

SPD 10 DataNet: Data Conservancy (UCLA-DC).

SPD 10.1 Overview

The Data Conservancy is a five-year project lead by Johns Hopkins University, funded by NSF's DataNet initiative by the Office of Cyberinfrastructure. A primary goal of DC is development of a repository that will archive data and facilitate collaborative access for a range of sciences. Its design will be based upon social science research of data curation and collaboration practices for science communities who are likely users of the DC repository. UCLA's researchers, administered through CENS, will conduct research on practices and data curation requirements for astronomy and astrophysics. The outcomes of their research will contribute to the design of the Data Conservancy system architecture.

SDP 10.2 Approach

UCLA Data Conservancy (UCLA-DC) is coordinating with social science teams at The Center for Informatics Research in Science and Scholarship (CIRSS) at University of Illinois, Urbana-Champaign, and the National Center for Atmospheric Research (NCAR), who are studying data practices and needs for a range of science communities expected to contribute their data to the Data Conservancy and who will be users of DC's products and services. UCLA-DC will extend their current work on scientific data practices and data curation requirements in embedded sensor networks to the fields of astronomy and astrophysics. We will examine initially the Sloan Digital Sky Survey (SDSS), and its relation to two subsequent sky survey projects: the Large Synoptic Survey Telescope (LSST) and the Pan-STARRS project. Also examined are the Infrared Processing and Analysis Center (IPaC), the International Virtual Observatory Alliance (IVOA) and Space Telescope Science Institute Archives (STScI) for relevant issues in data practices and metadata standards for astronomical objects and digital archives. Core questions for the three astronomy projects (SDSS, PAN-STARRS, and LSST) are focused on their design, structure, usage, and data practices, in support of data curation in particular. UCLA-DC will examine their history of development, core practices of data management and curation, hurdles overcome and remaining, and lessons learned that are instructive for related projects within and outside astronomy. Investigators will determine which forms of data are used, which are selected for sharing and curation and which are discarded, how they are curated, and expected future uses of curated data. We will employ a range of qualitative methods to address these goals including analysis of documented history of projects, oral histories and interviews with scientists, developers, and data managers, ethnography including observations of sites, and social network analysis of the involvement of key participants over time.

SDP 10.3 System(s) Description and/or Experiments

One central activity of the project will be the building and analysis of a relational database of documentation on project sites. The database will track project history, funding, personnel, timelines, and relationships among projects. The database will also incorporate for analysis data generated by our other methods. The database is built on an SQL network platform with a user interface operating in FileMaker Server.

SDP 10.4 Accomplishments

Specific accomplishments during the reporting period:

Borgman launched new graduate course on "Data, Data Practices, and Data Curation."

- Several meetings with main project partners, including project PI, SDSS director, and project partners at CIRSS and NCAR.
- Development and initial deployment of a relational database and document management system for organizational, ethnographic, and social network analysis.
- Established a relationship with the Caltech library, which helps manage large astronomy/astrophysics data centers and digital libraries, and who will aid in our contacts at Caltech and JPL.
- Had a facilities tour and conducted several interviews with personnel and astronomers at Caltech's Infrared Processing and Analysis Center (IPAC), to develop comparative materials for digital data management.
- Meetings with Microsoft Research, including VP Anthony Hey, to explore partnership directions with UCLA-DC and the DC project.

SDP 10.5 Future Directions

Our full interview and site visit schedule will begin in Summer 2010. Interviews at LA-area astronomy/astrophysics institutions will continue. Concurrently, we will be developing and populating our database with information about SDSS and other project sites, and extending its integration with our interview and ethnographic data, and social network analysis capacities. We will be coordinating activities with CIRSS, NCAR, and Microsoft partners.

All empirical results, use cases, and accounts of scientists' data practices and conceptualizations of data will be provided to the Data Conservancy Data Practices and Data Concepts groups with regular working sessions for joint interpretation. The UCLA-DC team will also collaborate with DC researchers at CIRSS and NCAR to integrate related findings, sharing results and interpretations of findings through the project wiki, conference calls, and scheduled face-to-face meetings. The overall analysis will produce a taxonomy of data practices and data attributes for assessment of curation needs and to facilitate curation activities.