Jun 22nd, 3:30 PM - 4:30 PM

The SDSS and e-science archiving at the University of Chicago Library

Barbara Kern
University of Chicago, bkern@uchicago.edu

Dean Armstrong
University of Chicago, dean@uchicago.edu

Charles Blair
University of Chicago, chas@uchicago.edu

David Farley
University of Chicago, d-farley@uchicago.edu

Kathleen Feeney
University of Chicago, kefeeney@uchicago.edu

See next page for additional authors

Presenter Information
Barbara Kern, Dean Armstrong, Charles Blair, David Farley, Kathleen Feeney, Eileen Ielmini, Elisabeth Long, Daniel Meyer, and Peggy Wilkins

THE SDSS AND E-SCIENCE ARCHIVING AT THE UNIVERSITY OF CHICAGO LIBRARY

Barbara Kern
The University of Chicago Library, USA, bkern@uchicago.edu

Dean Armstrong
The University of Chicago Library, USA, dean@uchicago.edu

Charles Blair
The University of Chicago Library, USA, chas@uchicago.edu

David Farley
The University of Chicago Library, USA, d-farley@uchicago.edu

Kathleen Feeney
The University of Chicago Library, USA, kefeeney@uchicago.edu

Eileen Ielmini
The University of Chicago Library, USA, eielmini@uchicago.edu

Elisabeth Long
The University of Chicago Library, USA, elong@uchicago.edu

Daniel Meyer
The University of Chicago Library, USA, arch@uchicago.edu

Peggy Wilkins
The University of Chicago Library, USA, mozart@uchicago.edu

Abstract
The Sloan Digital Sky Survey (SDSS) is a co-operative scientific project involving over 25 institutions worldwide and managed by the Astrophysical Research Consortium (ARC) to map one-quarter of the entire sky in detail, determining the positions and absolute brightness of hundreds of millions of celestial objects. The project was completed in October 2008 and produced over 100 terabytes of data comprised of object catalogs, images, and spectra. While the project remained active, SDSS data was housed at Fermilab. As the project neared completion the SDSS project director (and University of Chicago faculty member) Richard Kron considered options for long term storage and preservation of the data turning to the University of Chicago Library for assistance. In 2007-2008 the University of Chicago Library undertook a pilot project to investigate the feasibility of long term storage and archiving of the project data and providing ongoing access by scientists and educators to the data through the SkyServer user interface. In late 2008 the University of Chicago Library entered into a formal agreement with ARC agreeing to assume responsibility for:

- Archiving of the survey data (long-term scientific data archiving)
- Serving up survey data to the public
- Managing the HelpDesk
- Preserving the SDSS Administrative Record

This paper outlines the various aspects of the project as well as implementation.

The University of Chicago Library is excited to be involved in archiving and providing access to one of the most influential astronomical surveys ever conducted, and to be preserving the project's print and electronic administrative records. A co-operative scientific project involving more than 25 institutions worldwide, the Sloan Digital Sky Survey (SDSS) is managed by the Astrophysical Research Consortium (ARC) and funded by the Alfred P. Sloan Foundation of New York City, the National Science Foundation (NSF), the U.S. Department of Energy and the member institutions. During its eight years of operation (SDSS-I, 2000-2005; SDSS-II, 2005-2008) the project mapped one-quarter of the sky including 930,000 galaxies and more than 120,000 quasars. SDSS amassed over 100 terabytes of data, including spectra, images, and object catalogs. The impact of the public release of SDSS data has been wide ranging. Over 2,000 articles have been published based on SDSS data and have generated more than 70,000 citations in scholarly papers. At the same time, the popular SkyServer interface has made SDSS images and data accessible to the general public and has spawned interactive projects such as Galaxy Zoo that engages the public in helping to classify the shapes of galaxies captured in the survey.

A hallmark of this archiving project has been the way in which the SDSS project incorporated planning for the long-term viability of the data into its operations, committing both staff time and money to ensure a smooth hand-off of data as part of the project close out. In 2006, two years before the survey was scheduled to come to an end, University of Chicago faculty member and project director of SDSS-II Professor Richard Kron asked the Library to consider taking a role in providing both long term storage of project data and ongoing data access for scientists and educators through an SDSS interface. Since the start of SDSS in 2000, Fermi National Accelerator Laboratory (Fermilab) - a high energy physics facilities funded by Office of Science of the U.S. Department of Energy and located in Batavia, Illinois - was responsible for processing, storing and serving the data. While Fermilab wished to continue hosting and serving the data, there was recognition that archiving data was not a primary function of the Laboratory. Like other research libraries, the University of Chicago Library is concerned with preserving information. Our mission includes "ensur[ing] the preservation and long-lasting availability of Library collections and resources" [University of Chicago Library 2004]. In addition, there was a recognized need to address load-balancing and disaster recovery issues by sharing SDSS data both geographically and institutionally. By storing the data at multiple institutions a redundancy is created that protects the data in case of a power failure or other catastrophic event at one institution. In addition to Fermilab and the University of Chicago Library, Johns Hopkins University is involved in serving and archiving the data.

At the University of Chicago Library we have understood for some time that we need to better understand our current and future roles in data archiving efforts. Discussions like the one we had with Professor Kron about SDSS are now becoming familiar for us. Various research projects at the University of Chicago are accumulating vast amounts of data that remain valuable even after projects are completed, and questions of how to preserve and provide access to data are becoming paramount.

The SDSS provided an excellent opportunity for us to test our ability to preserve data from a large project. This involved considering investments in staff time, expertise, computer hardware, and requisite software. Questions of sustainability related to storage capacity and data migration were also considered. In addition, as we discussed the scope of the project and learned the significance of the SDSS project as a whole, we developed an interest in not only the data, but also in the print and electronic administrative records of the project. These records not only provide invaluable insights for future historians of SDSS project, but also provide core resources for understanding the context of the collection and processing of the data itself. Thanks to funding from ARC and the John Crerar Foundation, plus Library staff time and resources which we contributed to the effort, we were able to move forward with this project.

Our participation in this project can be divided into four sections: the Catalog Archive Server (CAS), the Digital Archive Server (DAS), the Administrative Records and the SDSS Help Desk.

The CAS provides access to the SDSS data through a series of databases accessible through an internet interface. This interface lets users search SDSS processed data with an extensive series of parameters. The CAS utility to researchers is that broad analysis of hundreds of millions of objects is available via simple SQL searches. It saves the tedious nature of data reduction from thousands of scientists and makes possible statistical associations from the data that would be otherwise be inaccessible.

The databases run under Microsoft SQL Server 2005, and the internet interface is IIS 6 running on Windows 2003 Server 32-bit.

Approximately 20TB in size, the CAS databases include:
1. TargDR7, containing all objects
2. BestDR7, for multiply-observed, containing only the best seeing data
3. SEGUE, an extension into the local Milky Way stellar fields
4. Stripe82, a repetitive and deep scan across the South Galactic Cap
5. RunsDB, objects outside the standard stripes and specialized areas

The interface allows for searching via a variety of methods, including point-and-click on a skymap; searching via RA/Dec coordinates or radius searches; or by direct SQL queries. Parameters in the database include Gunn ugriz magnitudes, redshift, date and time, SDSS stripe numbers, and spectral parameters, among many others. Objects with spectra offer a quick look-up spectrum with major lines identified.

Prior to our commitment to store the SDSS data, we conducted a successful pilot project with earlier data releases of the CAS (DR5). The pilot was designed to test our ability to install the data, maintain the servers, and make the data accessible to researchers. Having gained this experience, we were confident in moving forward with archiving the final data release of the CAS (DR7), and next turned to addressing specific cost issues. New hardware for hosting the CAS DR7 was estimated to cost $20,000-$25,000 depending on the specific server configuration. Instead of purchasing new hardware, we arranged to accept a transfer of existing SDSS DR7 servers from Fermilab to Chicago, and to reuse servers previously running the CAS DR5 pilot project. The internet server hardware is composed of two internet 1U and 2U OEM servers, both with 2GB of RAM and a 250GB RAID-1 hard drive mirror, one with a AMD Opteron 1.7GHz dual core CPU and the other with a Intel Xeon 2.8GHz CPU. The database hardware consists of four OEM 4U servers with similar specifications to the internet servers but with space for 24 SATA drives in each unit. The drives are connected to 3 3Ware RAID controllers in each server, with 8 drives per controller, each configured as a RAID-5 array. Each array is then combined as one large contiguous volume in the Windows OS. All the equipment is in one 42U rack and powered via APC Smart-UPSes models 3000XL and 2200. Gigabit connections to the Internet are provided.

The hardware/software management component of the project is managed by The University of Chicago Library's Administrative Desktop Systems Department. This group administers the Windows servers running CAS, provides security patching and general hardware maintenance.

The DAS includes almost 80 TB of processed data in flat file format. Composed of fits images, spectra and catalog tables, the DAS provides access to both a directory and interactive forms that allow the upload of data tables and web pages for browsing survey data products for all releases.

An ongoing role of the University of Chicago Library as curator for the DAS will be to ensure the long-term preservation (archiving) of the digital data. The equipment we selected to store the DAS is from Sun Microsystems (now part of Oracle). Our choice was based on the fact that we wanted the system to be as close as possible to a turnkey system or information appliance. The purpose of selecting a turnkey system was to help ensure that we could maintain close to 80TB of data over

the lifetime of the equipment. Recently we completed copying the DAS data from Fermilab's servers over the internet, a process which took many months due to the large amount of data. After confirming that the copy is complete and accurate, we will begin performing routine, automated checksums of the data to ensure data integrity.

When considering Library staff resources needed to archive the DAS we leveraged existing staff, and did not need to shift existing priorities or add staff. We treated this project as we would any other digital library development activity. In this we succeeded.

The DAS component of the project is managed by The University of Chicago Library's Digital Library Development Center.

The SDSS Administrative Records were an attractive addition to the University of Chicago's Special Collections Research Center (SCRC). The SCRC is the principal repository and steward of the University of Chicago Library's rare books, manuscripts, University Archives and the Chicago Jazz Archives. The mission of SCRC is to provide primary sources to stimulate, enrich, and support research, teaching and learning at the University of Chicago. As part of the University's engagement with the larger community of scholars and independent researchers, the SCRC also strives to make its resources available to a broader constituency.

For SCRC, the SDSS Administrative Records are invaluable research material in their own right, in part because they document the collaborative work of University of Chicago affiliated scholars and institutions. The SDSS records complement SCRC's holdings in the history of astronomy, astrophysics and allied sciences, including the personal and professional papers of Subrahmanyan Chandrasekhar, E. E. Bernard, W. W. Morgan, John A. Simpson, David Schramm and other scientists. Additionally, the records of University of Chicago departments and affiliated institutions, such as the Yerkes Observatory, are also part of these records. With the support of the American Institute of Physics, the SCRC has engaged in several recent projects to make these materials broadly accessible to researchers. This includes archival processing of the Simpson and Schramm Papers and digitization of almost 5,000 documentary photos from the records of the Yerkes Observatory.

Working with Professor Kron and Bill Boroski from Fermilab, Daniel Meyer, Eileen Ielmini, and Kathleen Feeney of SCRC identified and secured a portion of administrative records of SDSS. The records were identified through discussions with staff from SDSS-affiliated institutions and on- site visits to Fermilab and the Apache Point Observatory. The majority of the physical records received by SCRC were provided by Professor Kron and University of Chicago professor and SDSS founding director Donald York, plus Bill Boroski, Valena Sibley and John Galvan from Fermilab. A small amount of material was also received from the Apache Point Observatory. The records include engineering materials, meeting minutes, correspondence, photographs, construction documents, scrapbooks, posters and publicity materials, videotapes, CDs, floppy disks and other electronic storage media, and project proposals.

Additionally, SCRC has also received a collection of electronic files documenting SDSS operations from Robert Lupton of Princeton University. These include a "survey snapshot" of the SDSS listservs, an archive SDSS' bug-tracking GNATS database, and a snapshot of the project website along with associated documents.

The materials collected by SCRC are defined as SDSS administrative records because they generally document project operations rather than the scientific data collected. While the collection in its current form documents a wide array of SDSS activities and functions, the inter-institutional nature of the project, coupled with the challenges of preserving born-digital materials over time, make it unlikely that a comprehensive administrative record of the project will be assembled in any single archival repository.

In many ways, the SDSS Administrative Records are typical of modern collections received by SCRC. They contain materials in a variety analog and digital formats ranging from paper documents to videotapes to born-digital files delivered on CD or hard drive. The records present a special challenge in that the most comprehensive and complete portion of the collection is largely digital. Moreover, this collection was received in the form of digital "tarballs" - uninventoried and undescribed documents. An associated challenge is ensuring the long-term accessibility of files created or accessed through specialized software not widely used outside the astronomical community. This includes materials stored on obsolete media like older floppy disks. The hybrid paper and electronic records of SDSS are representative of those produced by contemporary research projects, and their receipt by SCRC provides an opportunity to develop tools and policies in its role as a steward of these records.

The non-digital portions of the SDSS records have been inventoried. With the exception of a small number of files currently restricted due to budgetary or personnel material content, these are accessible to researchers at the SCRC. The digital records received from Professor Lupton have been deposited in the Library's digital repository. They are currently the subject of a collaborative project between SCRC and the Digital Library Development Center. In the short term, this project is designed to establish procedures for discovery, description and preservation of electronic records. In the long term, the goal of this project is to develop an interface for unmediated researcher access to unrestricted files.

All of records will be arranged and described according to established archival principles. A guide describing the contents of the collection will be made available through the Archives andManuscripts Finding Aids Database http://ead.lib.uchicago.edu/. This database aid contains guides to more than 1000 collections held by SCRC. The finding aid will describe both the physical and the digital portions of the SDSS records. While access to materials in the digital repository currently requires mediation by Library staff, it is hoped that researchers will eventually be able to access electronic materials directly, by means of links provided within the online finding aid.

The SDSS Help Desk was previously managed by staff at Fermilab. However, with the closing of the SDSS project and Fermilab's overall involvement in the project reduced, the Library accepted management of the Help Desk operations. The Library already has a system in place for distributing email queries, and it was felt that this aspect of the project fell well within the Library's core competencies.

Prior to making a final commitment to managing the Help Desk we conducted a 6 month pilot that was completed in January 2008. The pilot provided ample opportunity for us to understand staff resources required, as well whether or not our existing email query management tool could be used for SDSS. Questions that come in through the SDSS Help Desk are substantially different from questions that the library typically handles. The primary purposes of the SDSS Help Desk are to assist users with the SDSS interfaces, and to help users interpret downloaded results. Responders need an educational background in the field of astronomy/astrophysics and while some librarians do have this qualification this was not the case at the University of Chicago Library. Even if the qualifications had been present, we would have been concerned about sustainability. If one individual has the knowledge and skills but leaves their position at the Library, it may become difficult to find someone equally able to respond to Help Desk queries. Ultimately, the Library SDSS Help Desk was modeled after the model established by Fermilab.

In this model, the SDSS Help Desk at Fermilab relied on a group of experts located both at Fermilab and at other institutions to answer user's questions. The University of Chicago's John Crerar Library Reference Department, which has assumed responsibility for managing the Help Desk, relies solely on experts at other institutions to respond to questions and provide subject knowledge. The list of experts was provided by Fermilab and ARC, and includes individuals from Princeton, Johns Hopkins, the United States Navy, Ohio State University, and the University of Washington.

The Library uses KnowledgeTracker, an email query management system by Compendium. All questions are directed to a single account (SDSS/Crerar) that is monitored and managed by staff. We ask individuals submitting questions to the Help Desk to self-identify a category for their question (CAS, DAS, Photometry, Astrometry, Spectroscopy, Publication permissions and policy, Educational Exercises). For each category there are 2-3 experts available to respond to questions. By asking users to self-identify a category it helps ensure that questions are distributed effectively. To respond to a question, experts reply to an email sent by staff via KnowledgeTracker. Staff then sends the response to the patron. Additionally, there are several experts who use KnowledgeTracker, and will respond directly and respond to patron queries.

The SDSS Help Desk receives approximately 10 questions per week, and requires approximately 5 hours per week of Reference Department staff time.

Our participation in archiving and serving the SDSS data, preserving the Administrative Record and managing the Help Desk has been a success to date. In 2008 we signed a Memorandum of Understanding with ARC outlining our role in the project and setting an initial 5 year timeline. In December 2013 we will consider how to move forward keeping in mind that long term storage is our goal. Significant issues we anticipate needing to address include migration of data from old media to new media and the complexity of the CAS interface to the data which relies on what will be increasingly outdated software and technology platforms. Understanding the life cycles of the various components of the SDSS archive and their relevance to scholars and the general public will help us create effective strategies for the ongoing maintenance of the archive.

The Library is looking forward to continuing our work with SDSS and with other data curation projects in the future.

References

University of Chicago Library (2004). *Library Mission, Vision and Values* Retrieved from www.lib.uchicago.edu/e/about/mvv.html