

# Public Institution Perspective (Research Library)

DMASM 2014  
Sayeed Choudhury



**Data**Conservancy



# Outline

- A model for describing storage, archiving, preservation and curation
- Our definition of preservation
- The Sloan Digital Sky Survey (SDSS)
- Similarities between our problems, challenges and opportunities
- Grand challenges (or opportunities)



# Data Management Layers

Layers	Characteristics	Implication for PI	Implication relative to NSF
Curation	Adding value throughout life-cycle	<ul style="list-style-type: none"><li>• Feature Extraction</li><li>• New query capabilities</li><li>• Cross-disciplinary</li></ul>	<ul style="list-style-type: none"><li>• Competitive advantage</li><li>• New opportunities</li></ul>
Preservation	Ensuring that data can be fully used and interpreted	<ul style="list-style-type: none"><li>• Ability to use own data in the future (e.g. 5 yrs)</li><li>• Data sharing</li></ul>	<ul style="list-style-type: none"><li>• Satisfies NSF needs across directorates</li></ul>
Archiving	Data protection including fixity, identifiers	<ul style="list-style-type: none"><li>• Provides identifiers for sharing, references, etc.</li></ul>	<ul style="list-style-type: none"><li>• Could satisfy most NSF requirements</li></ul>
Storage	Bits on disk, tape, cloud, etc. Backup and restore	<ul style="list-style-type: none"><li>• Responsible for:<ul style="list-style-type: none"><li>• Restore</li><li>• Sharing</li><li>• Staffing</li></ul></li></ul>	<ul style="list-style-type: none"><li>• Could be enough for now but not near-term future</li></ul>



# Definition of Data Preservation

- “Data preservation involves providing enough representation information, context, metadata, fixity, etc. such that someone other than the original data producer can use and interpret the data.”
  - Ruth Duerr, National Snow and Ice Data Center



Sloan Digital Sky Survey

Sloan Digital Sky Survey

+

www.sdss.org

☆


Google

Q

★

↓

🏠



# Sloan Digital Sky Survey

## Mapping the Universe

Home

SDSS-III

SDSS Data

DR10

SDSS Data DR9

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 [Data Release 8](#) in January 2011, [Data Release 9](#) in August 2012, and [Data Release 10](#) in July 2013. SDSS-III will continue operating and releasing data through 2014.

[Data Release 10](#) contains the first release of APOGEE infrared Galactic spectroscopy as well as cumulative updates to the BOSS optical extragalactic spectroscopy archive.

[Data Release 9](#) contains the first release of BOSS spectroscopy to the public as well as several significant updates to the cumulative SDSS archive.

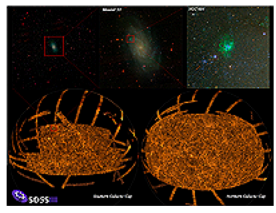
[Data Release 8](#) contains all images from the SDSS telescope - [the largest color image of the sky ever made](#). It also includes measurements for nearly 500 million stars and galaxies, and spectra of nearly two million. All the images, measurements, and spectra are available free online. You can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

The SDSS used a dedicated 2.5-meter telescope at Apache Point Observatory, New Mexico, equipped with two powerful special-purpose instruments. The 120-megapixel camera imaged 1.5 square degrees of sky at a time, about eight times the area of the full moon. A pair of spectrographs fed by optical fibers measured spectra of (and hence distances to) more than 600 galaxies and quasars in a single observation. A custom-designed set of software pipelines kept pace with the enormous data flow from the telescope. The two key technologies that enabled the SDSS, optical fibers and the digital imaging detectors known as CCDs, were the discoveries awarded the [2009 Nobel Prize in Physics](#).


During its first phase of operations, 2000-2005, the SDSS imaged more than 8,000 square degrees of the sky in five optical bandpasses, and it obtained spectra of galaxies and quasars selected from 5,700 square degrees of that imaging. It also obtained repeated imaging (roughly 30 scans) of a 300 square degree stripe in the southern Galactic cap.

### Images of the SDSS

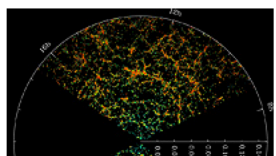
(click for more information)



#### The Final Survey



#### The Whirlpool Galaxy (M51)







# SDSS has...

- Raw Data
- Data Archive Server (DAS)
- Catalog Archive Server (CAS)
- Software
  
- Web-Based Data Documentation
- Publications
- Administrative Archive
  
- Content
  - Sloan Digital Sky Survey (SDSS) – Phase I & II
    - ~160 TB in ~80 million files
  - Researcher Content
    - Typically 5-200 GB in hundreds to thousands of files per article



# SDSS Lessons Learned

- <https://wiki.library.jhu.edu/x/eY1XAQ>
- <https://wiki.library.jhu.edu/display/sdss/sdss-lessons-learned>
- Tracking lessons based on data management layer stack model





# Similarities (1)

- SDSS scientists were pioneers who pushed the data envelope...
- ...without accounting for preservation
- While they represent the leading edge, increasingly science moves in this direction
- Existing data archives may not be well prepared to preserve new forms of science



## Similarities (2)

- Traditional approaches to preservation will maintain some record but not the full experience or story
- The use of machines, computers, code, operating systems, etc. is fundamentally changing the requirements
- Selection criteria is not only influence by potential value, but also potential processing needs in the future (provenance becomes critical)



## Similarities (3)

- While your industry may have different access requirements, there may be areas of overlap
  - Film becomes part of cultural heritage
  - Independent film makers, small archives are similar to scientists working with video
- At some core level, technology requirements may start to converge



# Grand Challenges

- Or are they opportunities?
- Influence storage and cloud vendors to meet preservation requirements in a transparent manner
- Demonstration of successful format migration



# Acknowledgements

- NSF Award OCI-0830976
- Sheridan Libraries and JHU financial support
- Tim DiLauro for SDSS slides
- Alex Szalay for Levels of Data slide
- <http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster192.pdf>
- <https://www.youtube.com/watch?v=F6iYXNvCRO4> -- data management layer stack model
- <http://dataconservancy.org>
- <http://www.dlib.org/dlib/september12/mayernik/09mayernik.html> -- DC blueprint document