# Proposal for the DISCOVER[1] Service Unit
# **Acquiring, Curating and Mining Scientific Datasets**

Version 3.0
April 15, 2008

## Executive Summary

Workers in most scientific disciplines are facing the challenges of managing the explosion in digital data. We propose a research service group that will facilitate data-driven science (DDS) at Cornell by developing archival and discovery tools across a variety of disciplines. The underpinning of discovery is an ensemble of data sets accessible for creatively motivated, cross-disciplinary analysis using an array of data mining and visualization tools. Most of the *types* of needed infrastructure exist but are too small in scale and inaccessible to the general research community. A system of hardware, software, middleware and human resources needs to be expanded and integrated to make DDS cost effective and accessible with low barriers of entry. Cornell's investments in the Cornell University Library, the Center for Advanced Computing, Cornell Information Technologies, the on-campus network backbone, and the National Lambda Rail along with support from domain science areas are essential parts of the system. Our goals are to develop a system for data curation that solves particular problems in specific research domains that also promotes cross-disciplinary studies. The system must scale from small to large data sets and use tools and protocols that are domain-independent where practical but allow domain-specific analyses without hindrance. Key to a usable system is the ability to browse metadata and to visualize selections from a multiplicity of data sets of high dimensionality.

The Discover Research Service Group (DRSG) will comprise research groups from several disciplines along with a small staff to provide the "glue" between research domains and the Center for Advanced Computing. It will survey current and projected needs of research groups and develop a roadmap for development of infrastructure and resources, particularly those that address common needs. Already, we have found a critical common need for data-management resources whose extent transcends that obtainable or developed by a single research group. The DRSG will conduct pilot projects in selected strategic areas that build upon current efforts in order to help develop a system of broad, general use. Activity will include development of data discovery portals using access-layer protocols now under development (e.g. Fedora Commons, Virtual Observatory). The DRSG will function, in broad terms, like the Computational Biology Service Unit (CBSU).

The DRSG will have a two year lifetime. In the first year, the DRSG will develop a white paper to assist and guide Cornell in planning data-driven cyberinfrastructure (CI), make recommendations for infrastructure, and commence pilot projects. The DRSG will facilitate collaborations on campus that respond to calls for proposals from the NSF and other funding sources. This work will continue in the second year. The project aims for a sustainable DRSG (or its follow-on) through a balance of grants for CI, cost recovery from domain science groups, and income-generating tools.

[1] DISCOVER = Data Intensive Science Organization for Virtual Exploration and Research

**1. Why a Discover Research Service Group?**

Empirical science is now almost invariably digital in nature. Sensors provide data at ever-increasing rates with the consequential explosion of data volumes. Assimilating data into knowledge is typically more challenging than acquiring the data. In addition, opportunities for discovery reside in the cumulative aspect of individual data sets and through the aggregation of data sets from different --- sometimes radically different --- sensors and research domains. This requires long-term curation, protocols for accessibility, tools for cross-disciplinary studies, and preservation, among other requirements.

Many research groups on campus find themselves in the position where their local resources are saturated with respect to handling the data explosion. There is also recognition that research groups face the same or similar problems in handling their data, so that a more collective approach will avoid repetition, promote the adoption of best practices, and be much more efficient and affordable. Of greatest import, however, are the synergistic opportunities that will emerge from use of protocols and portals that promote collaboration and discovery across disciplines.

We propose establishment of the DISCOVER Research Service Group (DRSG). It will identify and develop resources that promote data-driven discovery in Cornell's diverse research community. We will work with on-campus units that include the Cornell Center for Advanced Computing (CAC), Cornell Information Technologies (CIT), the Cornell University Library (CUL), and domain-science groups. The DRSG is the core of a broad collaboration that will bridge the disciplines of library and archival science, domain science, and computer and information science to address current and future challenges associated with curation and discovery. The DRSG will work closely with Cornell researchers to identify and implement common solutions to ensure maximal scientific and educational return, both today and in the future, of irreplaceable digital collections. This work will build on a solid foundation comprising collaborations between domain science groups and the CAC.

The overarching goal for the DRSG is to provide accessible paths for the curation (including preservation) and mining of scientific data. Data objects of all types --- text, images, time series, spectra aggregated over space and time, and others --- are now the content of digital libraries. The DRSG is less concerned with the acquisition of data, though it must consider how acquired data from specific domains can be accommodated.

Data curation --- the suite of activities required to manage digital research data over its entire life cycle --- is a dynamic and evolving field of research that major funding agencies such as the National Science Foundation acknowledge to be a central component of cyberinfrastructure (CI). Data are the material on which CI applications operate, and as such, must be properly formatted, well-documented, accessible and interoperable, and preserved over time. Cornell needs a system that will make data sets accessible physically in both space (over a wide network) and time (for the indefinite future) and also transparently, using modern web-based tools that are expected to evolve.

**2. Who Comprises the DRSG?**

The DRSG will consist initially of faculty and staff who have been involved actively in DDS and have expressed interest in widening the capabilities of on-campus infrastructure. The DRSG will of course be open to new members over its lifetime. In addition, the DRSG will include two new FTEs whose support we request from the Vice Provost for Research and whose job descriptions are given below. These FTEs will play key roles in accomplishing DRSG goals and serve research groups on campus. In addition, some of us are participating in a Cornell component of an NSF/Datanet proposal led by Johns Hopkins University, whose goals parallel ours but involve national and international collaboration. The DRSG plan has been communicated to the Datanet proposal team and it is their view that Cornell, in part via the DRSG, will be a testbed for data conservancy and it will also be a key site for maintaining specific data sets. If the proposal is successful, the Datanet project will provide a liaison person at Cornell who will bridge the DRSG and Datanet projects.

The FTEs associated with the DRSG (which we refer to as the "DRSG staff") will reside at and work with the CAC staff; they will interact with key players in the DRSG. In broad terms the DRSG staff will have expertise in several related scientific disciplines, library science and technical aspects of computers, networks, storage, databases, web-based tools and other software. Because the goal of the DRSG is to cross boundaries, the DRSG staff will necessarily have multi-disciplinary backgrounds, as implied by their job descriptions. Broadly, the DRSG staff will serve two primary functions. First, they will bridge the islands of expertise that currently exist on campus in research domains, in the Library, the CAC and elsewhere. Second, they will help develop explicit solutions for data curation and mining, both at the individual domain level and for the DDS system that we envision for Cornell.

We expect the stakeholders in the DRSG to comprise a large number of faculty and research groups across Cornell. To enable decision making and accomplishment of DRSG goals, we will establish a DRSG executive group, with initial membership from those listed on this proposal.

What can the DRSG do for Cornell stakeholders and how can Cornell research groups help the DRSG? The DRSG's broad goals are to identify a plan for data driven science and to begin the implementation of the plan by building outwards from current work in specific domains. By necessity, a critical mass of research groups needs to belong to or otherwise collaborate with the DRSG. Research groups will contribute their expertise and communicate requirements for their particular domains that will define the boundaries of a general use DDS system. The DRSG will identify the specifications and business model for curation and mining system(s) that are appropriate to generalization and it will conduct *bona fide* research activities needed for the development of the system(s).

## 3. Elements of a System that Promotes Discovery and Decision Making

To be of real use to domain researchers, solutions developed by the DRSG will require a balance between top-down organization and bottom-up activities. Overly top-down solutions cannot be forced on individual research groups nor are they optimal or even suitable in most cases. Invariably they will not be accepted. Mere aggregation of bottom-up solutions, however, would ignore the potential economies of scale and synergies that may be enabled by finding common solutions and by promoting communication between diverse groups. The DRSG will first survey domain research groups for common needs and potential solutions and then plan specific actions that will lead to a system of use to groups in multiple research domains.

Some general principles under which the DRSG will operate include:

- System design(s) need to grow out of science case studies and pilot projects.
- Open-source approaches are preferred to proprietary systems; however, they should be neutral with respect to operating systems and allow diverse points of entry.
- The overall approach should be light weight, agile, scalable and modular; it should allow diverse access methods to data, be they high-level data products or lower-level "raw" data.
- Algorithms for analysis and tools for visualization should emerge from research domains and collaborations between research domains enabled by the DRSG.

Discovery under our plan requires many components to enable curation and mining of data while providing long-term preservation as appropriate for the various discliplines. Some of these components already exist at some level but need to be integrated into a working system as well as enlarged in capacity and capabilities.

**Personnel to plan and develop the system and aid in its usage:**

Core staff is needed to provide expertise and to bridge research domains and groups.

**Storage including Disaster Mitigation:**

Short-term storage is essential for high-performance computing on large data sets; systems developed for simulations typically have too little storage.

Long-term curation of diverse data sets --- small to large, few to many in number --- requires a long term plan that responds to technology developments (tape, spinning disk, non-spinning disk, solid-state memory). Access tools, both within Cornell and to the outside world, are critical to our effort.

Many data sets are exceedingly expensive to reproduce if not outright irreplaceable. A storage system must mitigate the risk of loss, due to system failure or infrastructure damage (fire, flood, sabotage, etc.), implying the need for mirroring at a second site.

**High-speed networking:**

Bandwidth needs to grow both locally (on campus) and to state, national and international sites, enabling:

a. Flexibility in using computational resources on and off campus (moving data to the resources, etc.)
b. Scientific collaborations that require exchange of data and data products
c. Education (teaching and class-related research materials)
d. Outreach (Science@home, citizen science, etc)

**High performance processing:**

Computing resources reside in departments and labs, the CAC, at collaborating institutions and at Teragrid sites. The balance of usage of resources at these sites will undoubtedly evolve; the optimal configuration undoubtedly is domain and application specific and therefore needs to be flexible.

**Workflow:**

DDS involves intricate pipelines that need to handle raw sensor data and other kinds of data all the way to curated archives. These require a high degree of automation in order to handle large data sets and the multiplicity of data sets even if they are small. Expertise is needed for development or acquisition of workflow software that is most appropriate for a given project. A knowledge base of best practices is needed.

**Visualization:**

Discovery requires visualization of data at one or more levels in the continuum from raw data to data products and metadata. A DRSG system needs to accommodate the need for moving data to the infrastructure that provides the human interface. Our present view is that visualization is sufficiently domain specific that it is not the purview of the DRSG to provide specific tools but rather to provide the data hooks for enabling their development and usage. Moving data to the infrastructure has implications for network bandwidth. Where appropriate, any commonality of visualization tools should be identified and their availability accommodated.

Animations are a growing tool for communication of results both to scientific peers and for education and outreach. The DRSG should promote aggregation and dissemination of animations for these purposes because they are appropriate library objects.

**Databases:**

While database software is commonly available, expertise in database development that uses best practices for scalable systems is not. The DRSG should promote a diverse range of approaches to database development while also providing platforms for their storage and accessibility via the web. Elements of the system should recognize the common use of both proprietary software (MS SQL Server) and open-source software (MySQL). Databases are another example of the need for both centralized and distributed components. Any system needs to allow ready transfer between the two.

**Metadata management:**

Data characterization and catalogs are a necessary ingredient of cross-disciplinary studies. New approaches based on protocols and software developed under Fedora Commons, the National Virtual Observatory and the NSDL will be pursued by the DRSG in pilot projects.

**Scalability:**

The DRSG will consider approaches that are neutral with respect to data set or database size. Boundaries between small, medium and large datasets are artificial and time dependent and it would only be a hindrance to DDS if we were to focus on one or more extremes. However, practical approaches need to respond to what is affordable at any given time.

**Accessibility:**

Research domains have their own sociology and philosophy about data that are embedded in the prior investment in each community. Infrastructure for DDS must be based on the view that solutions will be accepted only if they are connectable to the protocols of a given community with low barriers of entry. They must be affordable.

**Interoperability:**

An agile DRSG system must be agnostic about computer OSs (Linux, OS X, Windows). Data formats cannot be forced; it must be recognized that data formats are often domain specific and have longevity. Best practice approaches may alter future usage in particular domains. To enable cross-disciplinary studies of multi-domain data, tools should be developed to provide translation from domain-specific formats to the common format needed for the studies. Universal formats, where appropriate, should be identified and promoted.

**Sustainability:**

The DRSG team understands that the initial subsidy provided by the Provost's office is for start-up funding only. Long term preservation of data obviously requires a sustained system of infrastructure and human resources. The pilot projects will provide insights into the infrastructure, expertise, functionality and costs associated with data-driven science. The projects will also help us outline a business model that balances the responsibilities of universities, individual research groups and departments on campus, and partner institutions that use our resources (as well as provide resources). The team expects that to be sustainable, support will come from a variety of sources, including cost-recovery from domain science groups, grants for cyberinfrastructure, and income-generating tools.

## 4. Activities and Needs in Domain Science Areas

**Astronomy: Fundamental Physics, Cosmology and Education/Public Outreach:**
(Brazier, Brown, Cordes, Giovanelli and Haynes)

Cornell faculty lead major survey projects using the Arecibo Observatory. These demand significant data management resources and also serve as prototypes for future surveys using other telescopes now under development with Cornell faculty leadership. NAIC staff play key roles in the data flow from Arecibo to the CAC and their activity is an important contribution to the DRSG activity we propose here.

**Science Goals:** Arecibo surveys aim for a comprehensive census of galaxies in the local universe to understand galaxy evolution and to use the spatial distribution of galaxies to probe cosmology. A deep survey for pulsars at Arecibo aims to find pulsars that are especially good for probing the physics of matter in extreme conditions, for testing theories of gravity and for use in the detection of gravitational waves from the early universe. The pulsar survey also allows detection of transient radio sources whose sources are not well understood. Another survey aims to probe the structure of the Milky Way galaxy, including turbulence in the interstellar gas. These surveys all use a multi-sensor receiver system, ALFA (the Arecibo L-band Feed Array) that produce large volumes of data. Their analysis leads to further, followup observations with Arecibo and other telescopes and to cross data-set analyses.

**Cyber-I Requirements:** Raw data set sizes are medium to large (hundreds of terabytes) and are to be archived indefinitely. Metadata and data products are put into databases accessible via the web for survey analyses and cross-analyses with astronomical data taken at other wavelengths across the electromagnetic spectrum. Raw data are of continued value because they can be mined using new algorithms as they develop, revealing new classes of astrophysical sources in some cases. Arecibo surveys involve national and international collaborations that require high-speed network access to the raw data and databases. These collaborations are formalized under several consortia.

The current system is a combination of Astronomy Department resources (workstations and modest storage) and a collaboration between Astronomy and the CAC under a CISE/RI grant that provides massive storage, a database server and a processing cluster. The collaboration involves database development, workflow of a processing pipeline and development of network access tools.

**Expected Growth of Astronomy Data Sets:** ALFA surveys will grow to at least 1 petabyte over the next five years as new instrumentation comes on line in 2008 and more telescope time is committed to the surveys. Follow-on Arecibo surveys will continue beyond the five-year period. Cornell faculty are involved in the development of new telescopes: the Cornell-Caltech Atacama Telescope and the Square Kilometer Array (SKA), with time frames of 5 to 15 years. Pathfinders for the SKA by international partners have time frames of 5 years and will provide data of interest to Cornell faculty for the same overall scientific goals as the ALFA surveys. Phase I of the SKA will be built in the next decade. We expect the overall growth of data to be in tens of petabytes over the next five to ten years and hundreds of petabytes afterward.

**Desired DISCOVER model:** Astronomical data curation and mining requires a system with all of the components described in section 4. Massive storage for long-term preservation as well as high-performance processing must be accompanied by large network bandwidth (10 Gb) on-and-off campus and easy input of data to the repository. Data mining requires access to raw data, data products and metadata via curated databases that themselves are development projects. The CAC is appropriate for algorithm and database development, immediate processing and database servers. The Library and CIT are appropriate for long-term preservation of data. Major development of data access portals is needed that uses new software infrastructure, including Fedora Commons and Virtual Observatory protocols. All systems need agility to promote innovation and scalability to meet the demands of future surveys. Cornell can play a major role in new telescope facilities as a data-conservancy center.

**Contributions to DISCOVER:** Astronomy faculty are already active in DRSG activities that provide the seeds for a more systematic approach and will contribute their time to the DRSG (Cordes and Haynes). NAIC management is responsible for long-term curation of ALFA survey results and supports Adam Brazier to develop databases and network access tools for data products. We will collaborate with Laboratory of Ornithology researchers on signal detection tool development. Astronomy faculty will also participate in writing of proposals to the NSF and other agencies, where appropriate, for cyber-infrastructure related to DRSG goals.

**Pilot Project Possibilities:** Growing the current archival, database and networking systems is a primary possibility. Processing of Arecibo pulsar data is an excellent test case for moving data on the NLR to Teragrid resources under the paradigm of "moving data to the computers" and also for multi-institutional collaborations on the data mining. Arecibo data also provide the basis for a joint astronomy-ornithology algorithm development project for detecting signals in the frequency-time plane. Despite the differences in sensors and motivation for data acquisition, statistical detection and characterization requirements are very similar, suggesting the potential for synergistic algorithm development. The ability to access both astronomy and ornithology data for performance tests is clearly a requirement for the pilot project.

**Crop, Soil and Atmospheric Sciences: The Cornell Computational Agriculture Initiative**
(van Es)

The Cornell Computational Agriculture Initiative is a collaborative effort between the College of Agriculture and Life Sciences and the Center for Advanced Computing (H. van Es, PI; Crop and Soil Sci.) and focuses on the application of high-performance computing to agricultural problems. A major effort relates to the development and application of high-resolution climate data, which has complex cyberinfrastructure requirements: Data streams from the National Weather Service are processed with models developed at the Northeast Region Climate Center at Cornell (A. DeGaetano, Earth Atm. Sci)), resulting in high-resolution climate data. These are warehoused at the Center for Advanced Computing and made available for data mining (P. Sullivan, Nat. Resources) to improve understanding of space-time patterns and error checking. The climate data are accessed for various stakeholder-focused applications through web-based services, including nitrogen management for maize (J. Melkonian; H. van Es, Crop and Soil Sci.). This involves the real-time application of a soil-crop dynamic simulation

model by farmers and consultants through a web interface, with high-resolution climate data input, to obtain nitrogen management recommendations. New applications will be developed for watershed N management (R. Howarth; Ecol.& Evol. Biol.), wine grape management (A. Lakso; Hort. Sciences-Geneva), use of geographic visualization methods, and use of spectroscopy data.

**Ornithology: Avian Knowledge Network, Nocturnal Flight Survey and Citizen Science:** (Kelling)

The Laboratory of Ornithology is the nexus for avian data from many sources. It is the primary site for the Avian Knowledge Network (AKN) that aggregates data on bird populations. It also promotes citizen science activities that provide data over a wide geographical area. Digital data are now produced by distributed streaming acoustic sensors at rates comparable to astronomy data. These provide input to understanding the effects of human activity on bird populations and input for decisions that may mitigate some of those effects.

**Science goals:** The Avian Knowledge Network characterizes bird populations spatially and temporally through aggregation of data into a large database with subsequent population analyses. The nocturnal flight call network of sensors aims to quantify bird migrations through blind species detection in digital frequency-time acoustical data, crucial input to the development of decision support tools for land managers.

**Cyber-infrastructure requirements:** Lab of O. is the primary center for the AKN with distributed nodes at partner institutions. Current data sets require about 10 TB and grow 30% per year. Storage at the Lab of O will be outgrown soon. The wide variety of data for the AKN, citizen science, and acoustical data is an important resource indefinitely and therefore requires long-term preservation as well as growth to accommodate new ventures.

**Expected growth of data sets and required resources:** The most demanding data source is from the streaming acoustic sensors that will produce tens of terabytes per year initially and that will increase as more sensors are put on the ground. The storage and accessibility requirements are very similar to the astronomical surveys, so the same issues of scalability and preservation apply.

**Desired DISCOVER model:** Essentially the same as the astronomy model.

**Contributions to the DRSG:** Lab of O participants will help define data curation requirements and provide lessons learned from their data management efforts. Algorithms for signal detection in acoustical data may have relevance to detection of astronomical transients signals and vice versa. Ornithology data are of wide interest to the public so success in management of relevant data will give visibility to Cornell's CI efforts.

**Pilot project possibilities:** Species occurrence data across the U.S. lend themselves to development of visualization techniques that use a CI system of data storage, databases and networking. Detection algorithms for bird calls in acoustical data can be co-developed with astronomers working on detection of transient signals. In both cases,

detection algorithms operate on the frequency-time plane after filtering in the spatial domain.

**Ornithology: the Loon Population Database**

(Walcott)

Charles Walcott and his collaborators, Walter Piper (Chapman University) and John Mager (Ohio Northern) study how loons establish and maintain their territories. The Common Loon, Gavia immer, breeds on freshwater lakes and is highly territorial. In a small lake one pair of loons will nest and, if all goes well, produce two chicks. By banding many hundreds of loons in a series of about 150 lakes surrounding Rhinelander, WI, the researchers are able to identify individuals and study the dynamics of their populations and the territorial interactions. For example, a young loon hatched in the study area returns after 3-4 years on salt water. There are three ways that it can establish a territory: It can colonize a vacant lake, replace an established breeder that failed to return from migration or it can displace a member of an established pair. Loons do all three of these things, but displacing an established breeder is the most common. If it is a female that intrudes, there is often a fight. The loser, whether resident or intruder, leaves and the winner mates with the territorial male. On the other hand, if it is a male that intrudes, the fights are much more severe; on about 30% of the occasions a male is killed. Interestingly, it is always the territorial male, never the intruding male.

This research involves noting which loons are seen on which lake, sampling their behavior for one hour and recording any sounds they make. This process has resulted in a large database; over 15 years of recordings, census data and behavioral observations. The researchers would like to make this database available to other investigators since it contains unique information on loon life history and behavior. The loon database is an example of a small-scale database that requires long-term curation.

**Physics & Theoretical & Applied Mechanics: the Cornell Insect Flight Database**

(Cohen, Wang)

Already in collaboration with the CAC, researchers in the Physics Department and in the Department of Theoretical & Applied Mechanics are creating a database which will store movies, data and theoretical analysis tools relating to the flight of insects. The project entails taking advantage of state-of-the-art experimental and numerical techniques currently being employed in the Cohen and Wang groups as well as the computational and data infrastructure tools available at the CAC to collect, archive and analyze insect flight data. In particular, Cohen has built a state-of-the-art visualization facility for recording untethered insect flight in 3D. The major advance of the facility is the full automation of the data collection which allows for a dramatic advance in the number of movies obtained for a given flight maneuver in a particular species of insect; data collected over a month's time will allow for the detailed analysis of 20 to 30 distinct maneuvers for a given species of insect. As experts in unsteady aerodynamics, Wang and her group are developing a suite of advanced computational tools for analyzing insect flight. The Insect Flight Database IFD will make available to the scientific community both the movies, the computational tools and their output for the purpose of studying insect flight.

**5. Activities in Library Science**

The Cornell University Library (CUL) will work with DRSG staff and domain scientists to assess the data needs, both met and unmet, for the above-mentioned projects and how the needs vary by discipline, domain, and type and size of data. In addition, the Library will solicit interest in participation from research projects in several disciplines at Cornell having data of significant size, time span, ongoing growth, and/or complexity requirements, and also seeking to identify improved means to disseminate and archive data. It will also need to consider image data (especially from high-throughput applications requiring automated access or processing), databases, and other data objects extending beyond single files or simple directory-based organization. Working with DRSG staff and domain scientists, the Library will discuss with each identified project their current data management methodologies and plans for publication/distribution, as well as any plans for submission to an existing long-term data repository.

Following this analysis, the Library will work with DRSG staff to assess existing CI data workflow and management tools, databases, and data storage solutions and their applicability to medium- to large-scale data in Cornell projects. It will be important to review current best practices for data management by long-term repositories (e.g., ICPSR, KNB) and current literature on grid-based approaches, science data-oriented web services, new approaches to database archiving, and techniques for interoperability and inter-process communication/workflow. The Library will gather information from Cornell participants to support the development of requirements for long-term digital preservation of data including policies in regard to storage architecture, quality control, scope and extent of preservation activities and priorities, preservation levels, and risk management. Fedora is the leading candidate repository platform for metadata, documents, and smaller datasets; the DRSG team will work with Fedora Commons researchers to evaluate and extend Fedora's support for managing medium to large-scale data resources

Working with staff at DRSG, the Library will also investigate leading open-source grid applications and tools (e.g., the Opal Toolkit, Kepler, and image processing/visualization tools for grid environments) and their potential for the specific data under consideration, as well as for more general applicability as components of a set of tools leveraging a scientific data repository (recognizing that this project will have only very limited resources for new code development). The focus will be on the potential for common approaches to data mining, extraction, and exchange that can leverage web services and standardized data formats, seeking to build specialization from a common base of standards only where necessary. It will be important to:

a) Assess standard costs and how they would change as startup effort is replicated with additional datasets

b) Explore feasibility of common approaches so that efforts can be prioritized in areas with potential scalability

c) Appraise the data archive as a trusted repository based on the National Archives and Records Administration’s guidelines, “Trustworthy Repositories Audit and Certification Criteria and Checklist” and the UK’s Digital Curation Centre’s toolkit, “Digital Repository Risk Assessment Audit Method.”

d) Investigate long-term financial and technical sustainability of systems and associated services in support of data-driven science at Cornell.

McCue et al. will consider the role of a "dCommons" for data: an institutional repository analogous to Cornell University Library's institutional repository — eCommons — but dedicated to distributing and preserving digital research data. Mann Library's current efforts are directed at small-scale data sets, and a dCommons could fill an important gap in infrastructure between small-science data sets and the very large data sets that are generally the basis for CI development. Such an infrastructure would be targeted at medium-sized data sets, or those data sets that are too large to provide access by treating them as a single (sometimes complex) digital object, but too small to justify dedicated and specialized platforms for distribution and archiving. Certainly some databases fall into this category, and research needs include the evaluation of emerging solutions for database preservation, as well as developing generalized solutions for ongoing access.

Collaborating with CIT & CAC, the Library will coordinate/facilitate the development of preservation requirements necessary to ensure the sustainability of data sets and associated applications and software tools. Digital preservation (archiving) involves a range of managed activities to support the long-term maintenance of bitstreams and continued accessibility of content. Bitstream preservation aims to keep the digital objects (data, metadata, identifier) intact and readable. It ensures bitstream integrity by monitoring for corruption to data fixity and authenticity; protecting digital content from undocumented alteration; securing the data from unauthorized use; and providing media stability. Preserving access is more involved and entails ensuring the usability of digital data retaining all quantities of authenticity, accuracy, and functionality necessary for its use.

## 6. Initial DRSG Model

Figure 1 shows a notional diagram as to how the DRSG will function. The Office of the Vice Provost for Research will sponsor the DRSG initially, which will be managed by a faculty core group from relevant departments and labs and will have staff comprising two FTEs funded by the OVPR (see below). The purpose of the DRSG is to determine and respond to the needs of research groups in particular science areas, as indicated. A small number of domain groups will be involved initially and we expect the number to grow. Implementation of DRSG recommendations initially will be at the CAC while we expect long-term storage and networking to be implemented through CIT. The Cornell University Library and Fedora Commons will play key roles in high-level cyber-infrastructure definition. All entities shown in the diagram will be involved with the development of partnerships both within Cornell and with other institutions.

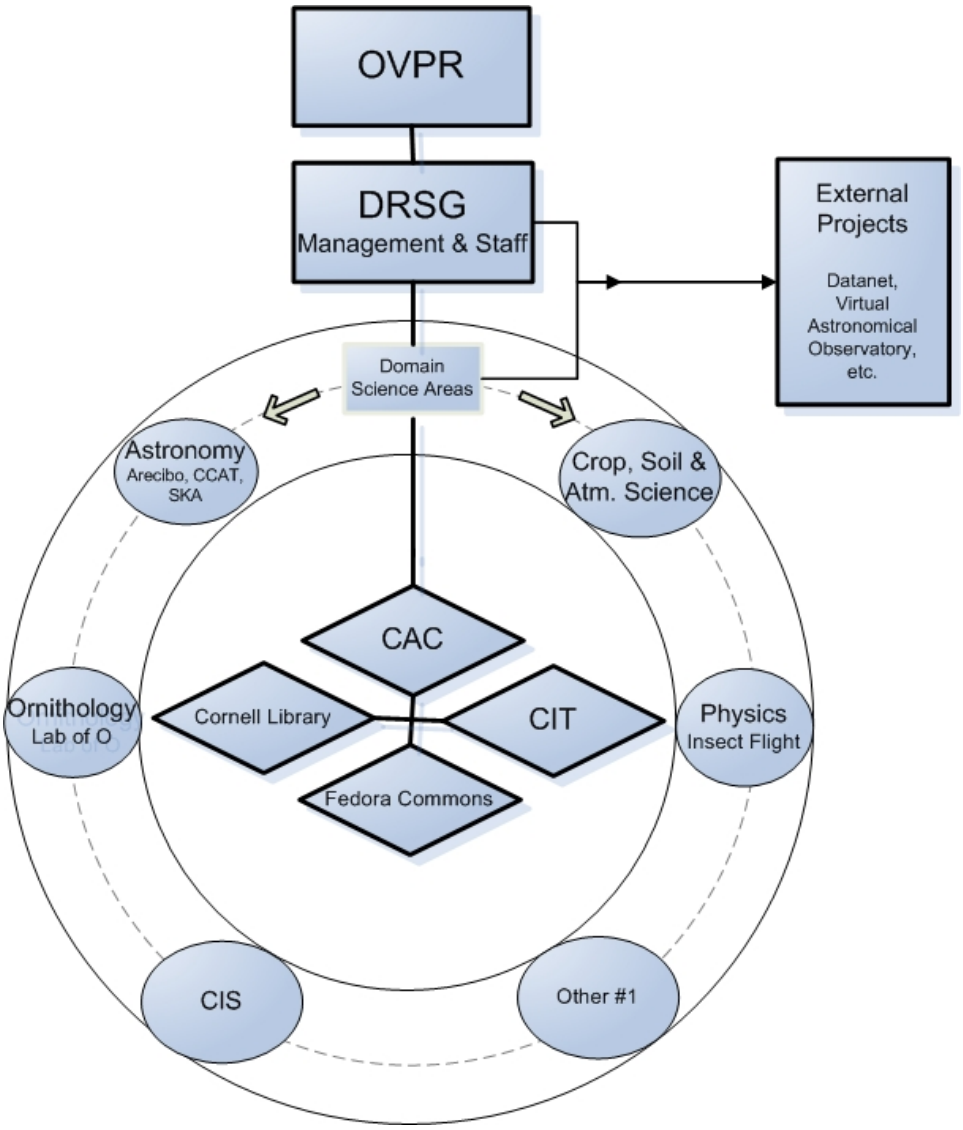


**Figure 1: Notional diagram for how the DRSG will work with domain science areas to implement data management systems working with the CAC, CIT, the Library and Fedora Commons. DRSG Management consists of a faculty core group. DRSG Staff comprises two FTEs who will reside at the CAC.**

Business model: We request startup support for the DRSG staff from the OVPR for a period of two years along with associated travel and miscellaneous expenses. A third person directly involved with the DRSG will be funded by Fedora Commons to work on one or more pilot projects that emerge from domain-science activities.

Additional funding and in-kind support will come from departments, labs, the Library, and domain science groups with resources that can be pooled and from grant proposals now being written or to be written over the next two years and beyond.

**Progress from Current Efforts:** Current effort on astronomical data sets has been done largely under an NSF CISE/RI grant to Computer Science (Demers, Gehrke) and Astronomy (Cordes). We are in the fourth year of a 5-year project that has acquired hardware for archival, processing and dissemination of data and data products from Arecibo. This has involved weekly meetings between Astronomy staff (Cordes, two research associates and a graduate student) and CAC personnel. Work has led to archiving protocols, monitoring of data integrity, development of databases and web-based tools for accessing data and data products, dealing with security issues associated with outside usage, and linkage to the Teragrid using Cornell's investment in the National Lambda Rail. We are in the process of getting a 10 Gb connection to the Space Sciences Building in order to facilitate workflow between Astronomy and the CAC. The current request will enable us to develop Virtual Observatory resources that will enable truly long-term usage of the data and data products. We also request resources so that tools developed under the CISE/RI grant with the CAC can be migrated to other astronomical surveys being done with Arecibo. This requires personnel that cannot be acquired under the CISE/RI grant.

The Cornell University Library has been engaged for a number of years in digital collection curation, long-term access, and preservation. The library has developed and provides ongoing support for digital collections, hosts the arXiv, and participates in multiple national and international digitization initiatives.

CUL has substantial expertise in creating and maintaining digital collections encompassing a wide range of formats, including still images, numeric data files, AV, GIS data files, and CAD files. Currently, the Library maintains and provides access to more than 40 open access repositories that are described in the Library's digital collections registry (http://rdc.library.cornell.edu). Through the development and maintenance of these scholarly information portals, the library has gained significant experience in digital curation, including negotiation with content owners, digital conversion, metadata creation, user requirements analysis, rights management, lifecycle management, and archiving. CUL has been at the forefront of research and training in digital preservation at an international level through offering a week-long workshop on digital preservation management and hosting an international conference. The Library's recent preservation agenda focuses on the development of an aDORe-based digital repository in addition to developing protocols for distributed preservation. On the storage architecture front, the Library has been collaborating with CIT in developing a storage strategy in support of this large-scale digitization initiative. This experience has greatly expanded the library's understanding of issues involved in large digital collections, including storage systems, media, and data management. The Library's digital preservation program acknowledges that preservation is not only a technical challenge but also has equally important organizational implications. The Library continues to explore organizational matters such as retention policies, business plans, IPR policies, and governance issues. The CUL Data Working

Group was established in early 2007 to review and discuss issues related to the curation of research data and has worked closely with the Office of the Vice Provost for Research on the Cornell University data retention policy, which is still in draft form.

Mann Library has also been awarded two NSF grants in this area including a Small Grant for Exploratory Research and a new project entitled, "Promoting the curation of research data through library-laboratory collaboration" (NSF award IIS-0712989, PI: Gail Steinhart, Co-PI: Janet McCue). In the latter project, Mann Library is developing services and infrastructure to support pre-publication sharing of data among selected collaborators, the creation of preliminary metadata early in the research process, and ultimately, the transfer of data and metadata to domain and/or institutional repositories, where data would be housed for the long term and made publicly accessible. This work includes development in two main areas:

- A metadata architecture that allows managers of data staging repositories to approach heterogeneous data and metadata in a more flexible way while still leveraging the significant investment that has already been made in discipline-specific metadata schemas.
- A set of services and procedures designed to facilitate the transmission of "publication-ready" data and metadata to domain and institutional repositories.

**Competitive Advantage for Funding – Specific Programs for Future Funding:** The establishment of the DRSG associated with the CAC is a significant institutional step that will serve as a cornerstone for future collaborations and extensions both within and outside Cornell, demonstrating institutional capabilities that will make Cornell a more effective competitor in multiple programs.

**DataNet** – The NSF Office of Cyberinfrastructure recently issued a call for pre-proposals: Sustainable Digital Data Preservation and Access Network Partners (DataNet). The goal of the program is to foster innovation in data curation as well as in organizational structures focused on this activity. Funding is up to $20 million per award (over five years), for a total of five awards to be made over a two year period. NSF program officers have repeatedly stressed the desirability of partnerships that integrate the capabilities of cyberinfrastructure with domain science, computer and information science, and library and archival sciences. It would enhance Cornell's competitiveness to have a track record of cooperation across these sectors, in addition to the activities already underway within each.

**IGERT –** the DRSG will provide opportunities for graduate students to adopt CI applications and data sharing. This should enhance future IGERT proposals coming out of the domains, as well as a DataNet proposal, because NSF recognizes that changing work practice/culture to promote data sharing and the use of CI is problematic in many disciplines. This could take the form of workshops or short courses for graduate students, individualized consulting, and computing accounts at CAC.

**National Virtual Observatory –** the NVO is being developed under an NSF/NASA collaboration. Both M. Haynes and C. Lagoze (CIS) served on the Advisory Committee for the NVO. From the Astronomy side, the NVO and our Arecibo data sets form the foundation under which Cornell can participate in astronomically-related cyberinfrastructure. We are in a position to provide unique, high-volume data sets to the worldwide community and consider the NVO initiative to be a likely source of follow-on funding for the DISCOVER RSG.

**7. DRSG Activities**

The DRSG will

1. Identify the infrastructure, expertise and functionality needed for research groups doing data-driven science (DDS) at Cornell. Characterize the issues related to curating and mining of data sets with a wide range of scale (small, medium and large).
2. Conduct pilot projects that build upon existing efforts by research groups on campus to develop core resources of general use for DDS. These will provide proofs of concept and a foundation for new strategic directions.
3. Develop a cyberinfrastructure white paper to assist and guide Cornell in planning for handling large research datasets in the short, mid, and long terms with respect to:
    a. overall requirements of Cornell research groups, including those that are domain specific and those that can be generalized:
        i. Astronomy
        ii. Atmospheric, Crop and Soil Science
        iii. Medical imaging
        iv. Ornithology
        v. Physics
        vi. Sociology (e.g. Weblab)
        vii. Other?
    b. networking requirements on-and-off campus (Teragrid and international)
    c. space requirements for computation and storage facilities
    d. disaster recovery
    e. curation and data storage, including migration to new media as they develop
    f. cyberinfrastructure reliability related to data access and processing for real-time applications
    g. use of cyberinfrastructure technology by off-campus stakeholders
    h. domain-specific requirements
    i. cross-domain collaborations, innovations and economies
    j. data standards and formats, metadata and translation between formats
    k. leveraging of Library resources and methodologies
    l. strategic collaborations with other institutions
    m. education, public outreach and citizen science
    n. best practices
4. Identify and facilitate responses to funding opportunities and collaborations with particular emphasis placed on Cornell's strengths.
5. Recommend Cornell investments that are essential for Cornell's leadership in DDS and competitiveness in national competitions.
6. Provide guidance about usage and cost-recovery of on-campus resources (e.g. CAC, CIT, Library) that will help Cornell researchers in the near term and help launch the plan in the white paper.

## 8. Personnel, Roles and Job Descriptions

Researchers who have participated in the formulation of the DRSG include:

| | |
|---|---|
| Jim Cordes | Astronomy (Co-PI) |
| Janet McCue | Mann Library (Co-PI) |
| Adam Brazier | NAIC |
| Bob Brown | NAIC |
| Itai Cohen | Physics |
| Jonathan Corson-Rikert | Mann Library |
| Art DeGaetano | Earth and Atmospheric Sciences |
| Riccardo Giovanelli | Astronomy |
| Martha Haynes | Astronomy |
| Steve Kelling | Lab of Ornithology |
| Carl Lagoze | CIS |
| David Lifka | CAC |
| Sandy Payette | CIS |
| Mirek Riedewald | CIS |
| Oya Rieger | Cornell University Library |
| John Saylor | Engineering Library |
| Gail Steinhart | Mann Library |
| Harold Van Es | Crop and Soil Science |
| Jane Wang | Theoretical and Applied Mechanics |

**Co-PIs:** Cordes will lead the large-data-set side of the DRSG, working with other domain groups to define the structure of a system that includes the elements described above. He will lead activity that involves astronomy data, working in consultation with his Cornell astronomy colleagues working on Arecibo surveys and on Virtual Astronomy Observatory utilities.

McCue will lead the library science aspect of tool development and the curation of data sets. She will be the primary Cornell University Library liaison to CAC and Fedora Commons, work with CAC staff to specify requirements for a dCommons data repository system, identify and refer researchers with candidate data sets for inclusion in the repository to CAC, and identify possible ongoing roles for library staff in supporting the system.

**DRSG Executive Group:** This group will initially consist of one representative from each domain research group along with the CAC Director and Library Director. The initial group will decide on a fair plan for representation that allows for growth as the DRSG evolves. The Executive Group will work with the CAC Faculty Oversight Committee and the OVPR to devise appropriate metrics to evaluate the progress of the DRSG by the end of year one and at the conclusion of the grant. These metrics may include, but are not limited to, submissions of proposals to outside agencies for funding; billable consulting hours, etc.

**DRSG Staff Job Descriptions**

Two full-time positions are required for the DRSG; one an Observational Sciences Data Manager and the other a Research Data Curator Specialist. Successful candidates for these

positions will be selected by mutual agreement of the co-chairs of the DRSG and the CAC Director, with input from the DRSG Executive group. These individuals will be appointed as staff in the Cornell Center for Advanced Computing and work closely with domain scientists, digital library and data specialists, and colleagues in the CAC.

**1. Observational Sciences Data Manager**

A candidate for this position should have a degree in one of the physical sciences, preferably applied physics, astronomy, geophysics, physics or possibly chemistry, and several years of experience in managing data for a large project. Useful areas of experience include database development using MS SQL Server or MySQL, web services, XML, web application programming (e.g. ASP.NET, php, and HTML/CSS), National Virtual Observatory protocols, high speed networking, Teragrid and NLR, storage resources, network security, statistical inference, workflow, image processing, data formatting and translation, visualization methods, C , C++, C#, perl and python programming.

The OS Data Manager will reside at the CAC but will work closely with Cornell research groups to proactively define data management needs, develop solutions involving CAC staff and resources, and identify new infrastructure needed for data management. S/he will identify areas of commonality between different scientific domains --- as well as differences--- in the early phases of the DRSG as part of a roadmap process for developing data-oriented CI.

**2. Research Data Curator Specialist**

A candidate for this position will work with researchers across a variety of disciplines in the life and earth sciences to improve access to scientific information. The Research Data Curator Specialist will survey and investigate new methods and tools to facilitate data-driven science in support of domain science; develop user-friendly interfaces for data discovery, acquisition, analysis, application and visualization; and coordinate the needs assessment, development, evaluation, and operation of a data repository to ensure the usability and archiving of data. The Specialist will reside at the CAC and will work closely with library and CAC staff in support of user-centered cyberinfrastructure (CI) development as well as foster partnerships with stakeholders to advance efforts to obtain funding for CI development.

Qualifications: MS in computer or information science. A working knowledge of high-performance computing workflows, visualization, and data mining tools is required. Strong communication and collaboration skills as well as strong analytical and project management competencies are required as well as one year experience in data management, database design, and web interface development. Preferred qualifications include a background in the life or earth sciences, knowledge of existing data and metadata standards in at least one scientific discipline, and experience in a research library. An understanding of digital preservation best practices and tools is also helpful.

**3. Fedora-Commons Liaison Specialist**

The DRSG will use Fedora Commons for one or more of the pilot projects to develop archives and accessibility tools that will be sustainable and scalable In addition to DRSG personnel who have expertise in some of the domain science areas and in hardware

technology, we need an FTE that is directly connected to the FC project. This FTE will be funded by the FC project itself and will spend time on FC aspects of the Datanet project now being proposed by a consortium led by Johns Hopkins, with Cornell and FC participation. The job description and work plan is now being discussed among FC, DRSG and Datanet participants.

In addition to these, we request 50% support for a high-level administrative aide, who can support report and proposal preparation.