Data is the New Currency

SLA-AGC 2014

Sayeed Choudhury

DataConservancy

Data Conservancy Objectives

- Data Conservancy is a community that develops solutions for data preservation and sharing to promote cross-disciplinary re-use.

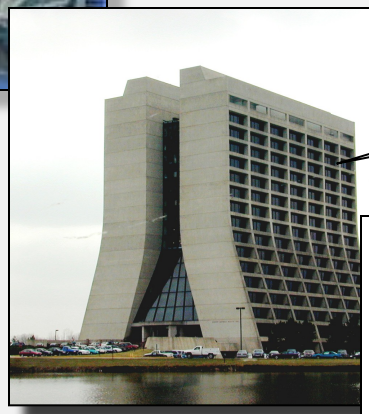
- Preserve – collect and take care of research data
- Share – reveal data's potential and possibilities
- Discover – promote re-use and new combinations

- Culmination of over a decade of experience with Sloan Digital Sky Survey (SDSS) data

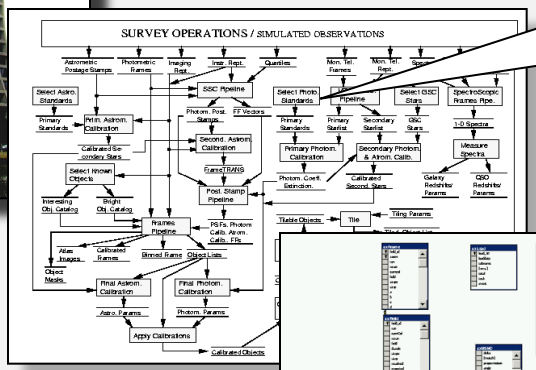
SDSS Data Flow (Levels of Data)

Pixel data collected by telescope

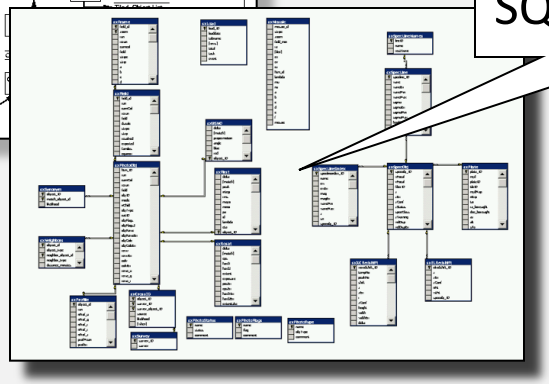
Sent to Fermilab for processing



Beowulf Cluster produces catalog



Loaded in a SQL database



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

Long Tail Researchers have...

Anything you could imagine.*

* And probably some that you could not

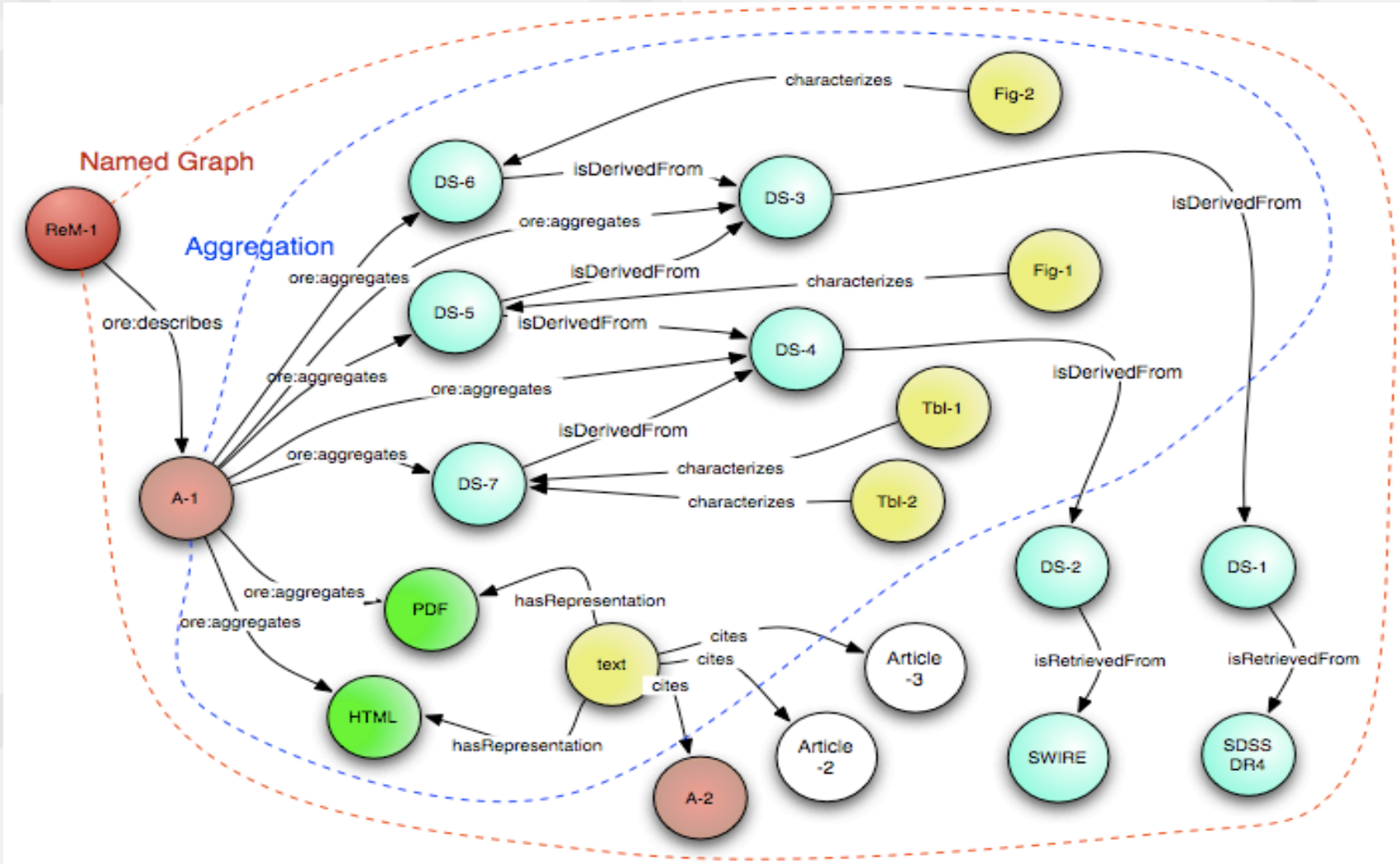
Side-by-Side

Concern	SDSS*	Long Tail Researchers
Sophistication	Greater expertise due to economies of scale.	Little or no expertise or sophistication.
Data Organization	Well structured and documented. Organized around data publication.	Ad hoc, but often organized around and triggered by article publication.
Formats	Community standard.**	Ad hoc.
Quality Control	Yes. And well documented. Strong project and community feedback.	Undocumented. Usually left to grad students with some minimal review at publication time.
Code management	Primarily CVS, but difficult to determine which version produced a given output.	Rare. Undocumented modifications are common.
Incentives	Ambivalent. Investigated only toward end of project.	Few. But more coming. More disincentives.

* For released data

** But overloaded

Information graph using OAI-ORE



Understanding Infrastructure: Dynamics, Tensions, and Design

Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures"

Paul N. Edwards
Steven J. Jackson
Geoffrey C. Bowker
Cory P. Knobel

January 2007

NSF Grant 0630263 • Human and Social Dynamics • Computer and Information Science and Engineering • Office of Cyberinfrastructure

…not a rigid road map but principles of navigation. There is no one way to design cyberinfrastructure, but there are tools we can teach the designers to help them appreciate the true size of the solution space – which is often much larger than they may think, if they are tied into technical fixes for all problems.

Currency

From Merriam-Webster dictionary –

- “the money that a country uses : a specific kind of money”
- “the quality or state of being current”
- “circulation as a medium of exchange”

Data as Money or Current

- "Data is the new oil" (stated in Qatar, European Commission, etc.)
- Data is the fourth factor of production (McKinsey)
- McKinsey estimates potential $3 trillion of economic value across seven sectors within US alone
- Todd Park estimates location sensitive apps generate $90 billion of value annually
- White House Office of Science and Technology Policy Executive Memorandum
- White House Open Government Initiative

Data as a Medium for Exchange

- The research community is composed of a diverse set of sub-cultures with their own rules of engagement and conduct. They are, however, bonded through a common vision, goal and purpose.
- Data are the means of exchange through which they can move beyond their sub-culture differences toward a common infrastructure and shared community.

Data as a Medium for Exchange

- The European community is composed of a diverse set of sub-cultures with their own rules of engagement and conduct. They are, however, bonded through a common vision, goal and purpose.
- Data are the means of exchange through which they can move beyond their sub-culture differences toward a common infrastructure and shared community.

Four Means of Exchange for EU

- Flow of goods and services (free trade)
- Flow of money (Euro)
- Flow of people (through work permits, immigration)
- Flow of research and ideas (open data)

Data as a Medium for Exchange

- The Arabian Gulf region community is composed of a diverse set of sub-cultures with their own rules of engagement and conduct. They are, however, bonded through a common vision, goal and purpose.
- Data are the means of exchange through which they can move beyond their sub-culture differences toward a common infrastructure and shared community.

Data as the New Currency

- Data are one mechanism by which we can build connections and shared culture.
- Libraries are cultural heritage institutions.
- "Stories are data with a soul."
  - Brené Brown
- "…Ultimately, preservation is about keeping people's stories alive for future generations."
  - Sayeed Choudhury

Acknowledgements

- NSF Award OCI-0830976
- Sheridan Libraries and JHU financial support
- Tim DiLauro for SDSS slides
- Alex Szalay for Levels of Data slide
- http://dataconservancy.org
- http://dmp.data.jhu.edu -- JHU Data Management Services
- http://www.dlib.org/dlib/september12/mayernik/09mayernik.html -- DC blueprint document
- SDSS lessons learned - https://wiki.library.jhu.edu/display/sdss/sdss-lessons-learned