



# Metacat

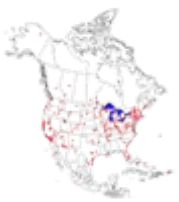
Duane Costa

LTER Network Office  
University of New Mexico

Adapted, with permission, from  
Matthew Jones

National Center for Ecological Analysis and Synthesis  
University of California, Santa Barbara

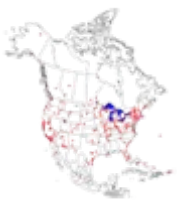




# Roadmap

- Part I: Introduction to Metacat capabilities
  - **Overview**
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

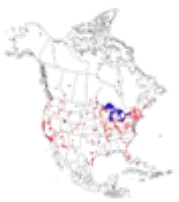




# Metacat

- Flexible storage system for storing and accessing metadata and data
  - Stores arbitrary metadata documents (requires XML)
  - Supports structured searches
  - Customizable web interface
  - Replication capabilities
  - Works on Linux, Windows, MacOS
    - Oracle, Postgres, MS SQL Server





# KNB Overview

## Clients



Web Browser

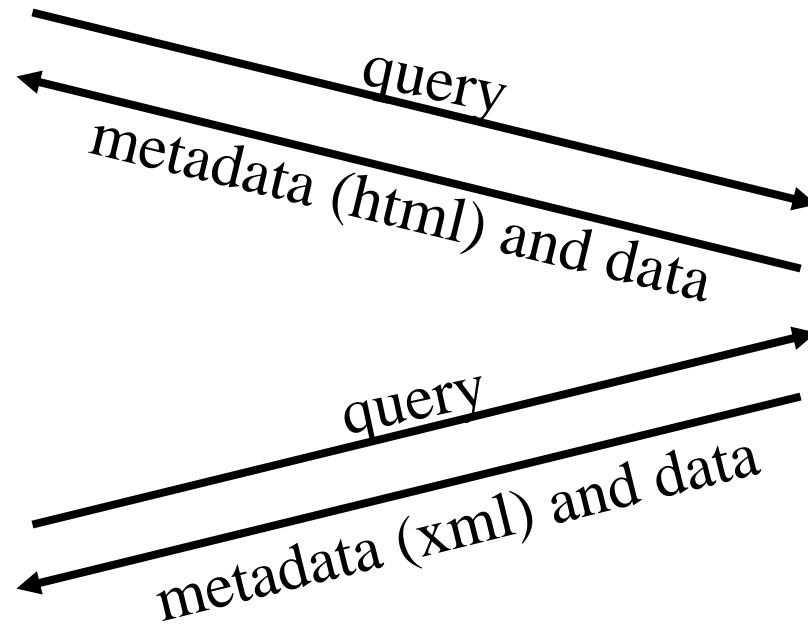


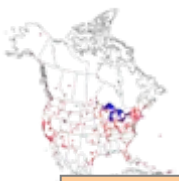
Morpho

Metadata  
(EML)

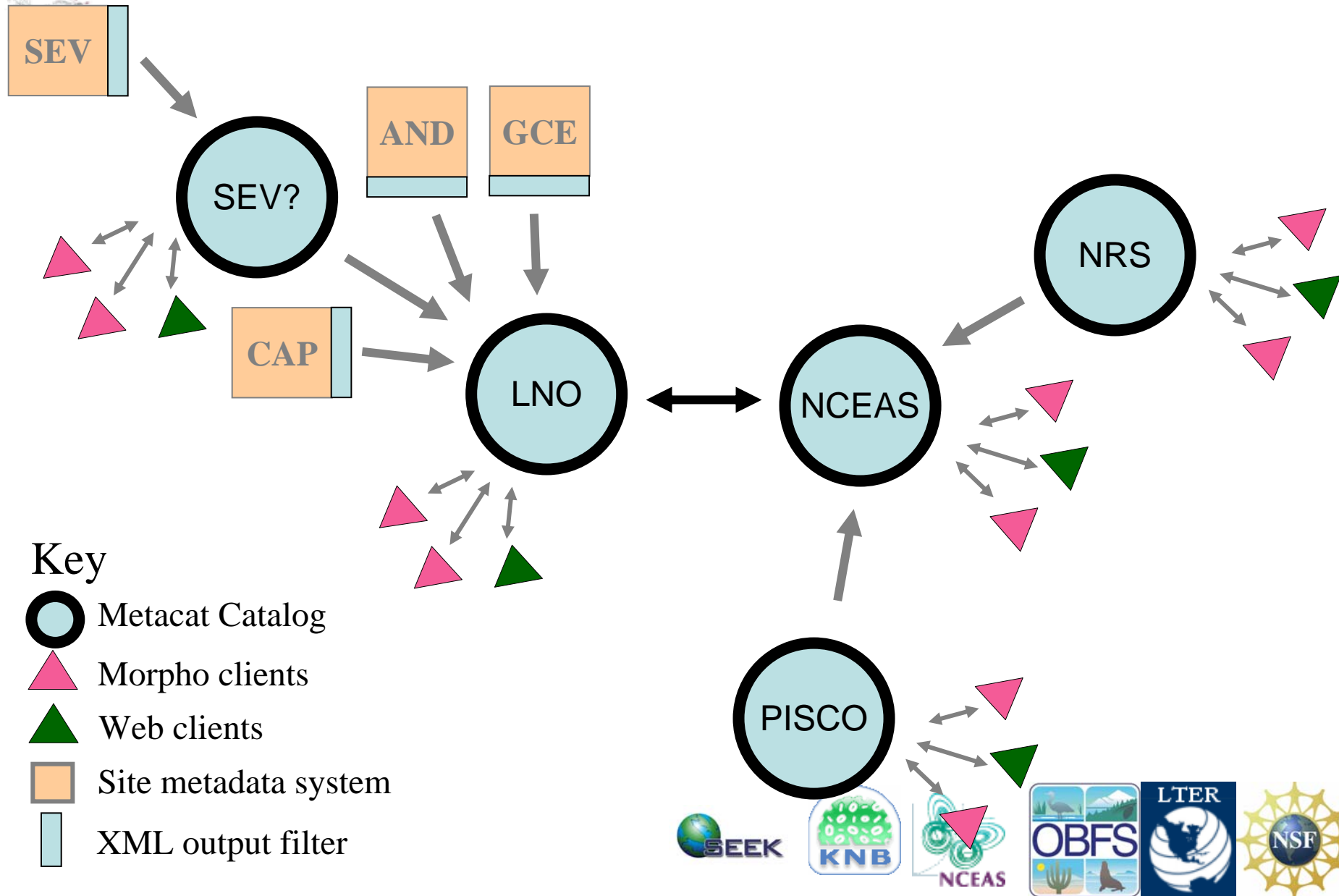
Data

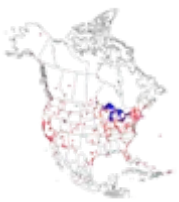
## Server



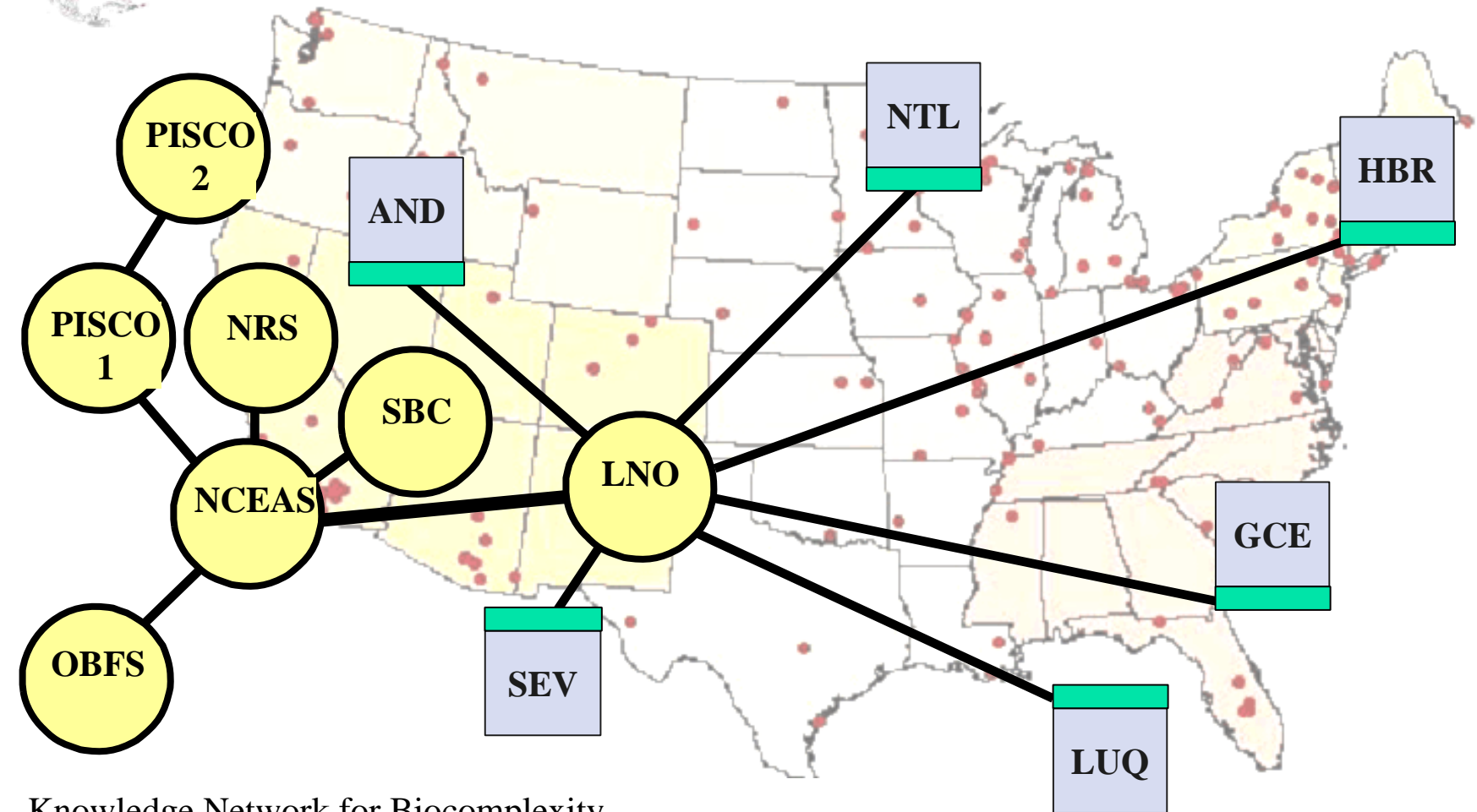


# Building the KNB network

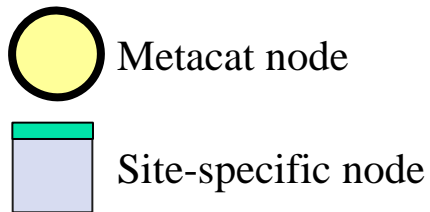


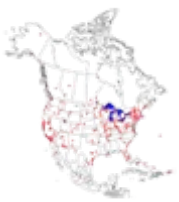


# Knowledge Network for Biocomplexity



Knowledge Network for Biocomplexity





# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - **Metacat web interface**
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting





# Metacat web user interface

**KNB :: The Knowledge Network for Biocomplexity - Mozilla**

## The Knowledge Network for Biocomplexity

The **Knowledge Network for Biocomplexity (KNB)** is a national network intended to facilitate ecological and environmental research on biocomplexity.

For scientists, the KNB is an efficient way to discover, access, interpret, integrate and analyze complex ecological data from a highly-distributed set of field stations, laboratories, research sites, and individual researchers.

### search for data on the KNB

**You ARE logged in (Logout).** You may search the KNB without being logged into your account, but will have access only to "public" data (see "login & registration")

Enter a search phrase (e.g. biodiversity) to search for data sets in the KNB, or click "advanced search" to enter more-detailed search criteria, or simply browse by category using the links below.

» advanced search «

#### Taxonomy

Amphibian, Bird, Fish, Fungus, Invertebrate, Mammal, Microbe, Plant, Reptile, Virus

#### Level of Organization

Molecule, Cell, Organism, Population, Community, Landscape, Ecosystem, Global

#### Ecology

Biodiversity, Competition, Decomposition, Disturbance, Endangered Species, Herbivory, Invasive Species, Nutrient Cycling, Parasitism, Population Dynamics, Predation, Productivity, Succession, Symbiosis, Trophic Dynamics

#### Measurements

Biomass, Carbon, Chlorophyll, GIS, Nitrate, Nutrients, Precipitation, Temperature, Radiation, Weather,

#### Evolution

Adaptation, Evolution, Extinction, Genetics, Mutation, Selection, Speciation, Survival

#### Habitat

Alpine, Freshwater, Benthic, Desert, Estuary, Forest, Grassland, Marine, Montane, Terrestrial, Tundra, Wetland

### login & registration

Logging into your account enables you to search any additional, non-public data for which you may have access privileges:

**You ARE logged in**

username:

organization:

password:

[create a new account](#)  
[forgot your password?](#)  
[change your password](#)

### Data Management Software

Morpho is easy-to-use data-management software. Use it to:





- query, view, retrieve and manipulate ecological data from the KNB network
- create, view and manipulate your own datasets, and specify access control to manage their availability

**Morpho: more information and download**

**Quick Download for:**  
Windows :: Mac OS X :: Linux

[:: more information about Morpho](#)

Sponsored and developed by:

National Center for Ecological Analysis and Synthesis   Texas Tech University   Long Term Ecological Research Network   San Diego Supercomputer Center

**Mozilla**

## Biocomplexity Data Search

Home

### search for data on the KNB

**You ARE logged in (Logout).** You may search the KNB without being logged into your account, but will have access only to "public" data (see "login & registration")

Enter a search phrase (e.g. biodiversity) to search for data sets in the KNB, or click "advanced search" to enter more-detailed search criteria, or simply browse by category using the links below.

» advanced search «

#### Taxonomy

Amphibian, Bird, Fish, Fungus, Invertebrate, Mammal, Microbe, Plant, Reptile, Virus

#### Level of Organization

Molecule, Cell, Organism, Population, Community, Landscape, Ecosystem, Global

#### Ecology

Biodiversity, Competition, Decomposition, Disturbance, Endangered Species, Herbivory, Invasive Species, Nutrient Cycling, Parasitism, Population Dynamics, Predation, Productivity, Succession, Symbiosis, Trophic Dynamics

#### Measurements

Biomass, Carbon, Chlorophyll, GIS, Nitrate, Nutrients, Precipitation, Temperature, Radiation, Weather,

#### Evolution






Adaptation, Evolution, Extinction, Genetics, Mutation, Selection, Speciation, Survival

#### Habitat

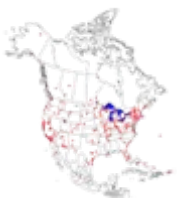
Alpine, Freshwater, Benthic, Desert, Estuary, Forest, Grassland, Marine, Montane, Terrestrial, Tundra, Urban, Wetland

### 456 data packages found

Title	Contacts	Organization	Keywords
<b>Datos meteorologicos</b>	Virinia Perez		
ID: VIR.4.1			
<b>Productivity, Diversity and Soil Data from two North American Grasslands</b>	Doe		
ID: bowles.450.1			
<b>Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed at GCE9_Hydro (Altamaha River near Rockdedundy Island, Georgia) from 25-Feb-2002 through 31-Dec-2002</b>	Sheldon Blanton	Georgia Coastal Ecosystems LTER Project	temperature sonde Sea-Bird salinity pressure mooring MicroCAT density ctd conductivity
ID: knb-lter-gce.87.4			





# Metacat UI is reconfigurable

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

**DATA CATALOG**  
Search

data catalog  
search

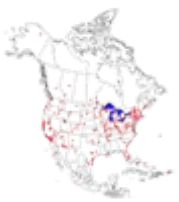
*Category Search:* *Other Search:*

Oceanographic Sensor Data    Microchemistry Data  
Intertidal Community Survey Data    Intertidal Recruitment Data  
Subtidal Community Survey Data    Subtidal Recruitment Data  
SBCLTER Demo Data

15 data packages found:

Title	Contacts	Organization	Associated Data
SBCLTER: Land: Hydrology: Precipitation Data (sbclter.992.3)	Melack	SBCLTER	SBCLTER Rain Gauge Locations.csv
SBCLTER: Land: Hydrology: SBCLPWD Precipitation Data (sbclter.718.12)	Melack	SBCLTER	SBCFC_Precip_Daily.txt SBCLPWD Rain Gauge Locations.csv
SBCLTER: Land: Hydrology: USGS Stream Discharge Data (sbclter.318.32)	Melack	SBCLTER	11114000day.txt 11118500day.txt 11118501day.txt 11119500day.txt 11119745day.txt 11119750day.txt 11119780day.txt 11119940day.txt 11120000day.txt 11120500day.txt 11120510day.txt 11120530day.txt 11120550day.txt USGS_Stream_Discharge_Locations.csv USGS_Discharge_Flag_Codes.csv
SBCLTER: Land: Stream Chemistry: Stream Samples (sbclter.379.27)	Melack	SBCLTER	river_chem_2001_DB.csv riverchemlocations.csv
SBCLTER: Land: Watershed Characteristics: GIS Layers (sbclter.385.6)	Melack	SBCLTER	SBCLTER Stream Sample Locations
SBCLTER: Land Ocean Reef: Foodweb Stable Isotopes.	Reed	SBCLTER	Isotope_Data.csv

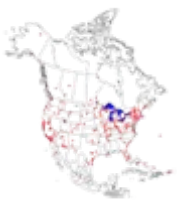




# Roadmap

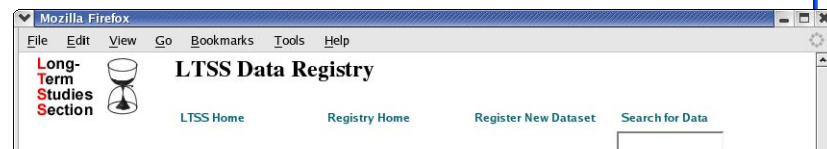
- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - **Registries and Repositories**
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

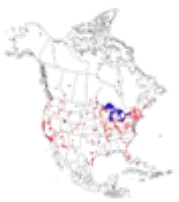




# Data Registries

- UC NRS Information System
  - NRS Network, NCEAS
- Resource Discovery Initiative for Field Stations (RDIFS)
  - LTER Network, OBFS Network, NCEAS, San Diego Supercomputer Center, University of Kansas
- Use metacat
  - Web-based metadata entry

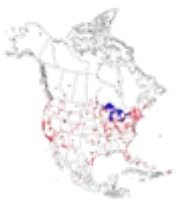




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - **Features**
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

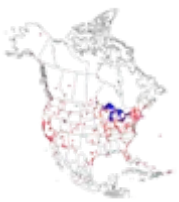




# Metacat features

- Metadata
  - Store & search any XML formatted metadata
    - Ecological Metadata Language (EML)
    - NBII Biological Data Profile
    - FGDC CSDGM
    - Site specific formats
  - Metadata validation
    - Configure to accept particular metadata formats
  - Enforces access control rules
  - Metadata conversion (using XSLT)
    - To HTML for presentation
    - To other metadata formats (e.g., NBII)
- Data
  - Storage
  - Access control



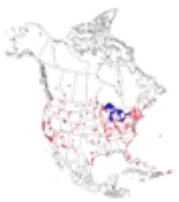


# Metacat implementation

- Java servlet for portability
  - Linux, Windows 2000, MacOS X
- HTTP access
  - Standard POST and GET queries
- Web HTML interface via XSLT transforms
  - Separates content from presentation
- Interfaces with RDBMS for storage
  - Oracle, PostgreSQL, (SQL Server) backend





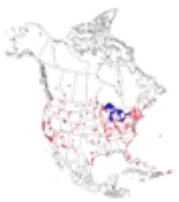


# Roadmap

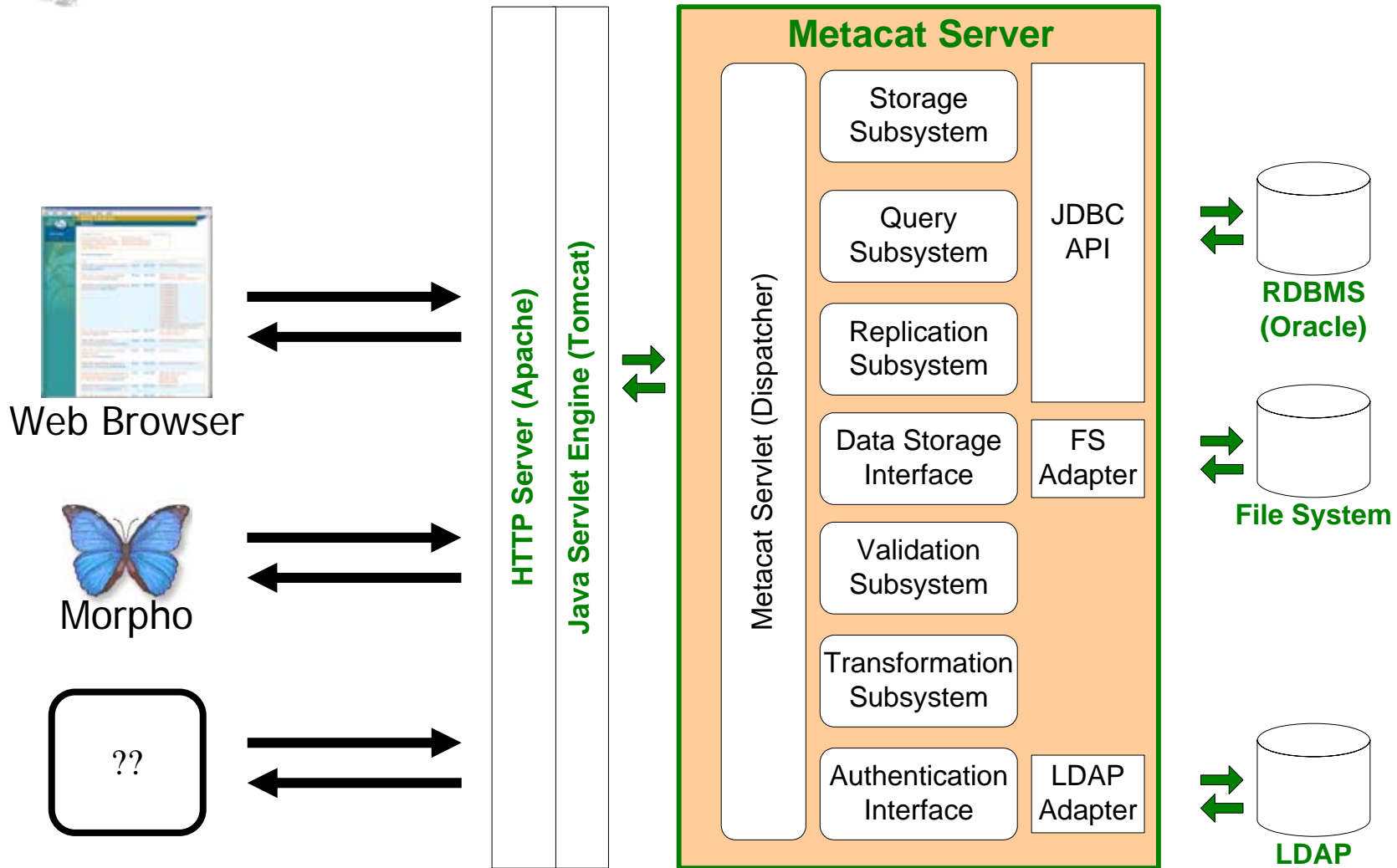
- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - **Architecture overview**
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

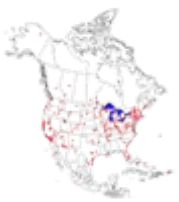






# Metacat architecture

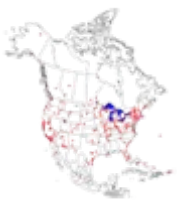




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - **Storage subsystem**
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

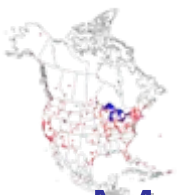




# Storage subsystem

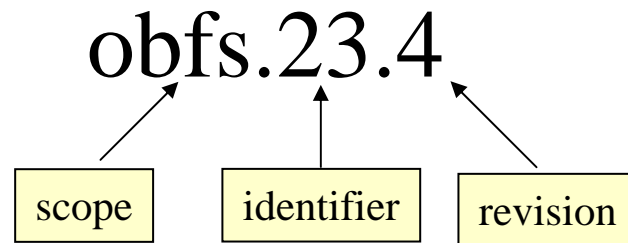
- Storage
  - XML metadata stored in relational db
    - Oracle, PostgreSQL, (SQL Server)
  - Data object storage on filesystem
- Data viewed as opaque objects
  - Assigned unique identifier
  - Metadata describes data structure & semantics





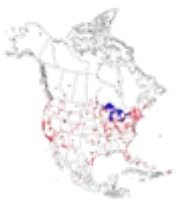
# Versioning and Identifiers

- Metacat prescribes a format for identifiers
  - Identifiers are a contract regarding uniqueness
  - Incorporates both 'identity' and 'version'
  - Two data streams with the same ID are defined as identical
  - 'insert' requires a unique ID
  - 'update' requires an existing ID with a new revision



- Will be adopting Life Science Identifiers (LSID)  
E.g., "**urn:lsid:lsid.ecoinformatics.org:obfs:26:3**"

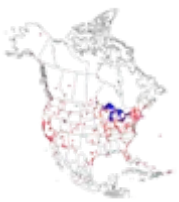




# Storage and retrieval actions

- Read
  - Download a document
- Insert
  - Put a new XML document in the database
- Update
  - Replace an existing xml document with a new version, incrementing the identifier
- Delete
  - Archive a document so that it does not show up in searches
- Upload
  - Put a binary or other non-xml in the file system

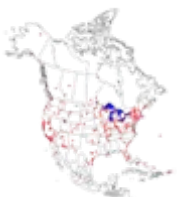




# Reading a document

- Simple HTTP GET or POST request
- <http://a.com/knb/metacat?action=read&docid=knb-lter-gce.109.5>
- Return document is in XML format by default
- Login is optional
  - If you don't login, you have 'public' privileges





# A simple web client

**MetaCat - Mozilla**

File Edit View Go Bookmarks Tools Window Help

File:///C:/Documents%20... Search

## MetaCat XML Loader

Upload, Change, or Delete an XML document using this form.

1. Choose an action: ☒ Insert ☐ Update ☐ Delete

2. Provide a Document ID

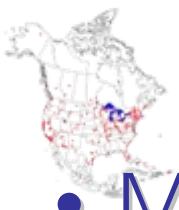
3. Provide XML text (not needed for Delete)

```
<?xml version="1.0"?>
<shoppingList>
  <item>milk</item>
  <item>oreos</item>
  <item>orange juice</item>
</shoppingList>
```

4. Provide DTD text for upload (optional; not needed for Delete)



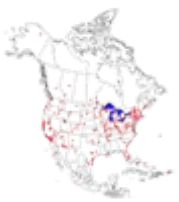




# Schema-independence

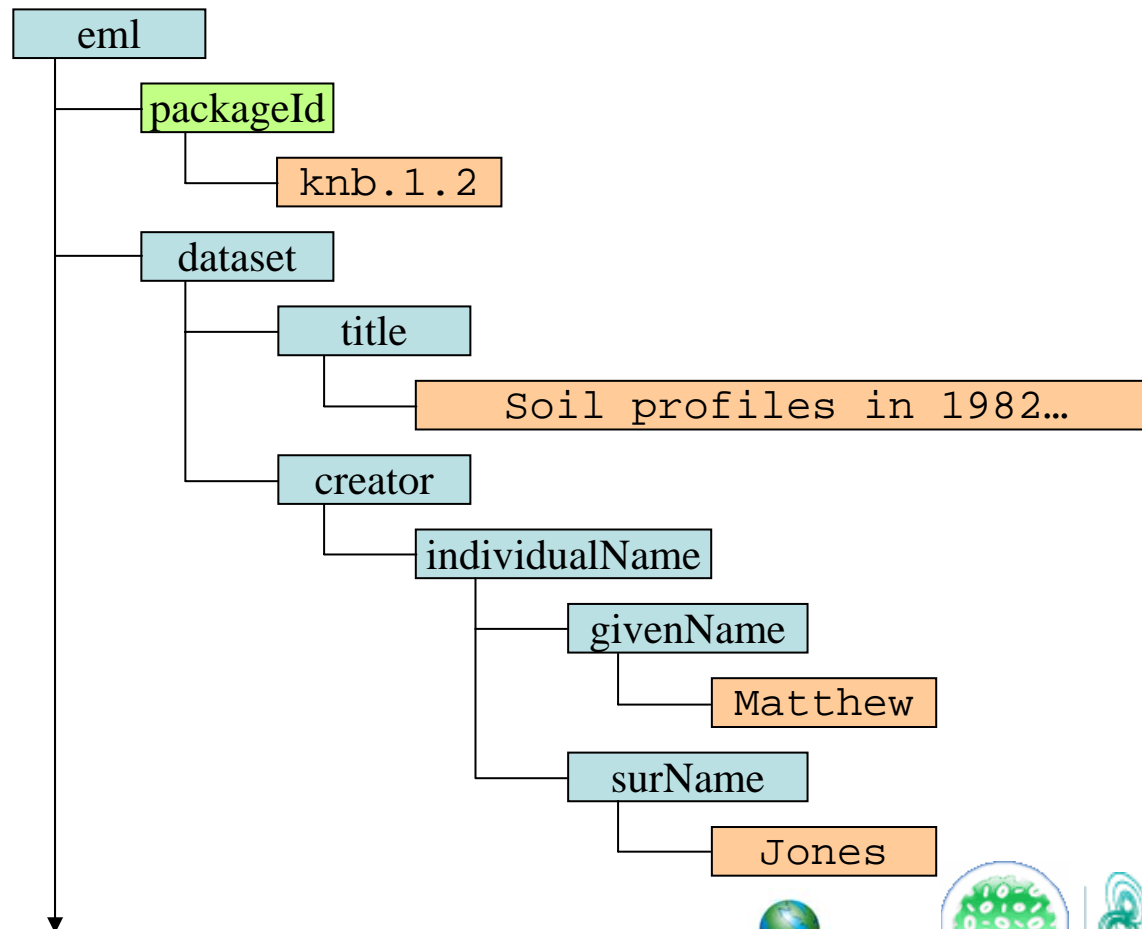
- Most relational db's support only one data model
  - Makes maintenance as models change expensive
- Metacat is “schema independent”
  - Any XML document, regardless of schema, can be stored without modifications to metacat
  - Metacat's data model follows the XML Document Object Model (DOM)
    - Thus, it models the XML structure rather than the data schema





# DOM

- DOM models hierarchical element and attribute structure of XML



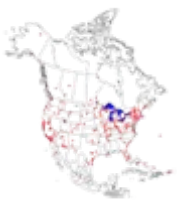
## Node types

element

attribute

text





# Yeah, so what?

- You can throw whatever you need into metacat (without schema or software changes)

- And you can query it

```
<?xml version="1.0"?>
```

```
<poll>
```

```
  <favoriteOS id="1">MacOS</favoriteOS>
```

```
  <favoriteOS id="2">Linux</favoriteOS>
```

```
  <favoriteOS id="3">Linux</favoriteOS>
```

```
  <favoriteOS id="4">Linux</favoriteOS>
```

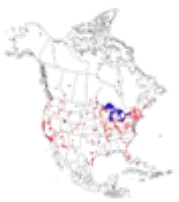
```
  <favoriteOS id="5">WinXP</favoriteOS>
```

```
  <favoriteOS id="6">Linux</favoriteOS>
```

```
  <favoriteOS id="7">Linux</favoriteOS>
```

```
</poll>
```

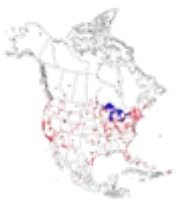




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - **Query subsystem**
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

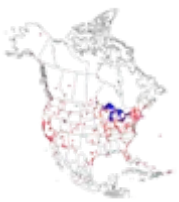




# Query subsystem

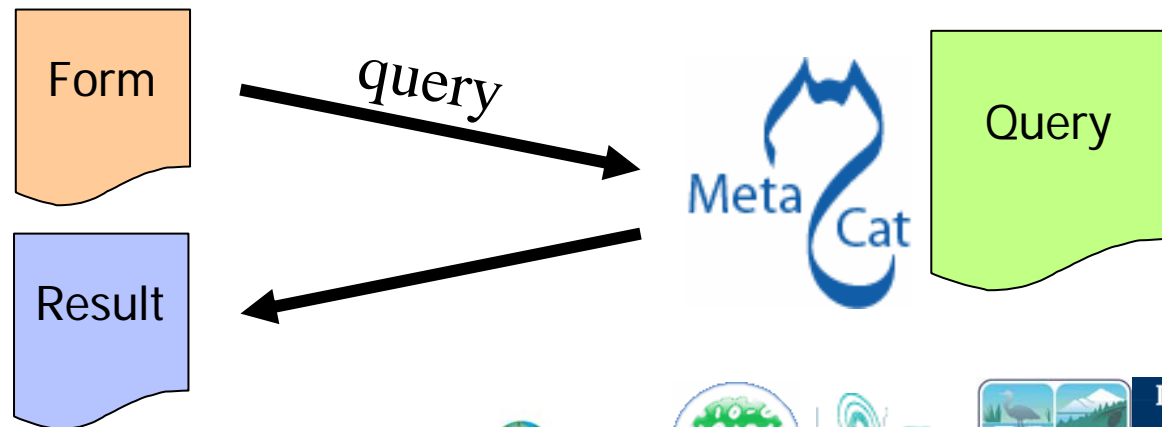
- Two means of submitting queries
  - Query action (query)
    - Query parameters passed as url-encoded form parameters (i.e., an html form)
    - Metacat builds a pathquery document automatically
  - Structured query action (squery)
    - Custom query syntax in xml format (pathquery)

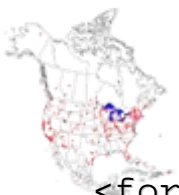




# HTML form queries

- Parameters passed as form elements
  - ‘Special’ fields
    - action, qformat, operator
    - returnfield, returndoctype
    - “anyfield”
  - Other fields create additional conditions
- Metacat builds the query from the fields





# Example HTML form

Form

Query

```
<form method="POST" action="@servlet-path@" target="_top">
```

Search for:

```
<input name="action" value="query" type="hidden">
```

```
<input name="operator" value="INTERSECT" type="hidden">
```

```
<input name="anyfield" type="text" value="" size="14">
```

```
<input name="organizationName" value="Organization of Biological Field  
Stations" type="hidden">
```

```
<input name="qformat" value="xml" type="hidden">
```

```
<input name="returnfield" value="creator/individualName/surName"  
type="hidden">
```

```
<input name="returnfield" value="creator/individualName/givenName"  
type="hidden">
```

```
<input name="returnfield" value="creator/organizationName" type="hidden">
```

```
<input name="returnfield" value="dataset/title" type="hidden">
```

```
<input name="returnfield" value="keyword" type="hidden">
```

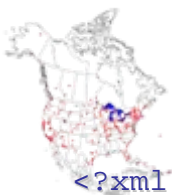
```
<input name="returndoctype" value="eml://ecoinformatics.org/eml-2.0.1"  
type="hidden">
```

```
<input value="Start Search" type="submit">
```

```
</form>
```







# Query Result Set Structure

Result

```
<?xml version="1.0"?>
```

```
<resultset>
```

```
<document>
```

```
<docid>knb.2.1</docid>
```

```
<docname>eml</docname>
```

```
<doctype>eml://ecoinformatics.org/eml-2.0.1</doctype>
```

```
<doctitle>Soil profiles from lower Yosemite Valley</doctitle>
```

```
<createdate>2000-06-10 12:54:07</createdate>
```

```
<updatedate>2000-06-10 12:54:07</updatedate>
```

```
<param name="/eml/dataset/creator/individualName/surName">Levings</param>\
```

```
<param name="/eml/dataset/creator/individualName/surName">Shriver</param>
```

```
<param name="/eml/dataset/keywordSet/keyword">strata</param>
```

```
<param name="/eml/dataset/keywordSet/keyword">mineralization</param>
```

```
</document>
```

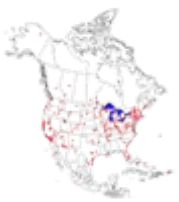
```
<document>
```

```
...
```

```
</document>
```

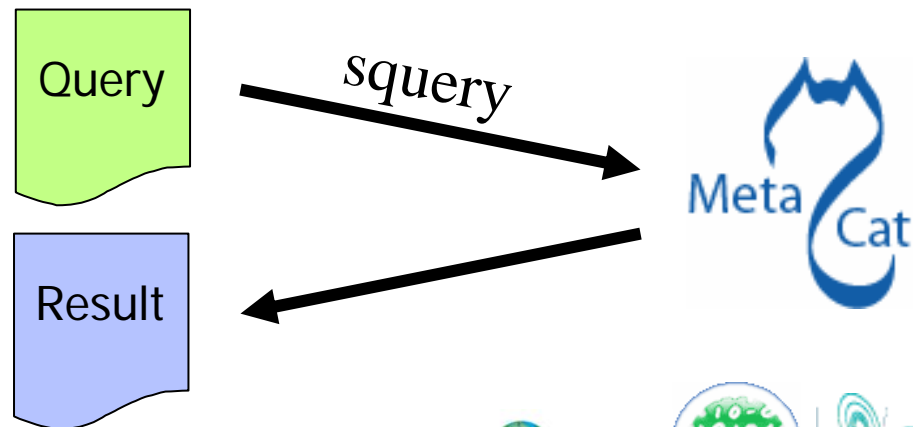
```
</resultset>
```

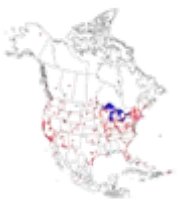




# Structured queries

- Pathquery syntax
  - Can build precise queries against arbitrary metadata schemas
    - Boolean combinations of conditions (AND, OR)
      - Uses Xpath-like syntax
    - Specify document types to search
    - Specify fields to return in resultset





# Query Conditions

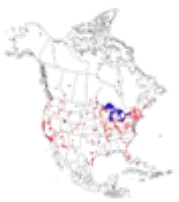
Query

- Language independent representation of a query structure
- Transformed into the appropriate native language of the data store

## Example:

```
<querygroup operator="UNION">  
  <queryterm searchmode="contains" casesensitive="false">  
    <value>soil</value>  
    <pathexpr>dataset/title</pathexpr>  
  </queryterm>  
  <queryterm searchmode="contains" casesensitive="false">  
    <value>nutrients</value>  
    <pathexpr>dataset/title</pathexpr>  
  </queryterm>  
</querygroup>
```





# Specifying the Resultset

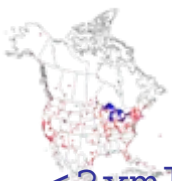
Query

- Specify the list of fields to be returned in the resultset
- Simple paths used to identify elements or document subtrees
- Effectively flattens the structure of the records, but allows generic representation (i.e, multiple standards)

## Example:

```
<returnfield>dataset/title</returnfield>  
<returnfield>creator/individualName/surName</returnfield>  
<returnfield>keyword</returnfield>
```





# Full Query Example

Query

```
<?xml version="1.0"?>
```

```
<pathquery version="1.2">
```

```
  <querytitle>Soil search</querytitle>
```

```
  <returndoctype>eml://ecoinformatics.org/eml-  
    2.0.0</returndoctype>
```

```
  <returnfield>creator/individualName/surName</returnfield>
```

```
  <returnfield>keyword</returnfield>
```

```
  <querygroup operator="UNION">
```

```
    <queryterm searchmode="contains" casesensitive="false">
```

```
      <value>soil</value>
```

```
    </queryterm>
```

```
    <queryterm searchmode="contains" casesensitive="false">
```

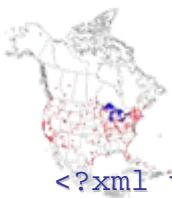
```
      <value>nutrients</value>
```

```
    </queryterm>
```

```
  </querygroup>
```

```
</pathquery>
```





# Query Result Set Structure

Result

```
<?xml version="1.0"?>
```

```
<resultset>
```

```
<document>
```

```
<docid>knbn.2.1</docid>
```

```
<docname>eml</docname>
```

```
<doctype>eml://ecoinformatics.org/eml-2.0.1</doctype>
```

```
<doctitle>Soil profiles from lower Yosemite Valley</doctitle>
```

```
<createdate>2000-06-10 12:54:07</createdate>
```

```
<updatedate>2000-06-10 12:54:07</updatedate>
```

```
<param name="/eml/dataset/creator/individualName/surName">Levings</param>\
```

```
<param name="/eml/dataset/creator/individualName/surName">Shriver</param>
```

```
<param name="/eml/dataset/keywordSet/keyword">strata</param>
```

```
<param name="/eml/dataset/keywordSet/keyword">mineralization</param>
```

```
</document>
```

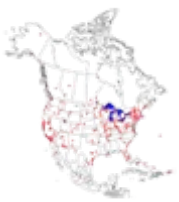
```
<document>
```

```
...
```

```
</document>
```

```
</resultset>
```



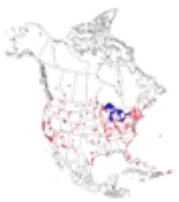


# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - **Transformation subsystem**
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting



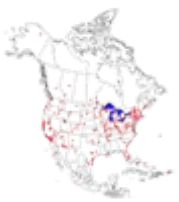




# Transforming a document

- Used to convert document before returning it
- Conversion uses XSLT
- Configuration of which style sheet to use is controlled by the skin via the 'qformat' parameter
- <http://knb.ecoinformatics.org/knb/metacat?action=read&docid=knb-lter-gce.109.5&qformat=ltss>
- Return document is converted and returned
- The '*skinname.xml*' file controls the mappings that determine which style sheet to use

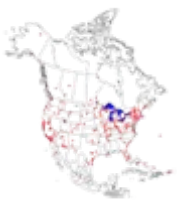




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - **Authentication subsystem**
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

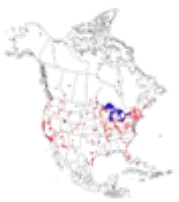




## Authentication subsystem

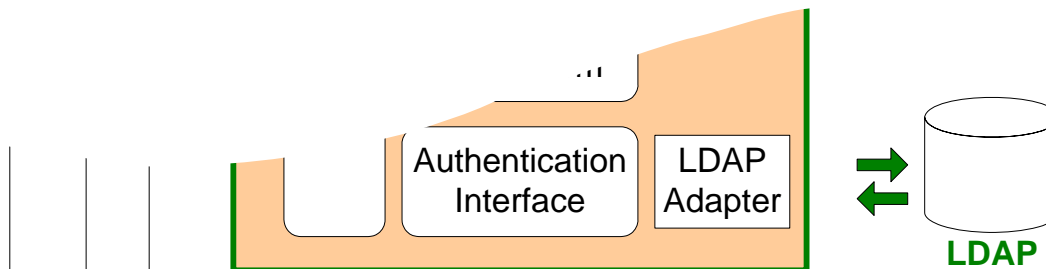
- Actions: 'login' and 'logout'
- Simple username/password system
- Successful login creates a session
  - Session ID tracked using an HTTP cookie

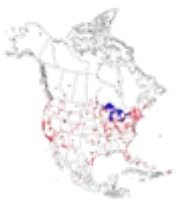




# Authentication plug-ins

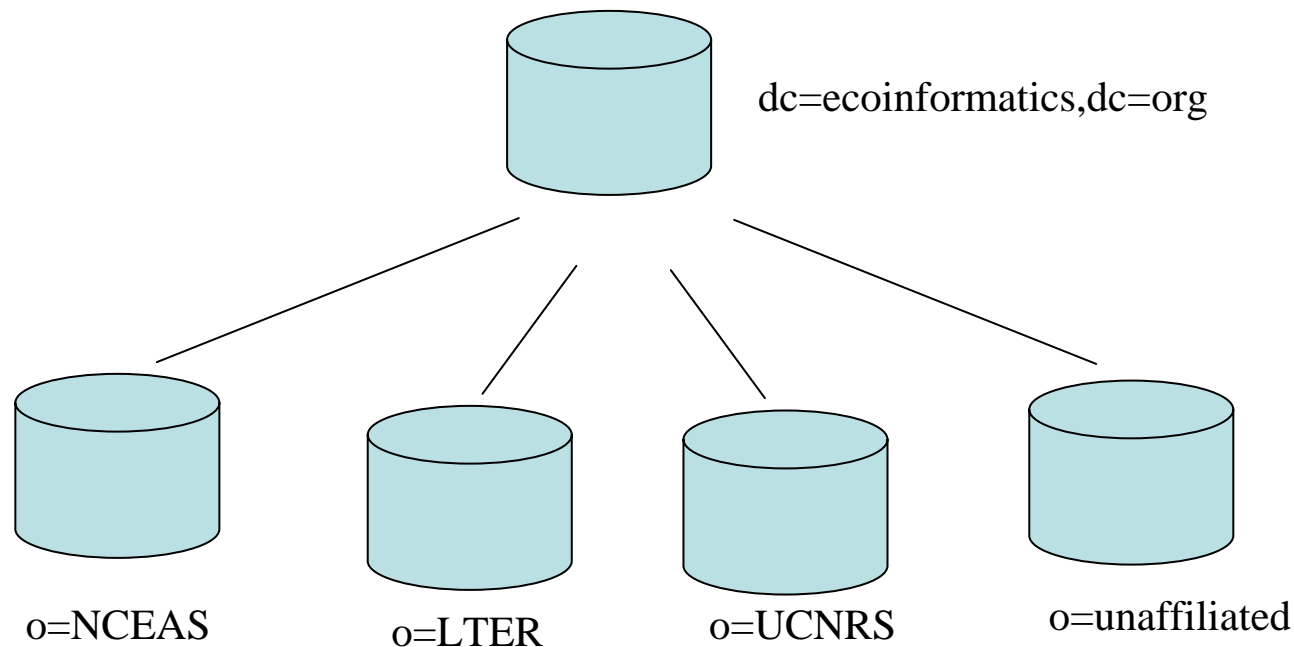
- Delegates authentication requests to a backend service via a plug-in
  - Lightweight Directory Access Protocol (LDAP)
  - Replaceable to interface with other systems
- Metacat administrator can choose which LDAP server to use

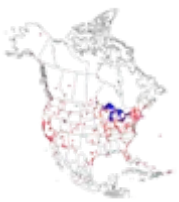




# Ecoinformatics.org LDAP

- Need for community-wide user identities
- Distributed system for participating institutions
- Root LDAP server refers requests to specific organizations for authentication

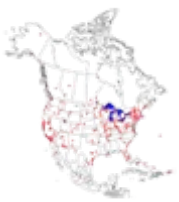




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - **Other subsystems**
  - Client API
- Part III: Metacat Advanced Topics
  - Replication and Harvesting



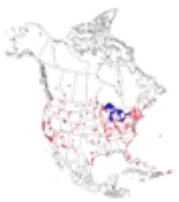


## Other subsystems

- **action=validate**
  - valtext, docid
- **action=setaccess**
  - docid, principal, permission, permType, permOrder, principal
- **action=getversion**
- **action=getlog**
  - ipaddress, principal, docid, event, start, end
  - <http://68.111.43.225:8080/knb/metacat?action=getlog&event=insert>



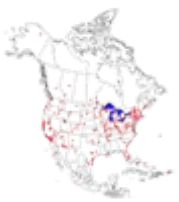




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - **Client API**
- Part III: Metacat Advanced Topics
  - Replication and Harvesting

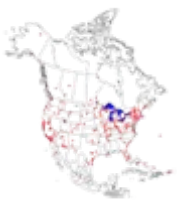




# Client API

- Application Programming Interface (API)
  - Defines language-specific binding for communicating with Metacat
  - Available in Java and Perl (partially available in Python)
- Allows development of new applications
- Allows integration of Metacat with existing applications
- Simple set of method calls

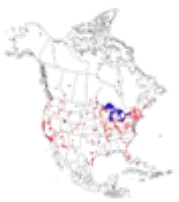




# Basic Client API

- `public String login(String username, String password)`
- `public String logout()`
- `public Reader read(String docid)`
- `public Reader query(Reader xmlQuery)`
- `public String insert(String docid, Reader xmlDocument, Reader schema)`
- `public String update(String docid, Reader xmlDocument, Reader schema)`
- `public String delete(String docid)`
- `public String upload(String docid, File file)`
- `public String upload(String docid, String fileName, InputStream fileData, int size)`





# Example use of client

```
String metacatUrl = "http://foo.com/context/metacat";  
String username = "uid=jones,o=NCEAS,dc=ecoinformatics,dc=org";  
String password = "neverHarcodeAPasswordInCode";  
try {
```

```
Metacat m = MetacatFactory.createMetacatConnection(metacatUrl);
```

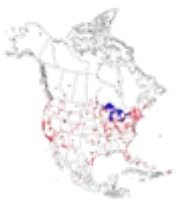
```
m.login(username, password);
```

```
Reader r = m.read("testdocument.1.1");
```

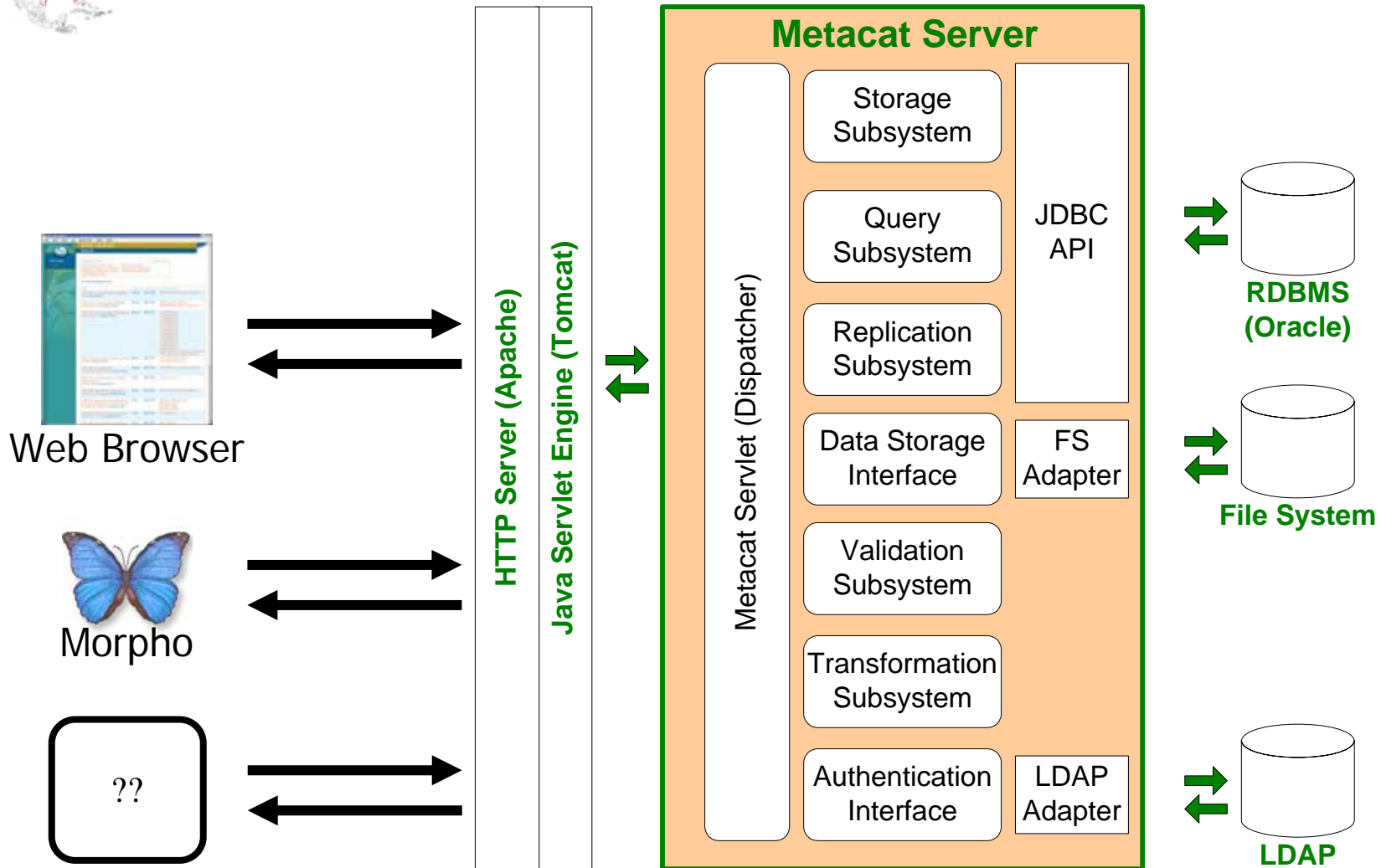
```
// Do whatever you want with Reader r
```

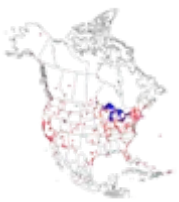
```
} catch (MetacatAuthException mae) {  
    handleError("Authorization failed:\n" + mae.getMessage());  
} catch (MetacatInaccessibleException mie) {  
    handleError("Metacat Inaccessible:\n" + mie.getMessage());  
} catch (Exception e) {  
    handleError("General exception:\n" + e.getMessage());  
}
```





# Review






# Documentation

KNB Software: Metacat: Metacat Tour Home - Mozilla

File Edit View Go Bookmarks Tools Window Help

 **Metacat Tour Home**

[KNB Home](#) [Data](#) [People](#) [Informatics](#) [Biocomplexity](#) [Education](#) [Software](#)

---

### Background Information



- [XML documents](#)
- [XML tree](#)
- [DOM API](#)

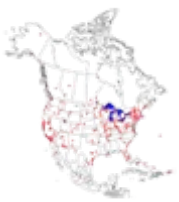
### Basic Functionality

- [Metacat Server](#)
- [Metacat Database](#)
- [Metacat Document Object Model \(DOM\)](#)
- [Metacat Client Programming Interface](#)
- [Writing XML](#)
- [Searching the Database](#)
- [Reading data from Metacat](#)
- [Event Log Reporting](#)

### Advanced Functionality

- [Document parsing](#)
- [XML Indexing](#)
- [Access Control](#)
- [LDAP for Authentication](#)

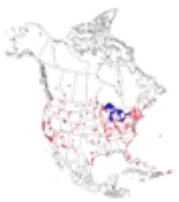


# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- **Part III: Metacat Advanced Topics**
  - **Replication and Harvesting**

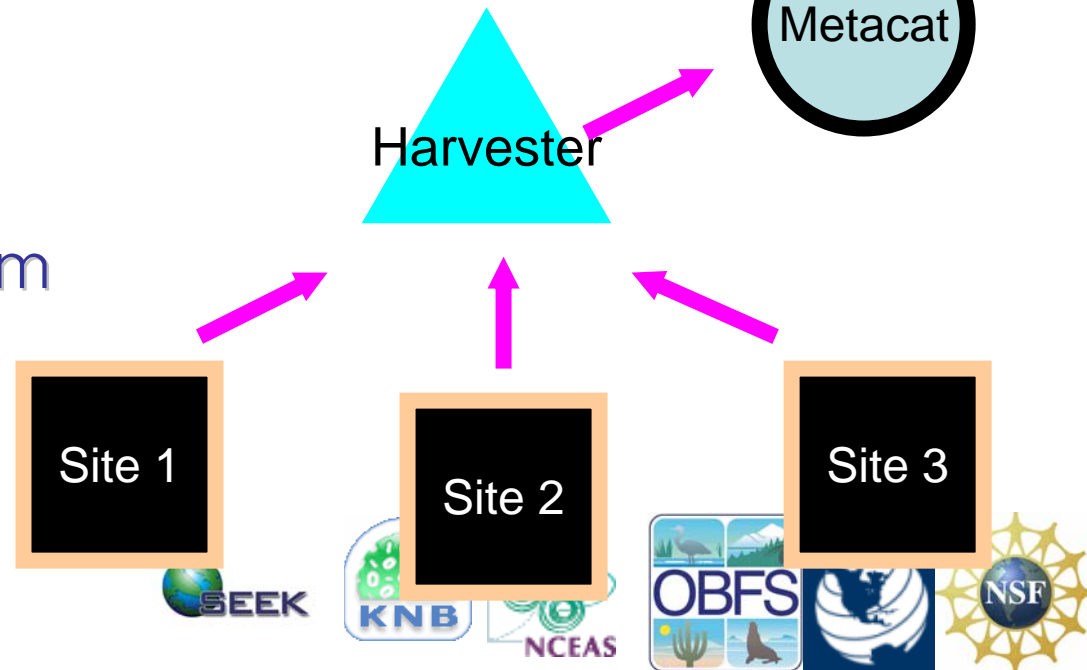
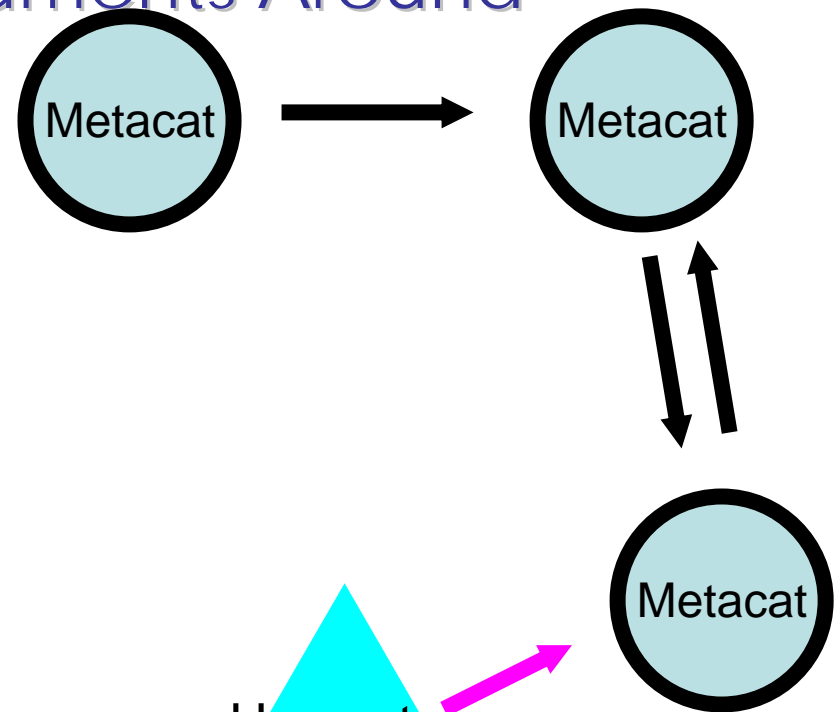


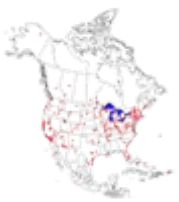




# Replication and Harvesting: Two Ways to Move Metacat Documents Around

- Replication
  - Synchronize content between 2 metacat servers
- Harvesting
  - Scheduled 'pull' of XML documents from web sources

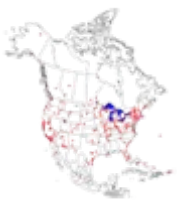




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- Part III: Metacat Advanced Topics
  - **Replication** and Harvesting

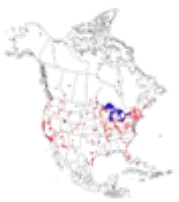




## Rationale for Replication System

- Distributed searches are slow, unreliable, **up-to-date**
- Centralized metadata searches are **fast, reliable**, potentially less up-to-date
- Metacat replication provides ***best of both***: centralized (**fast, reliable**) search of metadata that is always kept **up-to-date** via replication

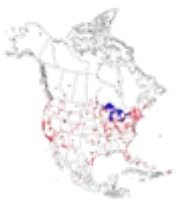




# Metacat Replication: Design Goals

- Data must remain consistent on each server
  - Metacat uses file locking to maintain consistency among multiple versions of documents
- Every document has a home server where the master copy of the document resides
  - Only a document's home server can give a lock to another server for that file to be altered
- Allow one-way replication
  - Some Metacat servers may want to share their data with other Metacat servers but not want to receive outside data onto their servers

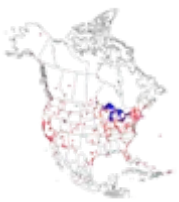




## Metacat Hubs and Non-Hubs

- A Metacat server that is a non-hub can *only* replicate documents whose home server is itself
- A Metacat server that is a hub can replicate *both* its own documents and documents that were replicated to it from other servers

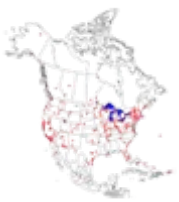




## Two Different Replication Mechanisms

- Event-based notification
  - Each replication server is notified when a document is inserted, updated, or deleted
- Delta-T monitoring
  - Checks each replication server on at regular time intervals, e.g. once every 30 seconds, once every 24 hours, or once per week





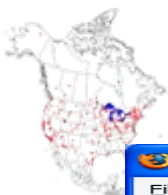
# Replication Table: xml\_replication

(Think *push*, not pull)

serverid	server	last_checked	repl cate	datareplicate	hub
1	localhost	null	0	0	0
2	aaa.xxx.edu/knb/servlet/ replication	2005-10-30 11:00:00	0	0	0
3	bbb.yyy.edu/ Metacat/servlet/ replication	2005-10-30 11:00:01	1	1	0
4	ccc.zzz.edu/knb/servlet/r eplication	2005-10-30 11:00:02	1	1	1







# Metacat Replication Control Panel

**MetaCat - Mozilla Firefox**

File Edit View Go Bookmarks Tools Help

http://knb.lternet.edu:8088/knb/style/skins/dev/replControl.html

Albuquerque weather... Books24x7.com Java 2 Platform SE v... KNB :: The Knowled... LTER Intranet - Serv...

Found 17 tickets

**Metacat Replication Control Panel**

[KNB](#) [Data](#) [People](#) [Informatics](#) [Biodiversity](#) [Education](#) [Software](#)

**Time Handler**

☒ Start Delta T:  seconds

☐ Stop

**Force Replication**

☐ Get All - bring all updated documents from remote hosts to this server

**Servers**

☐ Remove this server

☐ Add this server

Replicate xml doc To server (1 or 0)?:

Replicate data file To server (1 or 0)?:

Localhost is a hub to server (1 or 0)?:

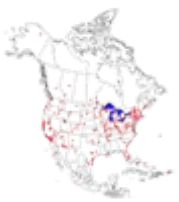
Download certificate from

[Refresh Server List](#)

server	last_checked	replicate	datareplicate	hub
localhost	null	0	0	0
knb.ecoinformatics.org/knb/servlet/replication	2005-01-26 08:03:49.0	1	1	1

Done



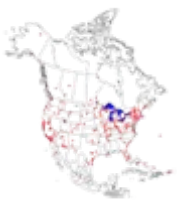


# Replication Security: Keys and SSL

- Replication in six easy steps (for Tomcat standalone)
  - *Step 1* Using **keytool**, I generate a key in my Java keystore.
  - *Step 2* Using **keytool**, I generate a certificate for the key that I can give to you.
  - *Step 3* I modify my Tomcat configuration to activate my SSL port, 8443, and tell Tomcat where to find my Java keystore
  - *Step 4* Using **keytool**, I import your certificate into my Java keystore. (You do the same with my certificate.)
  - *Step 5* I restart Tomcat.
  - *Step 6* I use the Replication Control Panel to add your server to my replication table. (You do the same in your replication table.)

*Now we're replicating!*

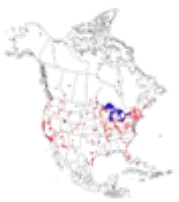




# Roadmap

- Part I: Introduction to Metacat capabilities
  - Overview
  - Metacat web interface
  - Registries and Repositories
  - Features
- Part II: Metacat Design and Architecture
  - Architecture overview
  - Storage subsystem
  - Query subsystem
  - Transformation subsystem
  - Authentication subsystem
  - Other subsystems
  - Client API
- **Part III: Metacat Advanced Topics**
  - Replication and **Harvesting**

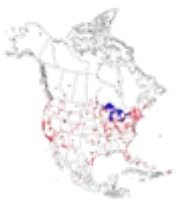




# Metacat Harvester

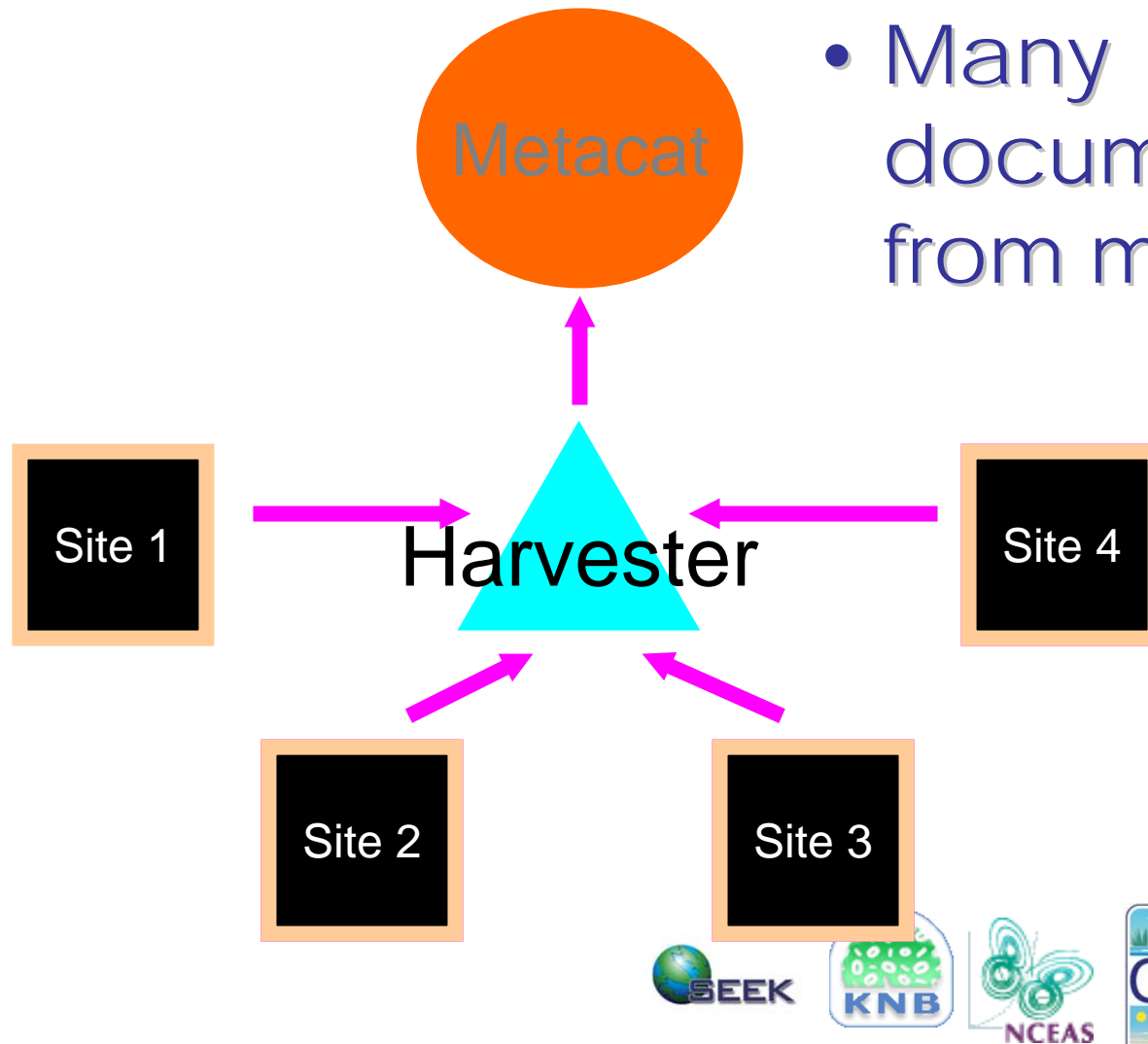
- Harvester provides a convenient mechanism for batch upload of EML documents to Metacat on a scheduled basis, potentially adding large numbers of documents to the Metacat repository
- Bundled with Metacat distribution (beginning with Metacat 1.4.0), but using Harvester is optional

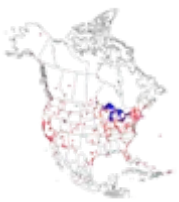




# Harvesting to Metacat

- Server-side pull
- Many documents from many sites

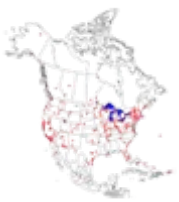




## Who Should Use Harvester?

- Your EML documents were created with a tool other than Morpho
- Your EML documents are dynamically generated
- Your EML documents are frequently revised and you'd like them to be automatically re-harvested



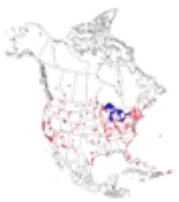


## Harvester Features

- Each site controls its own harvest schedule
- Generates and sends email reports after each harvest
- Logs Harvester operations in Metacat DB
- Works with dynamically generated EML







# Harvester Definitions

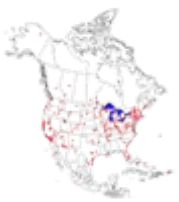
- Harvester Administrator

The individual who installs and manages Harvester (typically the same person who installs and manages Metacat)

- Harvest Site

A remote location from which Harvester can retrieve EML documents via HTTP; Harvester can retrieve from any number of different Harvest Sites





## Harvester Definitions (cont.)

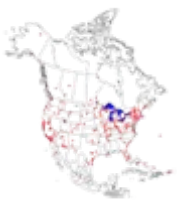
- Harvest List

An XML document, composed at a Harvest Site, that lists a set of EML documents to be harvested from that site

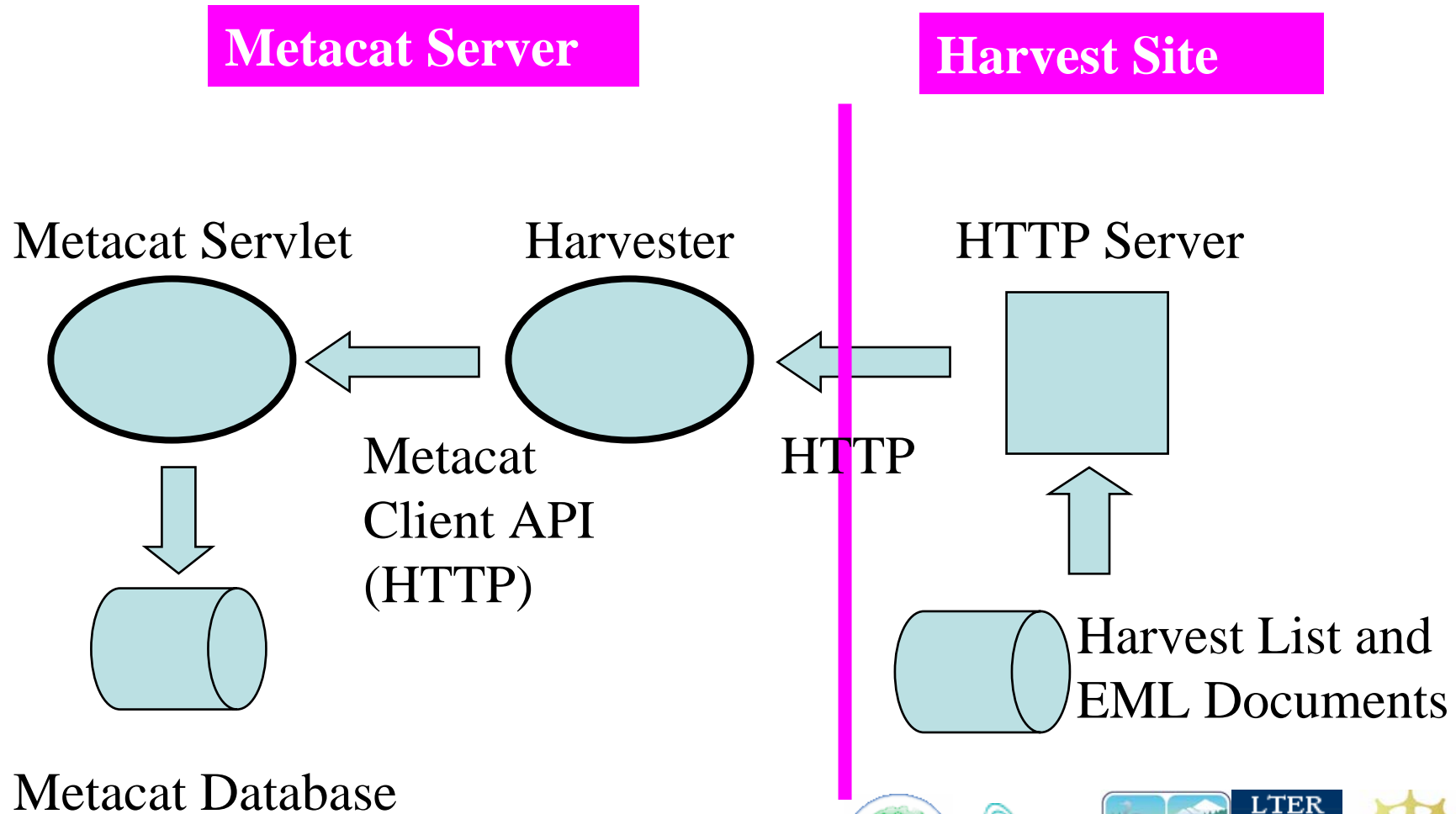
- Site Contact

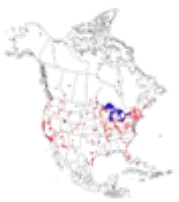
The individual at a Harvest Site who prepares the site's EML documents for retrieval, composes a Harvest List, and registers the site with Harvester





# Harvester Architectural Overview

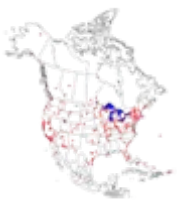




## Harvester Administration

- Configuring Harvester
- Running Harvester
- Reviewing E-mail Reports from Harvester

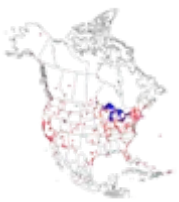




# Configuring Harvester: Settable Properties (in metacat.properties)

Property and Sample Value	Explanation
<u><a href="#">harvesterAdministrator= myaddress@unm.edu</a></u>	Send email to this address after every harvest
<u><a href="#">smtpServer= somehost.institution.edu</a></u>	Use this host machine to send email
delay=1	Wait 1 hour before starting the first harvest
period=24	Run a new harvest once every 24 hours
maxHarvests=90	Stop execution after completing 90 harvests

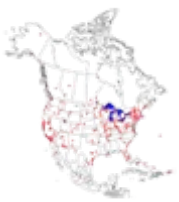




## Two Ways to Run Harvester

- In a terminal window
- As a background process (servlet)

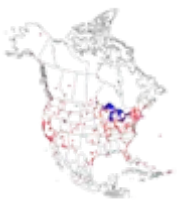




## Running Harvester in a Terminal Window

- Windows  
runHarvester.bat
- Linux/Unix  
sh runHarvester.sh
- Requires the Harvester Administrator to keep a terminal window open continuously, maintaining a connection to the Metacat server machine



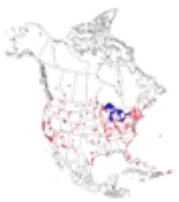


## Running Harvester as a Background Process

- Harvester can run as a Tomcat servlet, just like Metacat itself does
- No need to maintain a connection to the Metacat server machine in a terminal window



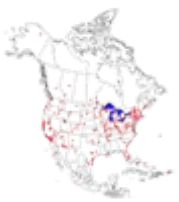




## Reviewing E-mail Reports from Harvester

- After every harvest, Harvester generates and sends an email report to the Harvester Administrator, summarizing the harvest results at each Harvest Site
- Harvester Administrator should review any reported errors, and work with the Site Contact to resolve them

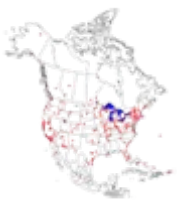




## Managing a Harvest Site

- Composing a Harvest List
- Registering with Harvester
- Reviewing Harvester reports to the Site Contact





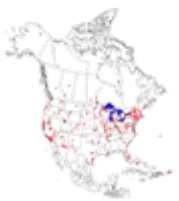
# Composing a Harvest List

Three items are specified for each document in the harvest list:

- docid (e.g. knb-lter-lno.8.1)
  - Scope knb-lter-lno
  - Identifier 8
  - Revision 1
- documentType  
eml://ecoinformatics.org/eml-2.0.1
- documentURL
  - Any URL that points to the EML document

<http://www.lternet.edu/~dcosta/remoteSensing/archive-lter-and-tm-19880723.xml>





## Composing a Harvest List (cont.)

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
<hrv:harvestList xmlns:hrv="eml://ecoinformatics.org/harvestList" >
```

```
<document>
```

```
<docid>
```

```
<scope>knb-lter-lno</scope>
```

```
<identifier>8</identifier>
```

```
<revision>1</revision>
```

```
</docid>
```

```
<documentType>eml://ecoinformatics.org/eml-2.0.1</documentType>
```

```
<documentURL>
```

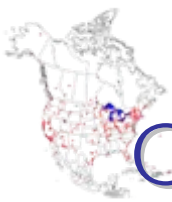
```
http://www.lternet.edu/~dcosta/remoteSensing/archive-lter-and-tm-19880723.xml
```

```
</documentURL>
```

```
</document>
```

```
</hrv:harvestList>
```





# Composing a Harvest List (cont.)

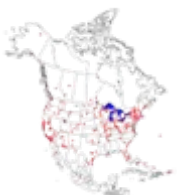
Harvest List Editor: harvestList.xml

Row #	Scope	Identifier	Revision	Document Type	Document URL
1	demoDocument	1	8	eml://ecoinformatics.org/eml-2.0.0	http://www.lternet.edu/~dcosta/document1.xml
2	demoDocument	2	6	eml://ecoinformatics.org/eml-2.0.0	http://www.lternet.edu/~dcosta/document2.xml
3	demoDocument	2	7	eml://ecoinformatics.org/eml-2.0.0	http://www.lternet.edu/~dcosta/document2.xml
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					

Cut Copy Paste Paste Defaults

- Harvest List Editor is a tool for composing and editing a Harvest List without looking at the underlying XML
- Harvest List Editor is included in the Metacat distribution, but is also available as a separate, downloadable client tool





# Harvester Registration Login

Metacat Harvester Registration Login - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://knb.lternet.edu:8088/knb/style/skins/knb/harvesterRegistrationLogin.html

Albuquerque weather... Books24x7.com Java 2 Platform SE v... KNB :: The Knowled... LTER Intranet - Serv... Metacat Data Catalog RT Login

Found 17 tickets Metacat Harvester Registration Login

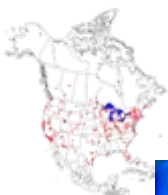
**Metacat Harvester Registration Login**

**Please Enter Username, Organization, and Password**

<b>Username</b>	<input type="text" value="LNO"/>
<b>Organization</b>	<input type="radio"/> NCEAS <input checked="" type="radio"/> LTER <input type="radio"/> NRS <input type="radio"/> PISCO <input type="radio"/> OBFS <input type="radio"/> Unaffiliated
<b>Password</b>	<input type="password" value="*****"/>

Done





# Harvester Registration

**Metacat Harvester Registration - Mozilla Firefox**

File Edit View Go Bookmarks Tools Help

http://knb.lternet.edu:8088/knb/harvesterRegistration

Albuquerque weather... Books24x7.com Java 2 Platform SE v... KNB :: The Knowled... LTER Intranet - Serv...

Found 17 tickets **Metacat Harvester Registration**

## Metacat Harvester Registration

Fill out the form below to schedule regular harvests of EML documents from your site.  
To register or changes values, enter all values below and click **Register**. To unregister, simply click **Unregister**.

Email address:

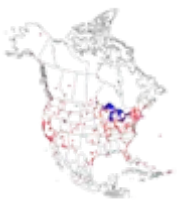
Harvest List URL:

Harvest Frequency  
Once every (1-99):

☒ day(s) ☐ week(s) ☐ month(s)

Done



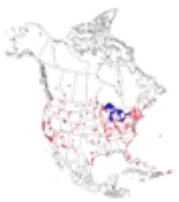


## Reviewing Harvester Reports to the Site Contact

- After each harvest at a site, Harvester generates and sends an email report to the Site Contact (as specified at Harvester Registration)
- Site Contact should attempt to resolve reported errors





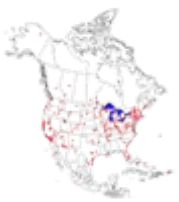


## Reviewing Harvester Reports to the Site

### Contact: Common Sources of Error

- documentURL in the Harvest List does not match location of the file on disk
- URL to the Harvest List that was entered during registration is incorrect
- Harvest List is not valid XML
- EML document that Harvester attempted to upload to Metacat is not valid EML





# Acknowledgements

This material is based upon work supported by:

The National Science Foundation under Grant Numbers 9980154, 9904777, 0131178, 9905838, 0129792, and 0225676.

The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.

The Andrew W. Mellon Foundation.

PBI Collaborators: NCEAS, University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research)

Kepler contributors: SEEK, Ptolemy II, SDM/SciDAC, GEON

