# Member Node Description: KNB Data Repository

Version 2.0 6/1/2015 Matt Jones

## General

| | |
|---|---|
| **Name of resource:** | KNB Data Repository (KNB) |
| **URL(s):** | http://knb.ecoinformatics.org/ |
| **Institutional affiliation(s):** | • National Center for Ecological Analysis and Synthesis, UC Santa Barbara<br>• Long-term Ecological Research Network (LTER)<br>• Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)<br>• Organization of Biological Field Stations (OBFS)<br>• Ecological Society of America<br>• South African National Parks<br>• UC Natural Reserve System<br>• Many others, representing several hundred field stations… |
| **Primary geographic location:** | Global, with primary nodes in Santa Barbara, CA |
| **Project Director & contact info:** | Matthew B. Jones, jones@nceas.ucsb.edu |
| **Technical Contact & contact info:** | Matthew B. Jones, jones@nceas.ucsb.edu |
| **Age of resource:** | Since 1999 |
| **Funding support:** | NSF, Mellon Foundation, Moore Foundation |
| **Proposed Unique Identifier:** | urn:node:KNB |

## Content

### Content description/collection policy (1 paragraph, domain and spatial/temporal coverage, uniqueness of content, exclusions, as applicable):

Data in the KNB Data Repository is principally from the ecological and environmental fields, but includes other data from allied disciplines as well, such as genomics data, physiology, behavior, evolution, economics, sociology, policy, and related fields. Data represents survey and monitoring data as well as many types of manipulated data from field and laboratory experiments.

### Types of data (complex objects, text, image, video, audio, other):

Tabular relational data, vector and matrix data, raster images, vector images, audio, and software and code from analysis and modeling.

### Data and metadata availability (rights, licensing, restrictions):

Intellectual rights subject to laws in the US, but otherwise set by contributor. The KNB requires a CC-BY or more permissive license for any data that are made public to allow the KNB system to legally redistribute data and metadata. The KNB system provides separate access control for metadata and data objects, and provides mechanism for provider to state intellectual rights requirements.

### Option for embargo (yes/no, duration):

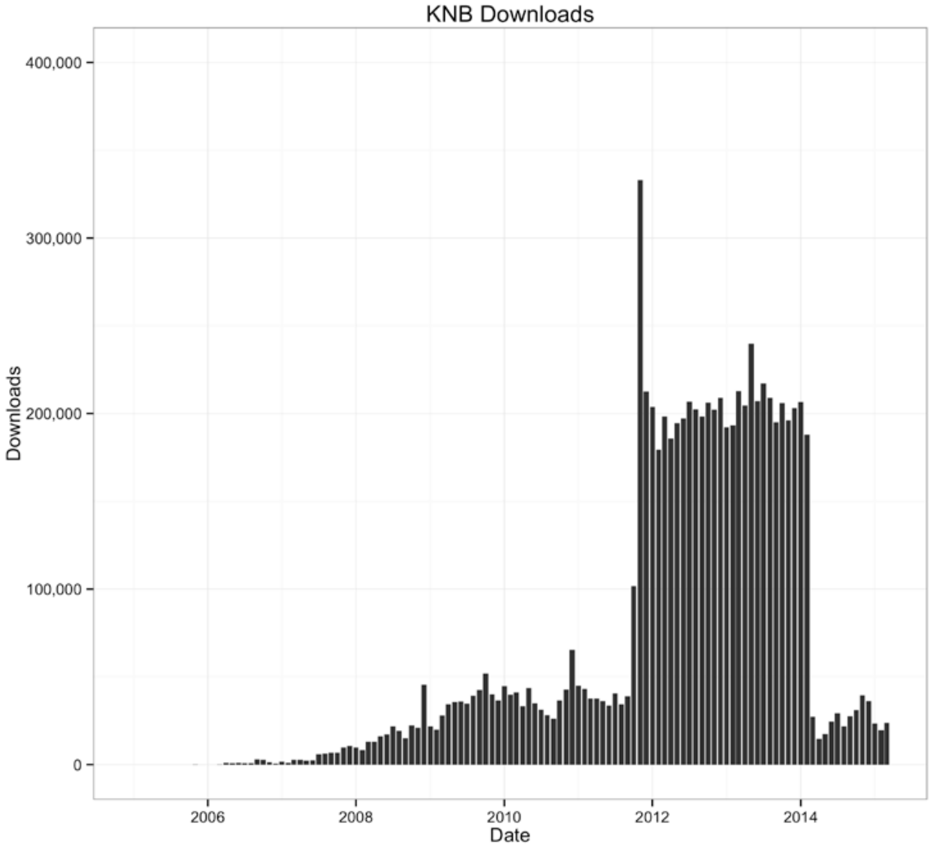
Yes, accomplished through fine-grained access control as specified by data contributor.

### Size of holdings (number and size of datasets, mean and median granules (files) per dataset):

The KNB hosts over 22,000 data packages containing more than 19,000 individual data files. Details of these data and metadata are continuously updated online through a set of statistics about the repository (see https://knb.ecoinformatics.org/#profile).

### Please describe recent usage statistics, if known, including information on annual data product downloads, annual number of users, annual number of data products used in publications:

The KNB tracks all data accesses and reports these to DataONE, and displays accesses in the web display for each data package so that contributors and users can assess how often data are accessed. Monthly data downloads typically range from 25,000 to 200,000 objects per month (see Figure):



## User interactions

### How does a user contribute data? (what can be deposited, how are data prepared, are specific software required, documentation/support available)

Data can be contributed in any format, but open standards like CSV, NetCDF, etc. are encouraged. Metadata can be in any format that is serialized in a valid XML document. Common metadata types include EML, BDP, MoML, ISO19139. Many specialized software packages for a variety of environments have been written to contribute data and metadata (e.g., web-based registries, Morpho, MATLAB scripts, etc.). Metacat includes two harvester tools, including an OIA-PMH harvester, and several open APIs for client tools to directly contribute data to the system.

### How does a user acquire / access data?

By using search services to locate metadata and data packages online, and then can download data either through the web or directly from within analytical tools such as R or Matlab.

### What user support services are available (both for depositing and accessing/using data)?

Mailing lists are maintained for all data submission portals, and there is a general knb-help@nceas.ucsb.edu available

Software tools have their own support mailing lists (morpho-dev@ecoinformatics.org)

### How does the resource curate data at the time of deposit?

Depends on the KNB partner. Some are caveat emptor submissions, others have dedicated data managers, and others have moderated submissions with peer-review and metadata revision before acceptance.

## Technical characteristics and policies

### Software platform description, incl. data search and access API(s):

The KNB is an interconnected set of Metacat data and metadata management servers that are linked in a hub and spoke model. The NCEAS site serves as the central index for all metadata on the KNB, and other Metacat instances are distributed worldwide. Contributors can choose to replicate data and metadata at other repositories within the DataONE network. DataONE Access APIs are web services using REST styles of interaction and are fully supported.

### Service reliability (including recent uptime statistics, frequency of hardware refresh, if known):

Uptime has always been > 364 days per year. Minimal scheduled outages occur for hardware refresh and OS upgrades.

Hardware updated approximately every 3 years as needed to maintain performance and reliability.

### Preservation reliability (including replication/backup, integrity checks, format migration, disaster planning):

Metadata and data fully replicated among two distributed sites (NCEAS and LTER) that are completely independent and geographically isolated. Servers are backed up daily.

### User authentication technology (incl. level of create/modify/delete access by users):

CIlogon identities are used to authenticate and utilize the DataONE API to manage data in the KNB. A distributed LDAP system is used to federate accounts across multiple institutions within the KNB itself.

Users authenticate to Metacat and then can specify fine grained (read/write/changePermission) access rules for both metadata and data objects in the system.

### Data identifier system and data citation policy, if available:

Each data package in the KNB can be assigned a DOI identifier, which allows citation of the dataset with a clear resolution mechanism to determine the current location of the data. Every data and metadata object contains a unique id from any number of identifier schemes (e.g., DOI, UUID, ARK, LSID, etc.), and strong versioning ensures long-term object availability (i.e., no objects can be overwritten, updates provide a new version of an object without removing the original). Data citation policy specified by contributors in intellectual rights statement, but citation is encouraged and provided in each record's HTML view.

### Metadata standards (including provenance):

- Ecological Metadata Language (EML)
- Biological Data Profile (BDP)
- ISO 19139
- Modeling Markup Language (MoML)

- others

## Capacity/services to DataONE

**At what functional tier will you initially be operating? (see http://bit.ly/MNFactSheet for definitions)**

□ **Tier 1: Read only, public content**
□ **Tier 2: Read only with access control**
□ **Tier 3: Read/write using client tools**
☒ **Tier 4: Able to operate as a replication target**

**If you can host data from other member nodes, what storage capacity is available?**
Yes. Terabytes of storage are available.

**Can you provide computing capacity to the broader network? If so, please describe.**
Limited. Some data subsetting and query services are provided, but no general computational services.

## Other Services

**What other services or resources (such as expertise, software development capacity, educational/training resources, or software tools) can be provided of benefit to the broader network?**
Extensive expertise in all aspects of federated system design and implementation, cross-organizational, coordinated, software development for data management tools, database and systems management.

Provide periodic training sessions on all aspects of informatics.