-0041.fit.gz g fpC-001889-g4-0041.fit.gz r fpC-001889-r4-0041.fit.gz i fpC-001889-i4-0041.fit.gz z fpC-001889-z4-0041.fit.gz	u fpBIN-001889-u4-0041.fit g fpBIN-001889-g4-0041.fit r fpBIN-001889-r4-0041.fit i fpBIN-001889-i4-0041.fit z fpBIN-001889-z4-0041.fit	u fpM-001889-u4-0041.fit g fpM-001889-g4-0041.fit r fpM-001889-r4-0041.fit i fpM-001889-i4-0041.fit z fpM-001889-z4-0041.fit	<p style="text-align: center;">Calibrated object catalogdrObj-001889-4-40-0041.fit</p> <p style="text-align: center;">Field calibration and statisticsdrField-001889-4-40-0041.fit</p> <p style="text-align: center;">Atlas imagesfpAtlas-001889-4-0041.fit</p> <p style="text-align: center;">Preliminary calib. and final PSF fitspsField-001889-4-0041.fit</p> <p style="text-align: center;">Uncalibrated object catalogfpObjc-001889-4-0041.fit</p>
42		u fpC-001889-u4-0042.fit.gz g fpC-001889-g4-0042.fit.gz r fpC-001889-r4-0042.fit.gz i fpC-001889-i4-0042.fit.gz z fpC-001889-z4-0042.fit.gz	u fpBIN-001889-u4-0042.fit g fpBIN-001889-g4-0042.fit r fpBIN-001889-r4-0042.fit i fpBIN-001889-i4-0042.fit z fpBIN-001889-z4-0042.fit	u fpM-001889-u4-0042.fit g fpM-001889-g4-0042.fit r fpM-001889-r4-0042.fit i fpM-001889-i4-0042.fit z fpM-001889-z4-0042.fit	<p style="text-align: center;">Calibrated object catalogdrObj-001889-4-40-0042.fit</p> <p style="text-align: center;">Field calibration and statisticsdrField-001889-4-40-0042.fit</p> <p style="text-align: center;">Atlas imagesfpAtlas-001889-4-0042.fit</p> <p style="text-align: center;">Preliminary calib. and final PSF fitspsField-001889-4-0042.fit</p> <p style="text-align: center;">Uncalibrated object catalogfpObjc-001889-4-0042.fit</p>
		u fpC-001889-u4-0043.fit.gz	u fpBIN-001889-u4-0043.fit	u fpM-001889-u4-0043.fit	<p style="text-align: center;">Calibrated object catalogdrObj-001889-4-40-0043.fit</p>

Basic Documentation about the SDSS

File Edit View History Bookmarks Tools Help

http://www.sdss.org/ Google

Sloan Digital Sky Survey



Sloan Digital Sky Survey

Mapping the Universe

- Home
- SDSS-III
- SDSS Data DR8
- SDSS Data DR7
- Science
- Press Releases
- Education
- Image Gallery
- Legacy Survey
- SEGUE
- Supernova Survey
- Collaboration
- Publications
- Contact Us
- Search

The Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) is one of the most ambitious and influential surveys in the history of astronomy. Over eight years of operations (SDSS-I, 2000-2005; SDSS-II, 2005-2008), it obtained deep, multi-color images covering more than a quarter of the sky and created 3-dimensional maps containing more than 930,000 galaxies and more than 120,000 quasars.

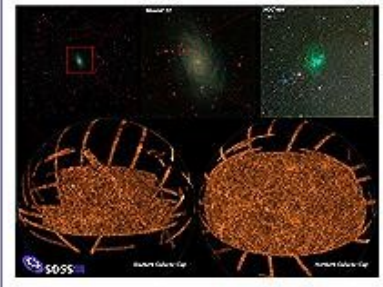
SDSS data have been released to the scientific community and the general public in annual increments, with the final public data release from SDSS-II occurring in October 2008. That release, [Data Release 7](#), is available through this website.

Meanwhile, SDSS is continuing with the Third Sloan Digital Sky Survey (SDSS-III), a program of four new surveys using SDSS facilities. SDSS-III began observations in July 2008 and released its first public data as [Data Release 8](#) to emphasize its continuity with previous SDSS releases. SDSS-III will continue operating and releasing data through 2014.


Data Release 8 contains all images from the SDSS telescope - [the largest color image of the sky ever made](#). It also includes measurements for nearly 500 million stars and galaxies, and spectra of nearly two million. All the images, measurements, and spectra are available free online. You can [browse through sky images](#), look up [data for individual objects](#), or [search for objects](#) anywhere in the sky based on any criteria.

Images of the SDSS

(click for more information)



The Final Survey



The Whirlpool Galaxy (M51)

CAS (Catalog archive server), an SQL-database interface to The SDSS catalog of over 350 million stars, galaxies and quasars.

Your SQL command was:

```
SELECT TOP 10
p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,
p.run, p.rerun, p.camcol, p.field,
s.specobjid, s.specClass, s.z,
s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
p.u BETWEEN 0 AND 19.6
AND g BETWEEN 0 AND 20
```

objid	ra	dec	u	g	r	i	z	run	rerun	camcol	field	specobjid	specClass	z	plate	mjd	fiberid
587725074995609782	147.32950414	0.02890341	17.963612	16.401266	15.575814	15.132483	14.847518	1239	40	4	180	75094093117521920	2	0.048089	266	51630	36
588848898824274111	146.51281644	-0.84577181	17.416025	16.201288	15.535498	15.084999	14.817842	756	44	1	205	75094093180436480	2	0.064857	266	51630	51
588848898824274160	146.57132235	-0.95719938	18.535383	17.050571	16.264538	15.798815	15.460427	756	44	1	205	75094093138493440	2	0.065066	266	51630	41
588848899898343534	147.32102993	-6.57628E-3	19.269053	18.788694	18.923374	19.10795	19.312033	756	44	3	210	75094093121716224	1	-2.27177E-4	266	51630	37
587725074458673337	147.17639124	-0.3540289	18.435514	17.251513	16.737843	16.500662	16.299618	1239	40	3	179	75094093079773184	2	6.32467E-3	266	51630	27
587725074458738699	147.25535095	-0.31932367	16.94639	16.502197	16.638393	16.812626	17.01622	1239	40	3	180	75094093075578880	1	-1.69719E-5	266	51630	26
587725073921474764	146.44833941	-0.71342989	19.012621	17.927675	17.341618	16.884993	16.702686	1239	40	2	174	75094093193019392	2	0.114681	266	51630	54
587725073921343533	146.1303268	-0.65936706	17.73105	17.747757	17.983236	18.189171	18.412859	1239	40	2	172	75094093377568768	1	1.0862E-4	266	51630	98
587725073921343577	146.19261402	-0.68841994	17.323654	16.169165	15.788399	15.68924	15.611839	1239	40	2	172	75094093381763072	2	4.13096E-3	266	51630	99
587725073921278122	145.93787679	-0.73395785	19.375374	17.69453	16.830414	16.393473	16.080275	1239	40	2	171	75094093515980800	2	0.142806	266	51630	131

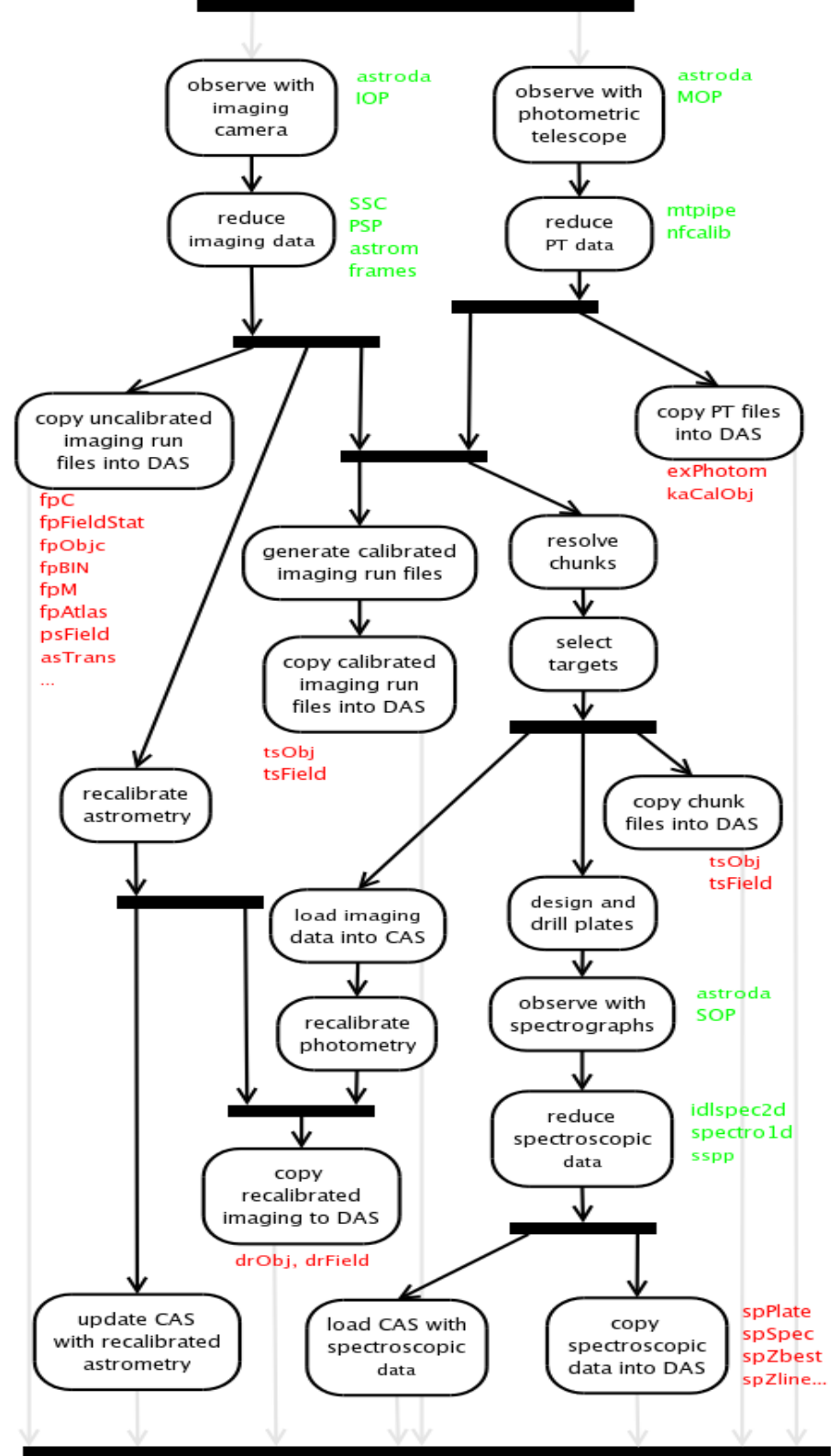
Use the button(s) below to upload the results of the above query to the DAS and retrieve the corresponding FITS files:

Upload list of fields to DAS

Upload list of spectra to DAS

Documentation example
Of how the SDSS software data
Processing pipeline work.

With data model, available at
<http://das.sdss.org>



Long term store copies of data to:

UofChicago Library (DAS and DR7 CAS)

Johns Hopkins Library (DAS and DR5 CAS)

ENSTORE tape robot (FITS files in .tar balls) at FNAL (has already been
Converted to new tape formats at least twice)

Active working copies of data to:

3 spinning (RAID disk) back ends of DR7 at FNAL, 3 front-end web servers

University of Portsmouth (DR7 CAS/CASJOBS)

Johns Hopkins Astrophysics (DR7 CAS/CASJOBS) [also DR8]

We believe unofficial copies of DR7 exist at UofWashington, Google, China, India, elsewhere.

What has helped us:

1. Standardized data storage formats: FITS, ASCII
 - Self-documenting (header keywords in fits)
 - Easily readable 'with type/cat'
 - Format itself documented, with tools to read the format if needed
 - Relatively compact (but not compressed)
 - Relatively simple (rectangular array tables)
2. Documentation with the data
3. Software source code for data processing
4. Published papers
5. Multiple identical copies distributed around the world
6. Library resources (librarians!) who take the data and documentation pretty much as-is and store it, and look for completeness and understandability.

What has hindered us:

1. Lack of documentation (code authors have moved on). Not budgeted for up front.
2. Lack of simple examples of step-by-step process for going from raw data to processed catalogs/images (lack of documentation)
3. Proprietary software or storage formats; active passwords Stored in documentation preventing distribution
4. No budget for copies built in up front.
5. Overly complex, impenetrable (after a few years) data formats (or compression algorithms)

Notes:

SDSS had major support from Fermilab's Computing division throughout the project in maintaining the data archive. Toward the end of SDSS-II (2008), the project director and manager (Kron and Boroski) made contact with Johns Hopkins Library and University of Chicago Library to plan for longer term storage. This was a forward-thinking move. Some funds were also put aside to continue support for the archive at Fermilab through 2013.

We were not able to fully reproduce the 'working data processing pipelines' in code – only 'in principle' with text and source code documentation

The next big astronomical survey which the Fermilab Experimental Astrophysics group is participating in is the Dark Energy Survey (DES). It is similar to the imaging portion of the SDSS, only in the Southern Hemisphere (different visible sky), and roughly two magnitudes deeper. The total data volume will be approximately 5x that of SDSS (100 TB images and catalogs, perhaps 250 TB total data volume). Long term plans are not yet set for DES, but we hope to build on our SDSS experience.

Steve Kent notes:

Astronomical surveys, including the predecessors of SDSS/DES/LSST, have a long history of archiving and preserving data for future use. [It should be noted that, in contrast to particle physics experiments, the subject matter of astronomical surveys - the sky - has a permanence that makes archival data of much greater value than might otherwise be the case. However, even data for transient objects, such as supernovae, are now archived such that they can be jointly reanalyzed along with new data sets.] Also, most data sets are sufficiently straightforward in nature (images of the sky, catalogs of objects) that one does not need detailed knowledge of the detectors or Monte Carlo simulations to make use of them.

The life expectancy of data can be decades or centuries, making the technical aspects of data preservation and dissemination an interesting challenge. Methods and practices are evolving continuously. Historically, astronomical survey data were published in the form of printed catalogs or observatory reports and were only replicated in limited numbers. Imaging data, if distributed at all, were done so as photographic plates or prints. The transition to digital technology has led to some curious paradoxes - even though digital technology can store and distribute data far more economically and widely, the lifetime of digital media is considerably shorter than that of the older, analog technologies. Further, the exponential growth in data storage capability means that one's strategies for the largest current and future surveys such as SDSS/DES/LSST will change during the lifetime of the survey. For example, at the beginning of SDSS, the data distribution mechanism was imagined to be on CDs, and only a small volume of the total dataset. By the end, all data were stored online and distributed by network, including all raw and processed data.

The astronomical field has defined a simple yet flexible standard file format (FITS) for storing images and tabular data, and all data archives make use of this format for flat files. Metadata included in the files provide basic provenance and calibration information. The SDSS, e.g., stored all binary data using the FITS format. Additionally, object catalogs are stored in an SQL database. (For these, there is no standard format or organization). Beyond the information stored in the files or databases themselves, there is extensive textual documentation of the file contents and the overall survey properties. For SDSS and DES, access to the data is provided to both the flatfiles and the catalog databases, in both cases by means of web interfaces. For the flat files, a form-based interface is provided to support simple queries that return a desired subset of the files. The server-side interface was designed to be portable and self-contained so it could be migrated to new architectures in the future. For the catalog databases, all access is again via web-based forms, but additionally one can provide straight SQL strings to implement complex queries. The database and web interface are sufficiently complex that one relies on the vendor (Microsoft) to support the server-side architecture going into the future.

For SDSS, public data releases were made annually and thus represented a snapshot of the data collected to that point and a particular version of the processing software. The last data release is the "final" release that forms the long-term archive. No reformatting or simplification was attempted - the public receive data in the same format as did the collaboration.

For SDSS, all datasets are maintained "live" on spinning disk. A complete backup copy is stored in a tape robot. For disaster mitigation, two additional copies have now been created offsite at university libraries, including replicas of the catalog databases and server. The long-term support plan for the archive is not yet established.

Regarding analysis software, all the SDSS software source code (except for one proprietary library) as was used for the final data release has been published. There is no support for compiling and running the code - the intent is that the code documents the algorithms used, not that it can be used to reprocess the data.

It is worth mentioning that there are multiple efforts to define standardized web-service-based interfaces to online archives - the umbrella organization for the efforts is the "International Virtual Observatory Alliance" (IVOA). DES has the intention of using these interfaces as part of its plan to provide public access to DES data, although none have yet been implemented.

Summary:

1. Long term storage (greater than one human lifetime) essential for progress in The astrophysical sciences.

2. Data formats should be convertible (to new type of tape or disk), simple (rectangular table, don't overuse compression), non-proprietary, and documented.

3. Raw data should (usually) be kept with code and instructions on how it was processed.

4. Keep at least one most-recent version of processed data, don't need all intermediate data.

5. Support your local (and remote) librarians. Keep multiple copies at different locations (recall Alexandria).