The Data Observation Network for Earth (DataONE) is a virtual organization dedicated to providing open, persistent, robust, and secure access to biodiversity and environmental data, supported by the U.S. National Science Foundation. DataONE is pleased to announce the availability of summer research internships for undergraduates, graduate students and recent postgraduates.

**Program Structure**

Up to eight interns will be accepted in 2011, each paired with one primary mentor and, in some cases, secondary mentors. Interns need not necessarily be at the same location or institution as their mentor(s). Interns and mentors are expected to have a face-to-face meeting at the beginning of the summer, and interns are encouraged to attend the DataONE All-Hands Meeting in the fall to present the results of their work. DataONE will pay all necessary travel expenses.

**Schedule**

- **March 15** - Application period opens
- **April 8** - Deadline for receipt of applications at midnight Pacific time
- **April 15** - Notification of acceptance. Scheduling of face-to-face kickoff meetings based on availability of interns and mentors
- **May 23** - Program begins*
- **June 27** - Midterm evaluations
- **July 29** - Program concludes
- **October 18-20** - DataONE All-Hands-Meeting, New Mexico (attendance encouraged)

* Allowance will be made for students who are unavailable during these date due to their school calendar.

**Eligibility**

The program is open to all undergraduate students, graduate students, and postgraduates who have received their masters or doctorate within the past five years. Given the broad range of projects, there are no restrictions on academic backgrounds or field of study. Interns must be at least 18 years of age by the program start date, must be currently enrolled or employed at a university or other research institution and must currently reside in, and be eligible to work in, the United States. Interns are expected to be available approximately 40 hours/week during the internship period (noted below) with significant availability during the normal business hours. Interns from previous years are eligible to participate.

**Financial Support**

Interns will receive a stipend of $4,500 for participation, paid in two installments (one at the midterm and one at the conclusion of the program). In addition, required travel expenses will be borne by DataONE. Participation in the program after the mid-term is contingent on satisfactory
performance. The University of New Mexico will administer funds. Interns will need to supply their own computing equipment and Internet connection. For students who are not US citizens or permanent residents, complete visa information will be required, and it may be necessary for the funds to be paid through the student’s university or research institution. In such cases, the student will need to provide the necessary contact information for their organization.

**Project Ideas**

Projects cover a range of topic areas and vary in the extent and type of prior background required of the intern. The interests and expertise of the applicants will, in part, determine which projects will be selected for the program. Off-list projects are also eligible, in which case potential applicants are strongly encouraged to contact the organizers and/or potential mentors with their ideas prior to applying. The titles of this year’s projects (see below for more detailed descriptions) are:

1. DATA MANAGEMENT: Best practices of data management for public participation in science and research
2. DATA MANAGEMENT: Online learning modules related to best practices throughout the data lifecycle
3. EDUCATION: Accessing and analyzing environmental data in the classroom
4. SOCIOLOGY OF SCIENCE: Understanding how scientists analyze data
5. DATA SCIENCE: How much ecological data is out there?
6. DATA SCIENCE: Tracking the reuse of 1000 datasets
7. PROGRAMMING: Subsetting and publishing “dynamic” scientific datasets
8. PROGRAMMING: Scientific workflow provenance repository and publishing toolkit
9. PROGRAMMING: Integrating loosely structured data into the Linked Open Data cloud
10. SCIENCE COMMUNICATION: Developing video animations for DataONE community engagement

**To Apply**

Application materials should be sent to internship@dataone.org by 11:59 PM (Pacific time) on April 8th, and should include a cover letter, resume and letter of reference all in PDF format. The applicant should send the cover letter and resume, while the letter of reference should be sent directly by its author.

1. The cover letter should address the following questions:
   - What DataONE Summer Internship projects are you most interested in and why?
   - What contributions do you expect to be able to make to the project(s)?
   - What background do you have which is relevant to the project(s)?
   - What do you expect to learn and/or achieve by participating?
   - What are your thoughts and ideas about the project, including particular suggestions for ways of achieving the project objectives?
   - How will participation in this program help you achieve your educational and career objectives?
Are there any factors that would affect your ability to participate, including other summer employment, university schedules, and other commitments?

2. The resume should include the applicant’s educational history, current position, any publications or honors, and full contact information (including phone number, e-mail address, and mailing address).

3. The letter of reference should be sent directly to internship@dataone.org and should be from a professor, supervisor, or mentor.

**Evaluation of applications**

Applications will be judged by the following criteria:

- The academic and technical qualifications of the applicant.
- Evidence of strong written and oral communication skills.
- The extent to which the applicant can provide substantive contributions to one or more projects, including the applicant’s ideas for project implementation.
- The extent to which the internship would be of value to the career development of the applicant.
- The availability of the applicant during the period of the internship.

**Intellectual Property**

DataONE is predicated on openness and universal access. Software is developed under one of several open source licenses, and copyrightable content produced during the course of the project will made available under a Creative Commons (CC-BY 3.0) license. Where appropriate, projects may result in published articles and conference presentations, on which the intern is expected to make a substantive contribution, and receive credit for that contribution.

**Funding acknowledgement**

The Summer Internships are supported by The National Science Foundation: "INTEROP: Creation of an International Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences" (NSF Award 0753138) and "DataNet Full Proposal: DataNetONE (Observation Network for Earth)" (NSF Award 0830944).

**For more information**

If you have questions or problems about the application process or internship program in general, please send e-mail to internship@dataone.org.

**Project Ideas**

1. **Best practices of data management for public participation in science and research**
   **Description:** The DataONE Citizen Science Working Group (CSWG) is working to organize and develop best practices for management of data and information for the
increasing number of local, regional and national projects that focus on “Public Participation in Science and Research (PPSR),” also called Citizen Science projects. The 2011 CSWG intern will assist in the inventory and description of data practices for PPSR projects, based on the response from an earlier survey conducted as part of the CSWG. The goals of the intern project are to develop a metadata description for key aspects of the data held by each group, and make this information available back to the CSWG as a small database. The intern will then help identify and document best practices for data management by PPSR projects, assist in vetting the best practice documents across the PPSR community, and work with CSWG to make the best practices available via the DataONE website as well as other outlets. Products will include a suite of best practices for data management by PPSR projects; in addition, the intern will be encouraged to give a formal presentation at a scientific, data management or PPSR conference or meeting.

Local work preferred, at Tucson or Ithaca, though remote work would be possible for outstanding candidates (though one trip for an organization meeting would be required).

Qualifications needed:
- Undergraduate or graduate student or equivalent; simple database management (e.g., MS Access) skills preferred; public engagement; writing; organization; small project management
- Skills to be learned: Metadata management; best practices template; database management; communications and outreach; project management
- Primary mentor: Jake Weltzin (USA National Phenology Network)
- Secondary mentor: Rick Bonney (Cornell Laboratory of Ornithology)

2. Developing online learning modules related to the best practices throughout the data lifecycle

Description: DataONE is developing online learning modules designed to educate DataONE users in various aspects of the data lifecycle. This project involves: 1) researching and acquiring software that can produce high quality online learning; 2) developing online learning modules using pre-prepared power point slides produced by the DataONE Community Engagement and Education Working Group; 3) adding content about data management 4) participating in a workshop hosted by DataONE to refine and add additional content to educational modules (July, 2011).

Qualifications needed: A science data management background; Familiarity with aspects of the data lifecycle; Ability to quickly learn new software; Some work in development of educational materials helpful

Skills to be learned: Creative ways to educate a varied audience on data lifecycle; familiarity in use of chosen software used to develop online learning modules; collaboration techniques with dispersed working group.

Primary mentor: Viv Hutchison (USGS NBII)
Secondary mentors: Stephanie Hampton (National Center for Ecological Analysis and Synthesis), Carly Strasser (National Center for Ecological Analysis and Synthesis)

3. Understanding how scientists analyze data

Description: Scientists use a wide variety of tools and techniques to manage and analyze data. However, to our knowledge no one has taken a systematic look at how scientists do their work. In this project, we will examine a large number of the scientific workflows that have been constructed. We will develop a way of categorizing workflows based on their complexity, types of processing steps employed, and other factors. The goal is to
develop new and significant understanding of the scientific process and how it is being enabled by science workflows.

**Qualifications needed:** Self-starter, determined, enthusiastic, willing to keep a research notebook up-to-date openly online. Experience with a modern programming language, statistics and data analysis, and R would be helpful.

**Skills to be learned:** Kepler and Taverna workflow languages, research methods, research analysis, keeping an open science research notebook, communicating research results. A peer-reviewed publication is envisioned.

**Primary mentor:** William Michener (University New Mexico)

**Secondary mentors:** Rebecca Koskela (University of New Mexico), Bertram Ludaescher (University of California Davis)

4. **Accessing and analyzing environmental data in the classroom**

**Description:** A graduate student intern will create an educational module for use in undergraduate classrooms – the module will be designed to teach basic principles in ecology or environmental science using data that are publicly available through the DataONE network. The student will work with mentors to choose appropriate data sets, questions and analyses, and create a simple program to access and analyze the data in R. The student will create documentation that accompanies the exercise, potentially in multimedia formats, to train instructors to use the exercise in classrooms.

**Qualifications needed:** Basic background in ecology or environmental science, and statistics is necessary. Experience implementing statistics in a scripted statistical package such as R, Matlab or SAS is necessary. Experience with online training materials and multimedia presentation – e.g., screencasts - is useful.

**Skills to be learned:** The student will hone skills in statistical analysis, programming in R, working with large data sets, and creating teaching materials. The student will gain a well-rounded perspective on the importance of all aspects of the data life cycle in environmental sciences, and build a diverse professional network with leaders in environmental informatics and data-driven environmental science research.

**Primary mentor:** Stephanie Hampton (National Center for Ecological Analysis and Synthesis)

**Secondary mentors:** Carly Strasser (National Center for Ecological Analysis and Synthesis), Amber Budden (University of New Mexico)

5. **How much ecological data is out there?**

**Description:** No one is certain how much ecological data exists, or how this amount compares to the volume of data currently housed in repositories such as Knowledge Network for Biocomplexity (KNB). It would be useful to determine this for designing infrastructure, but also as a call to arms for ecologists to start sharing this “dark data”. For this project, we will develop a method for estimating the amount of ecological data being generated, with a focus on “small science” projects. Initially this project will involve brainstorming about the best way to estimate such a complex figure, and the intern will then be tasked with producing the estimate using the decided upon methods. Potential methods for estimation may include sampling publications, surveying scientists, or exploring existing databases. We foresee that results from this project will be highly cited since such an estimate is useful for discussions about data sharing, data reuse, and repository development in Ecology.
**Qualifications needed:** Applicants should be graduate students, have a strong background in the field of ecology or environmental science, and have statistics experience. Experience using computer scripts for data retrieval would be helpful, along with programming experience in R and/or MATLAB. The intern will need to be creative and excited about tackling complex problems.

**Skills to be learned:** The student will be exposed to topics in data management, reuse, and archiving, and will learn to work with ecological databases. They will learn to work collaboratively on complex problems with several members of the DataONE team, and have the opportunity to write a peer-reviewed publication with the potential for high citation rates. Particular skills related to computer scripting, statistics, and data mining will be specific to the methods determined by the student and mentors.

**Primary mentor:** Carly Strasser (National Center for Ecological Analysis and Synthesis)

**Secondary mentor:** Stephanie Hampton (National Center for Ecological Analysis and Synthesis)

6. **Tracking the reuse of 1000 datasets**

**Description:** We believe that openly archiving raw data facilitates valuable reuse. Can we measure this? What contribution does data reuse make to the published literature? Who reanalyzes data? For what? Does this vary across disciplines and repositories? These questions are the focus of an exploratory study, "Tracking data reuse: Following one thousand datasets from public repositories into the published literature." In this internship you'll work directly with Heather to collect, extract, annotate, and analyze data to explore these important questions. See [http://bit.ly/cPsek0](http://bit.ly/cPsek0) for more info on the project.

**Qualifications needed:** Self-starter, determined, enthusiastic, willing to keep a research notebook up-to-date openly online. Experience with statistics, the academic literature, PubMed, ISI Web of Science, Python, R, and blogging would be helpful.

**Skills to be learned:** Research methods, research data collection, text extraction from the scientific literature, keeping an open science research notebook, communicating research results

**Primary mentor:** Heather Piwowar (National Evolutionary Synthesis Center)

**Secondary mentor:** Todd Vision (University of North Carolina Chapel Hill/National Evolutionary Synthesis Center)

7. **Subsetting and publishing “dynamic” scientific datasets**

**Description:** The Avian Knowledge Network (AKN) is a federation of bird monitoring datasets, the largest and most dynamic of which is eBird. Datasets such as these, that are constantly being edited and expanded, are challenging to incorporate into the DataONE framework because of the way they are currently published. This project involves researching issues around dataset subsetting and duplication to recommend a publishing approach that works for “dynamic” datasets. Expected outcomes: (1) Implement that strategy by migrating the AKN repository to a DataONE-integrated Metacat deployment, making AKN into a DataONE Member Node; (2) Produce a case-study article that captures the implementation process that could act as a guide to future Member Nodes making similar efforts.

**Qualifications needed:** metadata mapping; high level programming language (e.g., Perl, Java); SQL; shell scripting

**Skills to be learned:** data repository implementation; scientific data organization and
8. Scientific workflow provenance repository and publishing toolkit

Description: Scientific workflow systems are increasingly used to automate scientific computations and data analysis and visualization pipelines. An important feature of scientific workflow systems is their ability to record and subsequently query and visualize provenance information. Provenance includes the processing history and lineage of data, and can be used, e.g., to validate/invalidate outputs, debug workflows, document authorship and attribution chains, etc. and thus facilitate "reproducible science". We aim to develop (1) a provenance repository system for publishing and sharing data provenance collected from runs of a number of scientific workflow systems (Kepler, Taverna, Vistrails), together with (2) a provenance trace publication system that allows scientists to interactively and graphically select relevant fragments of a provenance trace for publishing. The selection may be driven by the need to protect private information, thus including hiding, abstracting, or anonymizing irrelevant or sensitive parts. Part (1) will be based on a DataONE-extension of the Open Provenance Model (D1-OPM) and leverage an earlier Summer of Code project. In particular, the provenance toolkit includes an API for managing workflow provenance (i.e., uploading into and retrieving from a data storage back-end). Part (2) will implement a new policy-aware approach to publishing provenance, which aims at reconciling a user’s (selective) provenance publication requests, with agreed upon provenance integrity constraints. For an existing rule-based backend, a graphical user environment needs to be developed that lets users select, abstract, hide, and anonymize provenance graph fragments prior to their publication.

Qualifications needed: For Part 1, applicants should have experience in SQL and Java or a scripting language (e.g., Python or Perl). For Part 2, programming of GUIs with Rich Internet Application (RIA) technologies (e.g., GWT) is a plus.

Skills to be learned: Collaborative open source software development using state-of-the-art languages and tools (databases, workflow systems, interactive information visualization).

Primary mentor: Bertram Ludaescher (University of California Davis)
Secondary mentor: Paolo Missier (Newcastle University)

9. Integrating loosely structured data into the Linked Open Data cloud

Description: The Linked Data conventions describe four principles that allow data of any kind and from any online source to form a global interconnected web of data: i) name every "thing" that has some data or information associated with it; ii) use HTTP URIs to do so; iii) provide useful information or data in Resource Description Framework (RDF) format to someone looking up such URIs; and iv) within information provided this way, link to other common "things", such as points or axes of reference, and use common vocabularies to attach meaning to links wherever possible. These seemingly simple principles have nonetheless been highly effective in facilitating the creation of large, globally distributed, and constantly growing aggregations of Linked Open Data (LOD), a universally applicable framework for machines and users alike to integrate, navigate, and discover data by following links that are semantically of interest. Trying to apply the
Linked Data principles to data holdings of non-specialized digital repositories, such as DataONE and many of its member nodes, is challenging. These data are often highly heterogenous, and not natively expressed in RDF, or a format structured enough that would lend itself to automatic conversion to RDF. Instead, they are typically represented in formats that are either loosely structured in an ad-hoc manner (such as spreadsheets), or according to one of a myriad of formats output by instruments or analysis programs. It is thus not clear what the universe of "things" to name is, what are common points or axes of reference, what kinds (semantics) of links are needed, and how data archived in this way can be exposed in RDF such that the conversion can be automated, yet is still useful for science-motivated discovery and integration. The idea of this project is to develop an exploratory prototype, and practical recommendations resulting from it, for how the heterogeneous and loosely structured data held in non-specialized DataONE member nodes can be exposed to the Linked (Open) Data cloud. The approach would consist of obtaining a sufficiently representative sample of data sets from DataONE's initial 3 member nodes (Dryad, KNB, and ORNL-DAAC), and using them as instance data for which to define the RDF predicate vocabularies, domain ontologies, resource URIs, and conversion mechanisms that are necessary to create a LOD representation of these data. This representation can then be uploaded to, navigated, and queried in either one of the web-based LOD browsers (such as URIburner), or for example in a local installation of OpenLink Virtuoso.

**Qualifications needed:** Knowledge of RDF and one of its widely used serializations (XML, N3). Familiarity with either C or Java programming, or a scripting language that has good support for RDF and OWL, will be needed. Familiarity with Linked Data, and experience with metadata vocabularies and domain ontologies in RDF and OWL will be very helpful.

**Skills to be learned:** Designing and executing an exploratory study through all phases. Identifying and communicating alternatives and their advantages and drawbacks. Developing practical semantic web resources for existing instance data.

**Primary Mentor:** Hilmar Lapp (National Evolutionary Synthesis Center)

10. ** Developing video animations for DataONE community engagement**

**Description:** DataONE wishes to develop a set of video animations to help explain DataONE’s value and capabilities to a range of audiences. Several topics have been identified for these short animations, a couple of storyboards have been developed, and one animation created. The intern will work with the mentors to continue building this set of animations according to the principles of ‘universal design’.

**Qualifications needed:** Applicants should have strong visual design skills and a high level of expertise in development of digital animation. Expertise in communicating scientific information to a variety of audiences is desirable.

**Skills to be learned:** Video / animation development; science communications.

**Primary mentor:** Paul Allen (Cornell Laboratory of Ornithology)

**Secondary mentors:** Amber Budden (University of New Mexico), Will Morris (Cornell Laboratory of Ornithology)

This information is also available at: [http://www.dataone.org/content/2011-summer-internship-program](http://www.dataone.org/content/2011-summer-internship-program)

Rebecca Koskela
Executive Director, DataONE
University of New Mexico
1312 Basehart SE
Albuquerque, NM 87106

Email: rkoskela@unm.edu
Office (M,W,F): (505) 814-1111
Cell & Office (T,Th): (505) 382-0890
Fax: (505) 246-6007