# The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation

**Sayeed Choudhury [(1)], Robert Hanisch [(2)]**

[(1)] ***Sheridan Libraries, The Johns Hopkins University***
*3400 N. Charles St., Baltimore, MD 21218 USA*
*EMail: sayeed@jhu.edu*

[(2)] ***Space Telescope Science Institute***
*3700 San Martin Drive, Baltimore, MD 21218 USA*
*EMail: hanisch@stsci.edu*

## ABSTRACT

The Data Conservancy (DC) is one of two awards through the US National Science Foundation's DataNet program. The goal of the DataNet program is to create "a set of exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning."

The DC embraces a shared vision: data curation is not an end, but rather a means to collect, organize, validate, and preserve data to address the grand research challenges that face society. *The overarching goal of The Data Conservancy is to support new forms of inquiry and learning to meet these challenges through the creation, implementation, and sustained management of an integrated and comprehensive data curation strategy.* DC will address this overarching goal with a comprehensive project comprising four interdependent threads: 1) infrastructure research and development, 2) computer science and information science research, 3) broader impacts, and 4) sustainability.

The DC is led by the Sheridan Libraries at Johns Hopkins University. Working with the Sloan Digital Sky Survey data and the US National Virtual Observatory, the Sheridan Libraries have developed an initial architectural design, data models and metadata profiles, and organizational models to support data curation. The DC will build upon these initial lessons learned from the partnership between the library and astronomy community and extend them into the life sciences, earth sciences, and social sciences. Use cases will provide the initial framework for technical requirements. A robust information science and computer science research agenda will highlight the scientific requirements and inform the development of a data framework for observations and a theoretical framework for data curation. These activities will guide the development of new curriculum at library and information science schools thereby building capacity for a new generation of data scientists.

One of the central tenets of DC's sustainability plan relates to the leadership role of the library. The Sheridan Libraries at Johns Hopkins University have established a leadership position in prototyping data curation systems and services, especially as they relate to astronomy. One of the key outcomes of DC will be a new model for libraries in the digital age. There are several fundamental implications for libraries in the realm of data curation as they relate to collections, services, and infrastructure. The North American Association of Research Libraries has already engaged the DC in its effort to consider these implications strategically as a means to transform the library's role and contributions toward building and sustaining data curation infrastructure.

Keywords: astronomy, life sciences, earth sciences, social sciences, libraries, curation

## THE DATANET PROGRAM

Beginning in 2006, the National Science Foundation (NSF) began planning for a focused program of data preservation and curation by co-sponsoring a workshop organized by the Association of Research Libraries entitled "To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering" [1]. The primary recommendations from this workshop were:

> *NSF should facilitate the establishment of a sustainable framework for the long-term stewardship if data. This framework should involve multiple stakeholders by:*
>
> - *Supporting the research and development required to understand, model, and prototype the technical and organizational capacities needed for data stewardship, including strategies for long-term sustainability, and at multiple scales;*
> - *Supporting training and educational programs to develop a new workforce in data science both within NSF and in cooperation with other agencies; and*
> - *Developing, supporting, and promoting educational efforts to effect change in the research enterprise regarding the importance of the stewardship of digital data produced by all scientific and engineering disciplines/domains.*

Several terms are particularly important in these findings. The first is the emphasis on "stewardship." That is, the approach to data curation and preservation (DCP) can be flexible and adaptable, taking into account the practices and culture within a scientific discipline, and there is no prescription for building large centralized data warehouses, utilizing cloud storage, or any other specific solution. The second is "sustainability." Solutions for DCP require both upfront and long-term investment, and any approach that does not deal with long-term operational expenses is not viable. The third is "multiple scales." DCP for today is a multi-scale problem, with some disciplines dealing in gigabytes and others in terabytes or petabytes. DCP for the future is multi-scale, as new experiments and simulations, bolstered by exponential growth in computation and data handling capabilities, lead to petabyte and exabyte-scale data management challenges. And fourth is "training and education." Not only do we need to develop a work force with skills in DCP, we need to enlighten the scientists who produce digital data of the value of preserving this information. A recent issue of *Nature* focused on this problem, noting that while some disciplines are taking DCP seriously, others are lagging seriously behind [2].

NSF's response to this report was to create a new program in the Office of Cyberinfrastructure entitled "Sustainable Digital Data Preservation and Access Network Partners," or "DataNet" for short [3]. NSF solicited proposals in two rounds, one that closed in March 2008 and another that closed in May 2009. The program envisions the creation of new (virtual) organizations:

> *The new types of organizations envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to:*
>
> - *Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;*
> - *Continuously anticipate and adapt to changes in technologies and in user needs and expectations;*
> - *Engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and*
> - *Serve as component elements of an interoperable data preservation and access network.*

NSF allocated $100M to the DataNet program, with each of 5 potential awards to be funded at a level of up to $20M over a five-year period. Choudhury is the PI for one of the two successful proposals from the first round, entitled *The Data Conservancy.* Work is just getting underway in the fall of 2009.

## THE DATA CONSERVANCY COLLABORATION

We understood that to succeed in meeting the goals of the original "Test of Time" report and fulfill the expectations of the DataNet program that we would need to assemble a team that was both interdisciplinary and multi-faceted. Furthermore, we wanted to define the DCP approach in terms of science use cases that would be developed directly within the research community. We will be paying particular attention to use cases that span traditional discipline boundaries in order to test just how far one can push data interoperability.

The collaboration draws upon discipline-based data management experience and expertise in astronomy (US National Virtual Observatory), geophysics, atmospheric sciences, and biological sciences. It also includes individuals with technical and programmatic experience from research libraries, library science programs, and information technology. And finally it includes people who have created and studied large, distributed collaborations, as we aware that working in such an environment has both benefits and challenges.

### Common Motivators

Some of the most importance scientific questions facing society today require integration of data from subdisciplines within a discipline, and/or from different disciplines altogether. Several science use cases serve as examples.

- Using the Astrophysics Data System bibliographic database, an astronomer notes that a new article has appeared in the Astrophysical Journal in which the discovery of the most distant cluster of galaxies is reported based on observations using the Hubble Space Telescope. The astronomer follows the link to an online journal article and finds an image of the cluster. The journal image is a JPEG file and lacks information about the location, scale, or true flux values in the original data. Eager to compare this cluster with others, she searches for the original data, using the approximate coordinates embedded in the name of the cluster with the Virtual Observatory's data discovery tools. The original calibrated observations are found in the online archives of the Hubble Space Telescope, but upon inspecting these files, it is clear that the published image differs from the originals because of a non-standard rotation and resampling to a different spatial scale. Also, an unusual artifact in the original data has been removed through processing, or the published image has been scaled in a manner that obscures this problem in the data. This makes it impossible to verify the numerical measurements from the article or to perform a systematic comparison with data from other observations and in other bandpasses.
- A team of seismologists, meteorologists, volcanologists, geographers and sociologists collaborate to prevent fatalities and property loss due to massively destructive fast-moving volcanic landslides. These result from the fact that volcanoes are inherently unstable structures that are susceptible to collapse or "flank failure." Mitigating the effect of these landslides involves a mixture of risk assessment, forecasting, and deployment of early warning systems. The cross-disciplinary team focuses on the relationship between intense rainfall, seismicity, land-use practices and landslides, both from a historical perspective and a contemporary one. This investigation requires searching for records of historical events, compiling field observations on ground characteristics where landslides have occurred, and interviewing people living in regions of recent events. Quantitative data from seismographic and meteorological networks are gathered and analyzed. Topographic and geologic maps are used to identify at-risk areas. From this information a risk assessment model is developed that can be used to identify populations around the world that are at risk. This model can then be the basis for an education and outreach program and an early warning system that uses precipitation data from the WMO, seismographic data the Federated Digital Seismographic Network, and observations from local populations to make landslide predictions.
- The UN has adopted the Global Emitted Carbon Tax, and, to meet its goals, Brazil must determine the net impact of its land use policies. Examining the net effects of deforestation and associated increases in nutrient runoff, staff at the Instituto Oceanografico da Universidade de Sao Paulo access NOAA ESRL Carbon Cycle Greenhouse Gases for high sensitivity data on the shifting atmospheric gas levels, and the World Oceanographic Data Centers for changing levels of nitrogen, phosphorus, carbon and iron by remote underwater sampling stations in their exclusive economic zone. Real-

time GOES east/west composite satellite imagery is monitored for phytoplankton blooms, while the navy collects ocean samples for taxonomic identification, population densities, genetic makeup via molecular barcodes and chemical analysis of coccoliths and diatoms. The dominant taxa are plugged into The Data Conservancy, providing access to data on the historical deposition of phytoplankton from the Ocean Drilling Program cores at the National Geophysical Data Center in Boulder. They modify the use of these data in the US JGOFS synthesis model in collaboration with geochemists at the Oak Ridge National Laboratories model archive. Land-use models and the loss of carbon sequestration due to deforestation are coupled with prediction of the ocean carbonate sink in their report on net carbon dioxide balance to the UN's Global Change Amelioration Workforce. The high-level analyses and models are deposited in The Data Conservancy, tied directly to the underlying data to support future global and longitudinal analyses.

The first case highlights a simple but widespread problem in scientific research: the data upon which conclusions are based are often not available to others for inspection or re-analysis. The other two cases have a broader reach, calling for both data accessibility and data integration from disparate sources.

## Common Denominators

A system that enables data discovery, access, and meaningful re-use across disciplinary boundaries requires a common language for describing data. The Data Conservancy will develop a cross-disciplinary data framework for observations that will include core attributes such as identity, date/time stamp, and location and accommodate discipline specific attributes. Finding the commonality in diverse scientific data sets, yet exposing necessary details to expert users, is a fundamental research challenge of this project. The data model will be graph-based with embedded semantic (ontology-derived) attributes. It will allow the aggregation of information units from distributed sources. The components of observations will be definable statically or dynamically via predicates such as a SQL or SPARQL query on a database or triple-store. This common data model will provide the template for common representation of cross-disciplinary scientific information, either prospectively via authoring or data collection applications, or retrospectively via data mining from heterogeneous data stores. It will also permit joins across observations from different disciplines and data stores, a key to our goal of facilitating interdisciplinary investigations.

## Warehouse vs. Clearinghouse

The Data Conservancy approach to DCP builds on the experiences of the virtual observatory development efforts in astronomy. The virtual observatory [4,5] is a federation of distributed data collections. The federation is created through a common interface layer for metadata management and aggregation (the "registry"), data access, database query, and distributed storage and job execution. Data storage itself is distributed among national observatories, data centers, and research groups. There is no central data warehouse at all, and thus the virtual observatory functions as a clearinghouse.

The Data Conservancy needs to provide a clearinghouse role, but also recognizes that many important data collections and individual data sets have no permanent home. Thus, another important aspect of this effort is to develop a "black box" storage environment that can be deployed easily within various organizations and environments. For example, a university research library might host a local repository for which it takes on curation and preservation responsibilities, but it would then "publish" its collections to the Data Conservancy through broadcast and replication of its collection metadata. Organizations might also agree to shadow each other's collections in order to assure availability. The Data Conservancy itself intends to host a repository that various science communities can utilize for upload and sharing. Part of the data ingest process, whether into an institutional repository or a common Data Conservancy repository, is provision of collection and object-level metadata that conforms with the common data model. A schematic view of this architecture is shown in Fig. 1.

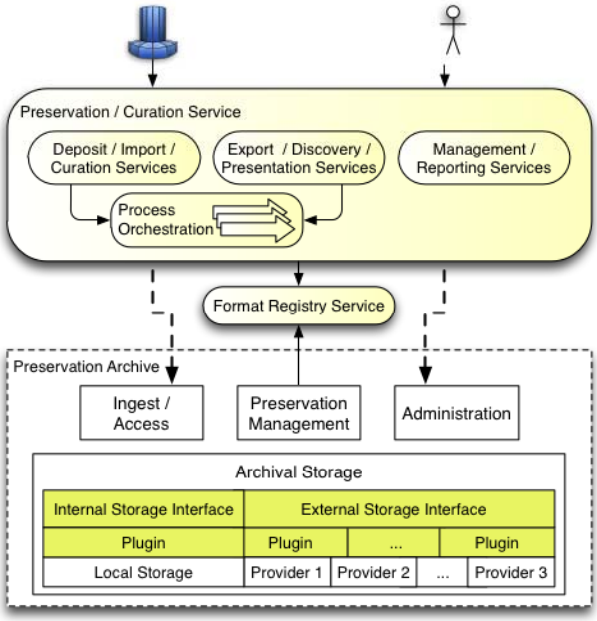


Figure 1: Data Conservancy architecture.

## PROJECT ORGANIZATION

We have organized the Data Conservancy project in a way that we believe addresses our key goals: being responsive to needs of the science community (both as data consumers and data providers), finding sound technical solutions to DPC that scale from local repositories to a national network, and understanding the total cost of ownership and models for sustainability (Fig. 2).

We have a Science Definition Group that engages various scientific domains to understand use cases. Working groups will be formed in different disciplines and will operate for a few years, to be replaced with other working groups with increasingly demanding usage scenarios.

Four technical teams focus on Sustainability, Broader Impacts, technology research (Information Science/Computer Science), and infrastructure development. Their work is overseen by a full time executive director.

Internal oversight is provided by the Partners' Council. The Partners' Council is composed of senior representatives of each member of the collaboration who represent each organization's interests and take responsibility for each organization's commitments to the project.

External oversight is provided by the Visiting Committee. The Visiting Committee is composed of external scientists and technical experts, including individuals with experience in managing large-scale distributed development projects. The Visiting Committee assesses progress against the project plan, advises on the effectiveness of project management, and makes suggestions for changes in strategy and priority.

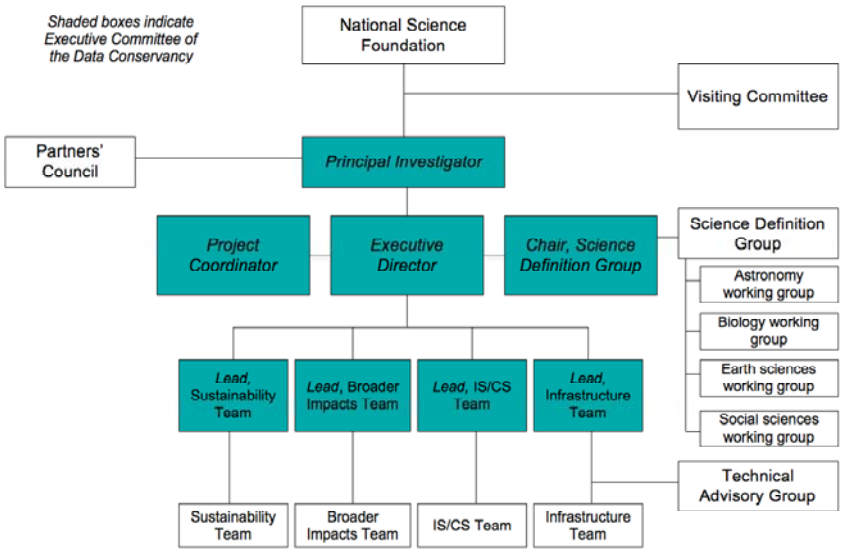


Figure 2: Data Conservancy organization.

## SUMMARY

The NSF DataNet program is providing support for 5 major development initiatives in data preservation and curation, of which the Data Conservancy project is one. The Data Conservancy team is interdisciplinary and multifaceted, and rooted in the university research library at The Johns Hopkins University. We believe that the library, with its tradition of information stewardship and experience in metadata curation, is the natural home for data preservation and curation efforts in the digital age. By basing the Data Conservancy program on genuine science use cases provided by active interdisciplinary researchers, we believe we can implement a truly useful system. By working within the domain of the university library, and in collaboration with scholarly publishers, we believe we can achieve a system that is sustainable and scalable.

## REFERENCES

[1] – A. Friedlander, P. Adler: "To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering." http://www.arl.org/bm~doc/digdatarpt.pdf

[2] – Nelson, B.: Nature 461, 160 (2009)

[3] – http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

[4] – US National Virtual Observatory (NVO). http://us-vo.org

[5] – International Virtual Observatory Alliance (IVOA). http://www.ivoa.net/

We would like to acknowledge major contributions to this paper and the Data Conservancy project from Tim DiLauro (Sheridan Libraries, The Johns Hopkins University), Mark Evans (Tessella Technology & Consulting, PLC), Ruth Duerr and Siri Jodi Khalsa (National Snow and Ice Data Center, University of Colorado-Boulder), David Patterson (Encyclopedia of Life, Marine Biological Laboratory), Patricia Romero-Lankao (Institute for the Study of Society and Environment, National Center for Atmospheric Research), and Mary Marlino (Library, National Center for Atmospheric Research).