Member Node Description: Northwest Knowledge Network

Version 1.0 3/12/15 Luke Sheneman

General

Name of resource:	Northwest Knowledge Network (NKN), University of Idaho
URL(s):	www.northwestknowledge.net
Institutional affiliation(s):	University of Idaho
Primary geographic location:	Moscow, Idaho, 83844
Project Director & contact info:	Dr. Paul Gessler, NKN Director, (paulg@uidaho.edu) Dr. Luke Sheneman, Technology Manager, (sheneman@uidaho.edu)
Technical Contact & contact info:	Ed Flathers, Data Portal Developer, (flathers@uidaho.edu) Dr. Luke Sheneman, Technology Manager, (sheneman@uidaho.edu) David Vollmer, Systems Administrator, (dvee@uidaho.edu)
Age of resource:	Since 2011
Funding support:	Institutional University Support, Grant Funding (NSF, USGS, etc.)
Proposed Unique Identifier:	urn:node:NKN

Content

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

NKN manages diverse collections of research data across many disciplines, with a strong emphasis on earth science, ecology, environmental, climate, and hydrology. We focus on archiving data specific to Idaho and the Pacific Northwest region. Much of our data is collected under recent or current federally funded grants such as NSF EPSCoR, USDA REACCH, or USGS Climate Science Center. Our largest data collections include raw and processed regional LiDAR products, downscaled climate scenario data (netCDF), statewide high resolution orthoimagery data, and legacy stream ecology datasets.

Most of our hosted data is unique and is not archived elsewhere. We specifically exclude human medical records and financial data of any kind that incur special privacy considerations.

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

Our data portal is data format agnostic and we support the archival of any kind of data format. To date, the bulk of our collections are represented as netCDF, raw LiDAR (.LAS), or TIFF.

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

NKN allows multiple licensing schemes for our published data and metadata. In some cases, this is determined by the project that funded the collection of the data. Our default license is the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license.

Option for embargo (yes/no, duration):

All uploaded and submitted data is scanned for malware held for content review by data portal administrators. All data/metadata requires administrator approval before being published and made available to the public.

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

With the upcoming ingestion of several Idaho NSF EPSCoR downscaled climate prediction datasets, NKN will be hosting approximately 100 datasets (with many more planned for ingestion/archival in the near future). In addition, through NSF EPSCoR we collaborate with New Mexico and Nevada. We will begin replicating data and metadata between these three institutions soon. Current granules per dataset: mean: 671, median: 336

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:

Our data access has increased steadily over the past three years. Access to downscaled climate data and statewide orthoimagery data consititute the majority of our recent data demand and access. NKN has over 450 registered users in different contexts, many of which are registered users of our data portal. We do not maintain any statistics available on the annual number of data products used in publications.

User interactions

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

The primary method for depositing data is through the NKN data portal. This portal is a customized version of the ESRI Geoportal Server that serves as our metadata catalog and provides a web-based upload tool for contributing data assets. Any registered user can upload a data resource. Datasets comprised of multiple files (e.g. a shapefile) should be zipped prior to uploading. All uploaded data are automatically inspected for malware/viruses and rejected if they are infected. Users then describe uploaded datasets using an online metadata editor that enforces a certain level of completeness and is based on the ISO-19115-2 standard. Once a dataset has been uploaded and documented through a complete metadata record, the user can request that the data be uploaded. A site administrator must inspect and approve data/metadata prior to publication. Once an inspector inspects and approves a dataset, that metadata will be published in the catalog and users can download data objects.

If users have very large or complex data to contribute, they can work with NKN staff to help perform the bulk import and metadata work on the backend. This can sometime be more efficient than using the upload tool and metadata editor.

How does a user acquire / access data?

Users perform searches against our metadata catalog in our data portal (ESRI Geoportal Server). Metadata records provide information on access mechanisms for data (download URL, or web service). Currently, all published data is publicly accessible.

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

NKN staff is available to help users at every level of using our data portal. We provide contact information on our portal, including support email address and phone number. We have and will continue to provide custom hands-on support for users with special needs (e.g. uploading very large or complex collections).

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

By design, all submitted data must be accompanied by standard metadata. Our preferred metadata standards are ISO 190115-2, FGDC, and EML. Before a metadata record (and associated data) is published and searchable, our online metadata editor enforces a certain level of completeness. Before data is published, an NKN administrator must inspect and approve the data/metadata. All uploaded data resources are supplied with an internal unique NKN unique identifier and stored to disk.

Uploaded content is synchronously replicated to an external datacenter at Idaho National Laboratory (INL) to better preserve these data in the event of catastrophic failure.

Technical characteristics and policies

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

We use a customized version of the ESRI Geoportal Server for our metadata catalog. We added an upload tool (PLUPLOAD) and integrated it with Geoportal Server. Data search is done through the ESRI Geoportal Server (which uses Solr/Lucene indexing). We expose our catalog through HTTP and through OGC-CSWweb services (as implemented in ESRI GeoPortal).

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

We have deployed Nagios monitoring software to monitor our entire enterprise and improve our security and uptime profile. From our experience, the NKN infrastructure and related service layers are very robust. Since inception, we have had no significant hardware failures that have materially impacted our data portal availability. We periodically, but infrequently, take down the portal to perform scheduled OS or software updates.

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

We maintain distributed server and storage hardware in two geographically distinct places (Moscow, ID and Idaho National Laboratory in Idaho Falls, ID). Incremental backups of all critical data are automatic and occur nightly. In addition, our Storage Area Network (SAN) provides block-level snapshots that perform hourly, nightly, weekly, and monthly snapshots at BOTH locations. Our GeoPortal also computes checksums for all data objects within our repository and tracks these in a database to ensure data integrity at the file level.

While we have a very robust and distributed architecture, we currently have no formal written disaster recovery plan.

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

We use LDAP for our authentication technology, and specifically the OpenLDAP platform. Our portal integrates with LDAP for authentication. Our LDAP system is administered by our dedicated systems administration staff using command-line and web tools. User registration, account creation, and password management has been integrated into our GeoPortal.

Data identifier system and data citation policy, if available:

We are currently using internal UUIDs to uniquely identify data resources within our portal. NKN has a subscription to the California Digital Library EZID system to provision and allocate Digital Object Identifiers (DOIs). One member of our technical staff serves as the dedicated EZID liaison and manually provisions DOIs as needed for specific data sets within NKN. NKN has plans to automatically allocate and assign DOIs for each "published" data resource using the EZID web service API. Once a data resource is published, we consider it immutable and citable and will be assigned a global DOI. We expect automatic DOI minting to be integrated with our portal by the end of the 2015 calendar year.

Our data citation policy (and complete legal framework for our portal) is available in our Terms of Service that is accessible here:

https://www.northwestknowledge.net/terms_of_service

Metadata standards (including provenance):
We currently require all uploaded resources to use a standard metadata format. In particular, we encourage the ISO-19115-2 standard and our online metadata editor is built around this standard. We also accept and injest other metadata standards such as FGDC CSDSM, EML, and Dublin Core.

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?
Our intention is to begin by allocating 5TB as a replication target. We can allocate up to 15TB to DataONE if needed.

Can you provide computing capacity to the broader network? If so, please describe.
We currently have unused server capacity, especially at Idaho National Laboratory (INL). We would be willing to spin up one or more virtual machines and make compute resources available to the DataONE network if needed/desired.

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?
NKN has an excellent working relationship with the University of Idaho Library and we work with library faculty and staff on issues of resource archival, curation, and metadata management. NKN has particular expertise in these areas due to our relationship with the UI Library. This University library/cataloging expertise may set NKN apart from many other data management groups and presents a unique resource that might be of interest to the larger DataONE network.

NKN has considerable expertise in IT systems architecture and has deployed a robust, scalable, distributed, network and server infrastructure. We would like to help DataONE leverage this architecure and this expertise in whatever way makes sense.

We have expertise and capacity in software development, especially with respect to building data management websites using Geoportal Server and the Drupal content management system. Our current development staff is fully committed on projects, but we have the capacity to develop and extend our own portal and tools.