



# Enabling metadata discovery

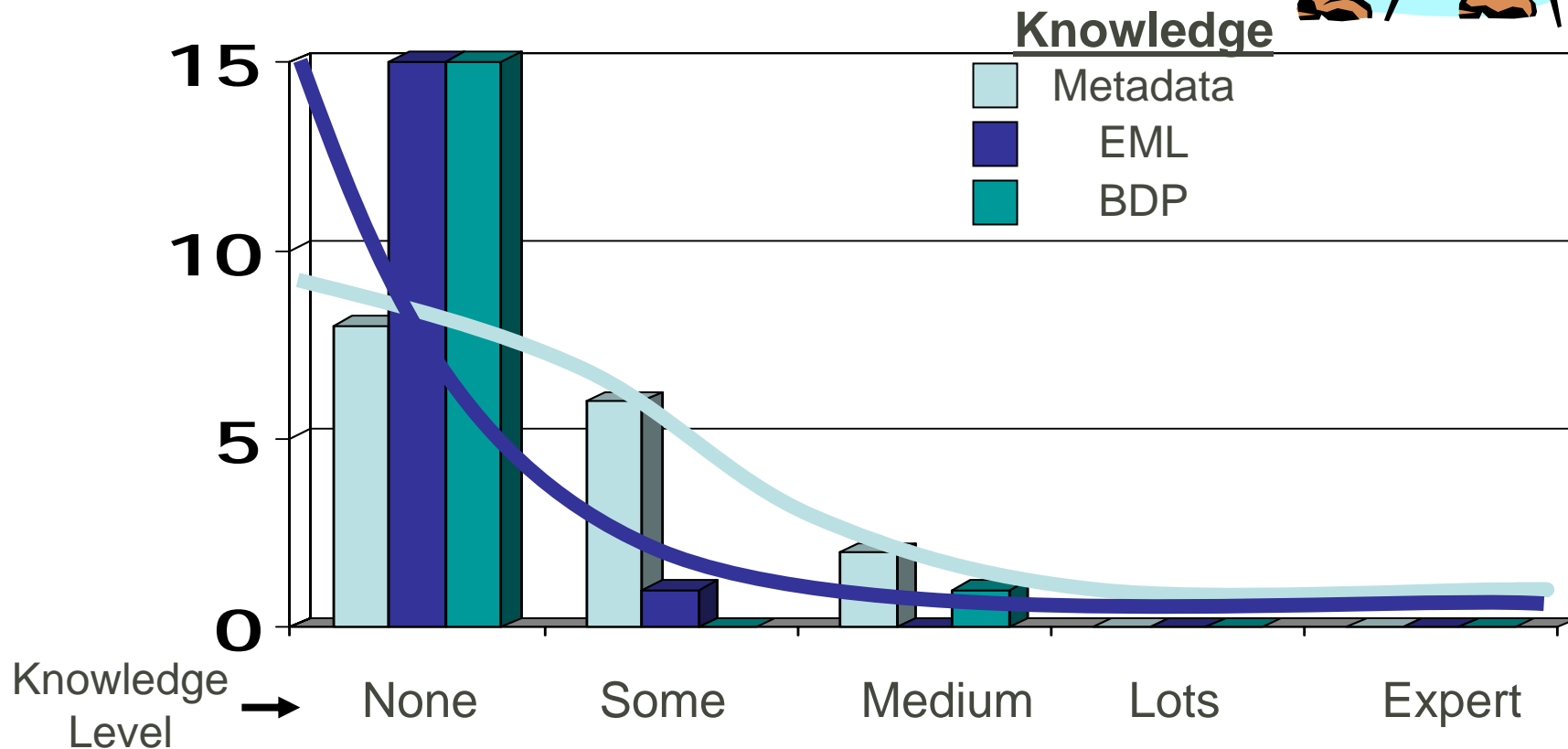
Practical metadata  
standardization tricks

Iñigo San Gil  
LTER Network office  
[isangil@lternet.edu](mailto:isangil@lternet.edu)





# How about YOU?





# Talk Outline

Metadata  
standardization  
(EML)

EML Conversion to & from  
the BDP standard

Dissemination  
of metadata





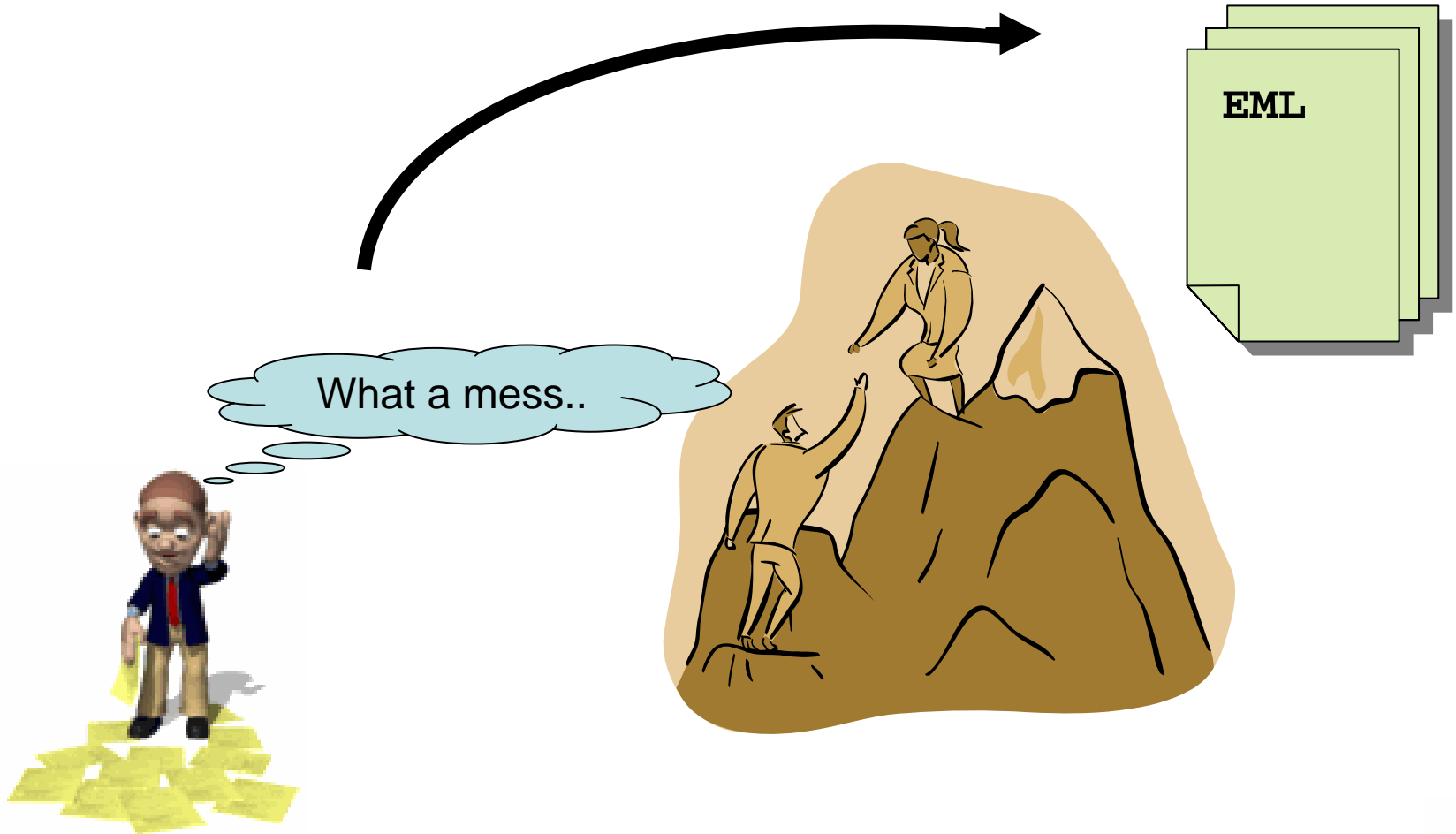
# Why enable metadata discovery?

- Expand research
- Forge alliances
- Easier management
- Common standard
- Interoperability





# Goal: standardize metadata





# Metadata standardization topics:

- How much, how rich and how is data stored
  - Just a few? Consider Morpho or an XML editor
  - A bunch? Databases: MS Access, SQLs, Paradox...
  - Flat text files: Structured Folders (or not)
- The LTER case. 26 sites, 26 forms of metadata
- Case study: transforming legacy metadata





# Standardization. Questions about your metadata

How wide and deep is your metadata ? do we have it in a database? do we have flat text files?

Morpho



**Morpho**

Data Management for Ecologists

Next talk –Will Tyburczy

XML editor



**Perl** (*Practical Extraction and Report Language*)








Other tools ... *sorry, not time for the showcase here!*





# Metadata Tool Matrix

	Open Source Proprietary	On-line Stand Alone	Platform
XML Spy	Home ed - \$0 Professional ~ \$1000 Enterprise ~ \$3000	Stand Alone	
Morpho	Free	Stand Alone On-line*	Any
oXygen	Academic \$50 Professional \$200	Stand Alone	Any
Notepad, Word		Stand Alone	
Emacs, vi, ...	Free	Stand Alone	Unix 
Databases	SQLs { \$0 - \$\$\$} Oracle {\$50* - \$1000+} MSAccess 	Stand Alone On-line	Mixed
NBII tools	Free	Stand Alone On-line	Mixed







# XML Editors -- XMLSpy

File Edit Project XML DTD/Schema Schema design XSL/XQuery Authentic Convert View Browser WSDL SOAP Tools Window Help

Project: Examples  
- Orig-Chart  
- Expense Report  
- International  
- Purchase Order  
- SOAP Debugger  
- WSDL Editor  
- IndustryStandards  
- XML-based Website  
- Tamino  
- XQuery  
- XSLT2

eml

attributes

dataset

ds:DataSetType

attributes

alternatIdentifier 0..∞

shortName

title 1..∞

creator 1..∞

metadataProvider 0..∞

associatedParty

pubDate 0..∞

language

series

abstract

keywordSet 0..∞

additionalInfo 0..∞

intellectualRights

distribution 0..∞

coverage

purpose

maintenance

contact 1..∞

publisher

pubPlace

methods

project

access

dataTable

spatialRaster

spatialVector

Components

Element

- access
- attribute
- attributeList
- citation
- dataTable
- doc:description
- doc:example
- doc:lineage
- doc:module
- doc:moduleDocs
- doc:summary
- doc:tooltip
- ds:dataset
- eml
- methods
- otherEntity
- party
- physical
- projectionList
- prot:protocol
- researchProject
- spatialRaster
- spatialReference
- spatialVector

by Type by Namespace

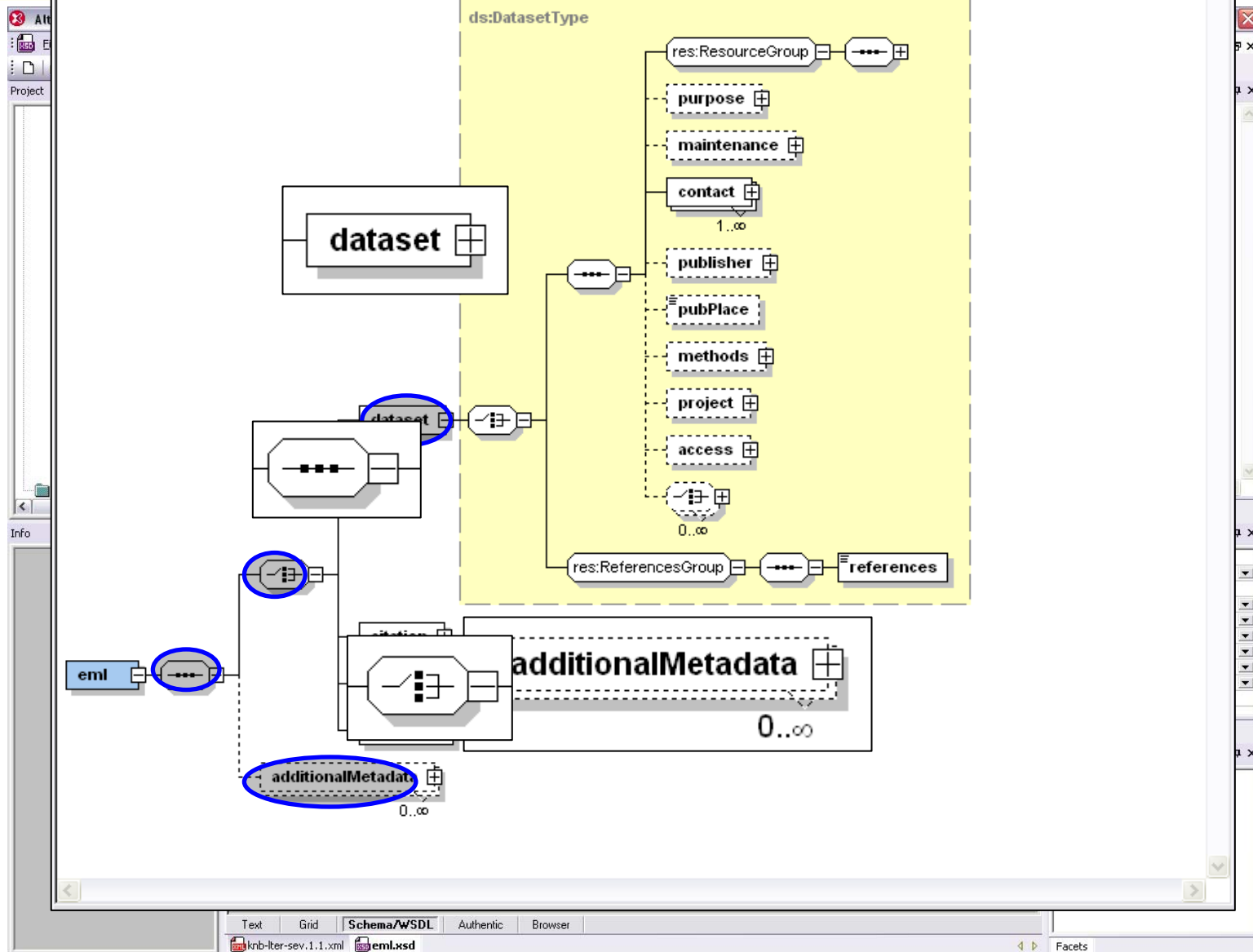
Details

name	creator
isRef	<input type="checkbox"/>
minOcc	1
maxOcc	unbounded
type	rp:ResponsibleParty
content	complex
derivedBy	
mixed	
nullable	
block	
form	
id	

Text Grid Schema/WSDL Authentic Browser

eml

Details





# XMLSpy -- Document View

Altova XMLSpy - [knb-lter-sev.1.1.xml]

Project: myEML \*  
XML Files: knb-lter-sev.1.1.xml  
XSL Files: XQuery Files  
HTML Files: DTD/Schemas: C:\eml-2.0.1  
Entities

Info: Element Model dataset choice

Text Grid Schema/WSDL Authentic Browser

knb-lter-sev.1.1.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XMLSpy v2005 sp1 U (http://www.xmlspy.com) by Mark Servilla (UNM Dept of Biology, LTER Network Office) -->
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.0.1" xmlns:ds="eml://ecoinformatics.org/dataset-2.0.1" xmlns:doc="eml://ecoinformatics.org/documentation-2.0.1" xmlns:cit="eml://ecoinformatics.org/literature-2.0.1" xmlns:prot="eml://ecoinformatics.org/protocol-2.0.1" xmlns:res="eml://ecoinformatics.org/resource-2.0.1" xmlns:sw="eml://ecoinformatics.org/software-2.0.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 C:\eml-2.0.1\eml.xsd" packageid="K...">
  <dataset>
    <title>Sevilleta LTER NPP Quad...
    <creator>
      <individualName>
        <givenName>Este</givenName>
        <surName>Muldaivn</surName>
      </individualName>
      <organizationName>New Me...
      <organizationName>Departm...
      <organizationName>Univers...
      <address>
        <deliveryPoint>1 Universi...
        <city>Albuquerque</city>
        <administrativeArea>NM</administrativeArea>
        <postalCode>87131</postalCode>
        <country>USA</country>
      </address>
      <electronicMailAddress>mul...
    </creator>
    <abstract>
      <section>
        <para>Net primary produc...
        Estimates of NPP are im...
        community to a wide ra...
        primary production acro...
        and pinyon-juniper woo...
        study focuses on estim...
        losses to death and de...
        is sampled three times...
        are obtained from perm...
      </para>
    </section>
    </abstract>
    <keywordSet>
      <keyword>ANPP</keyword>
      <keyword>biomass</keyword>
      <keyword>Sevilleta</keyword>
      <keyword>LTER</keyword>
    </keywordSet>
    <contact>
      <individualName>
        <surName>/>
      </individualName>
    </contact>
  </dataset>
</eml:eml>
```

Elements

- alternatIdentifier
- shortName
- title
- references
- creator
- metadataProvider
- associatedParty
- pubDate
- language
- series
- abstract
- keywordSet
- additionalInfo
- intellectualRights
- distribution
- coverage
- purpose
- maintenance
- contact
- publisher
- pubPlace
- methods
- project
- access
- dataTable
- spatialRaster
- spatialVector
- storedProcedure
- view
- otherEntity

of carbon consumption and fixation. as spatial and temporal responses of the hed at the Sevilleta LTER to monitor net grama dominant grassland, juniper-savannah are important in estimating NPP, the described the change in plant mass, including any above ground biomass production (ANPP) ites. In addition, volumetric measurements ss and volume.

NSF

CAP NUM SCRL



# XML Editors – oXygen

The screenshot displays the oXygen XML Editor interface with the Logical Model View of an XML Schema. The main workspace shows a hierarchical tree structure of the schema elements. The 'eml' root element contains an 'appinfo' element, which in turn contains 'doc:module', 'doc:tooltip', 'doc:summary', and 'doc:description'. The 'dataset' element is a complex type that contains a 'DatasetType' element, which is a base type for 'Dataset'. The 'DatasetType' element contains a 'res:ResourceGroup' element, which is a sequence of elements: 'purpose', 'maintenance', 'contact', 'publisher', 'pubPlace', and 'methods'. The 'res:ResourceGroup' element is a sequence of elements, and its cardinality is 1..∞. The 'purpose' element has a cardinality of 0..1, 'maintenance' has a cardinality of 0..1, 'contact' has a cardinality of 1..∞, 'publisher' has a cardinality of 0..1, 'pubPlace' has a cardinality of 0..1, and 'methods' has a cardinality of 0..1.

On the left, the 'Outline' pane shows the project structure with 'newProject.xpr' and 'eml.xsd'. Below it, the 'Outline' pane lists the schema components, including the 'eml' root element and its children.

On the right, the 'Attributes' pane shows the attributes of the selected element, and the 'Model' pane shows the namespace information.

At the bottom, the 'Full Model View' pane displays the XML Schema definition (XSD) code for the 'dataset' element. The code is as follows:

```
<?xml version="1.0"?>
<dataset>
  <doc:tooltip>Ecological Metadata
</doc:tooltip>
  <doc:summary>A collection of EML metadata
</doc:summary>
  <doc:description>The "eml" element allows
  </doc:description>
  <xs:appinfo>
    <xs:annotation>
      <xs:complexType>
        <xs:sequence>
          <xs:choice>
            <xs:element name="dataset" type="ds:DatasetType" />
            <xs:annotation>
              <xs:appinfo>
                <doc:tooltip>Dataset Resource</doc:tooltip>
                <doc:summary>A resource that describes a data set, which can
                include one or more data entities such as data tables.
              </doc:summary>
            </xs:appinfo>
          </xs:choice>
        </xs:sequence>
      </xs:complexType>
    </xs:annotation>
  </xs:appinfo>
</xs:appinfo>
</dataset>
```



# Data management models

Databases:

MySQL, MSAccess, Paradox, Metacat, etc



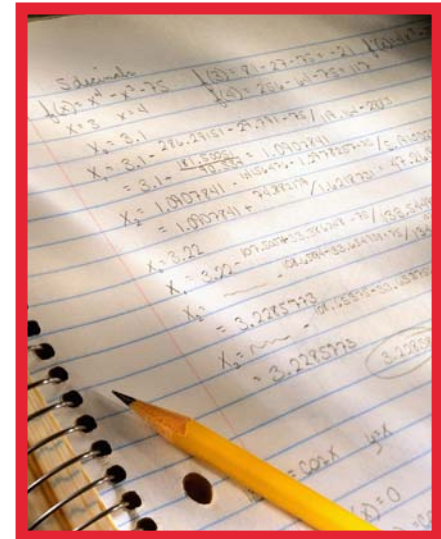
Spreadsheets:

Excel, etc



Text files:

Flat ASCII, Word documents.

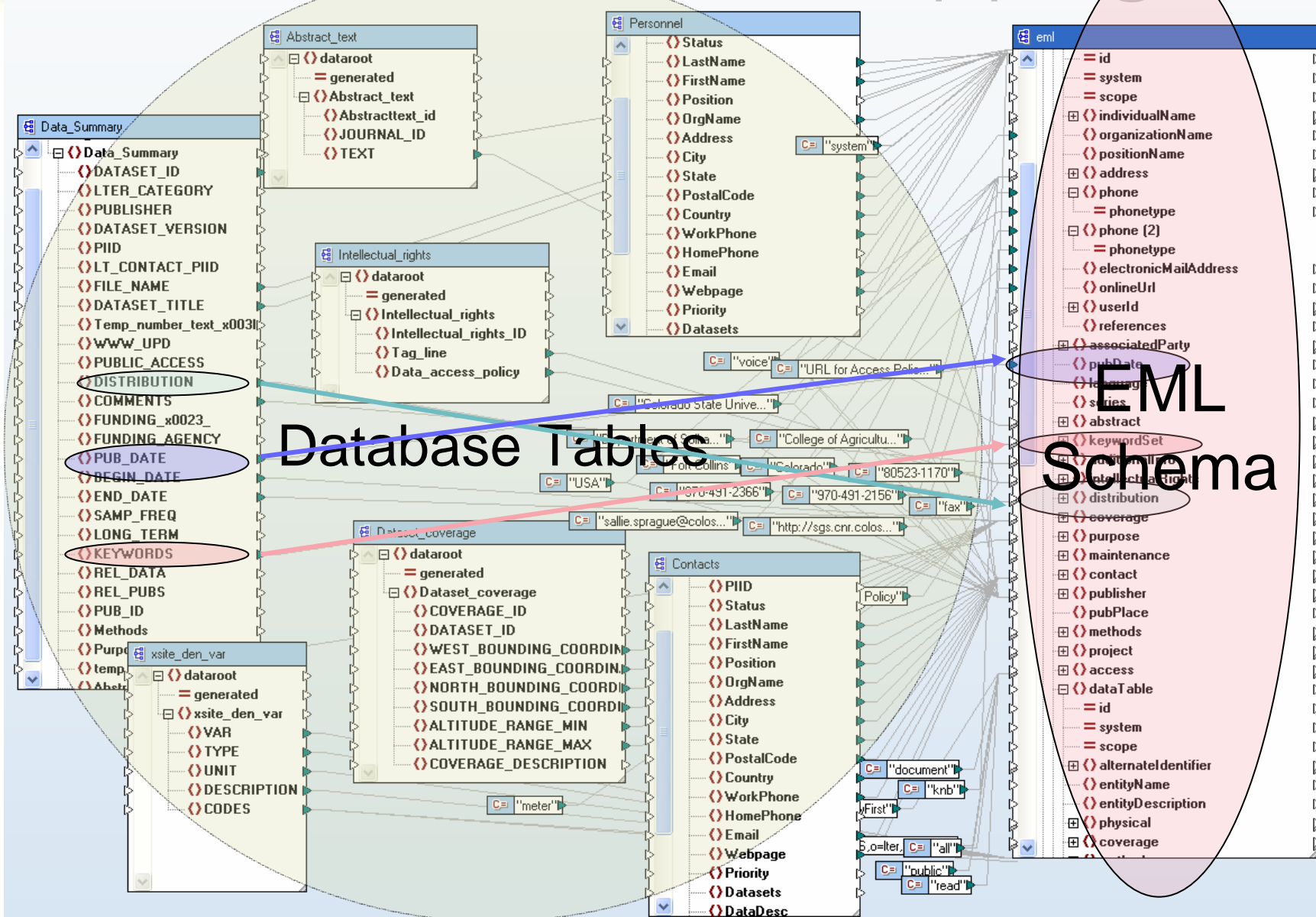


Combos of the above





# Database to EML mapping





# Text to EML ma

Useful when text sources are

## ARCTIC LTER DATABASE

(1) FILE NAME: 2004e5chlr.txt

(2) YEAR: 2004

(3) PI: Dr. Anne Giblin

(4) OTHERS: Chris Crockett, Lori Winters, Sam Kelse

(5) BRIEF DESCRIPTION OF DATA FILE: Corrected Chlorophyll a for lake E5 during the summer of 2004

(6) KEYWORDS: Chlorophyll a

(7) SITE TYPE: AQUATICS-LAKES

(8) RESEARCH LOCATION: Toolik Field Station, North Slope, Alaska.

(9) EXPERIMENTAL DESIGN AND METHODS:  
Water samples are collected at a minimum of the epi, meta, and hypolimnion, samples at other depths are collected as needed dependent on lake

```
<?xml version="1.0" encoding="UTF-8"?>  
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2  
C:\eml-2.0.1\eml.xsd" packageId="" system="">
```

```
<dataset>
```

```
<title> </title>
```

```
<creator>  
<individualName>
```

```
<title> </title>
```

```
<creator>  
<individualName>
```

```
<abstract>
```

```
<section>
```

```
<para>Corrected Chlorophyll a  
for lake E5 during the summer  
of 2004</para>
```

```
</section>
```

```
</abstract>
```

```
<surName/>  
</individualName>  
</contact>
```

```
</dataset>
```

```
</eml:eml>
```





# The LTER network

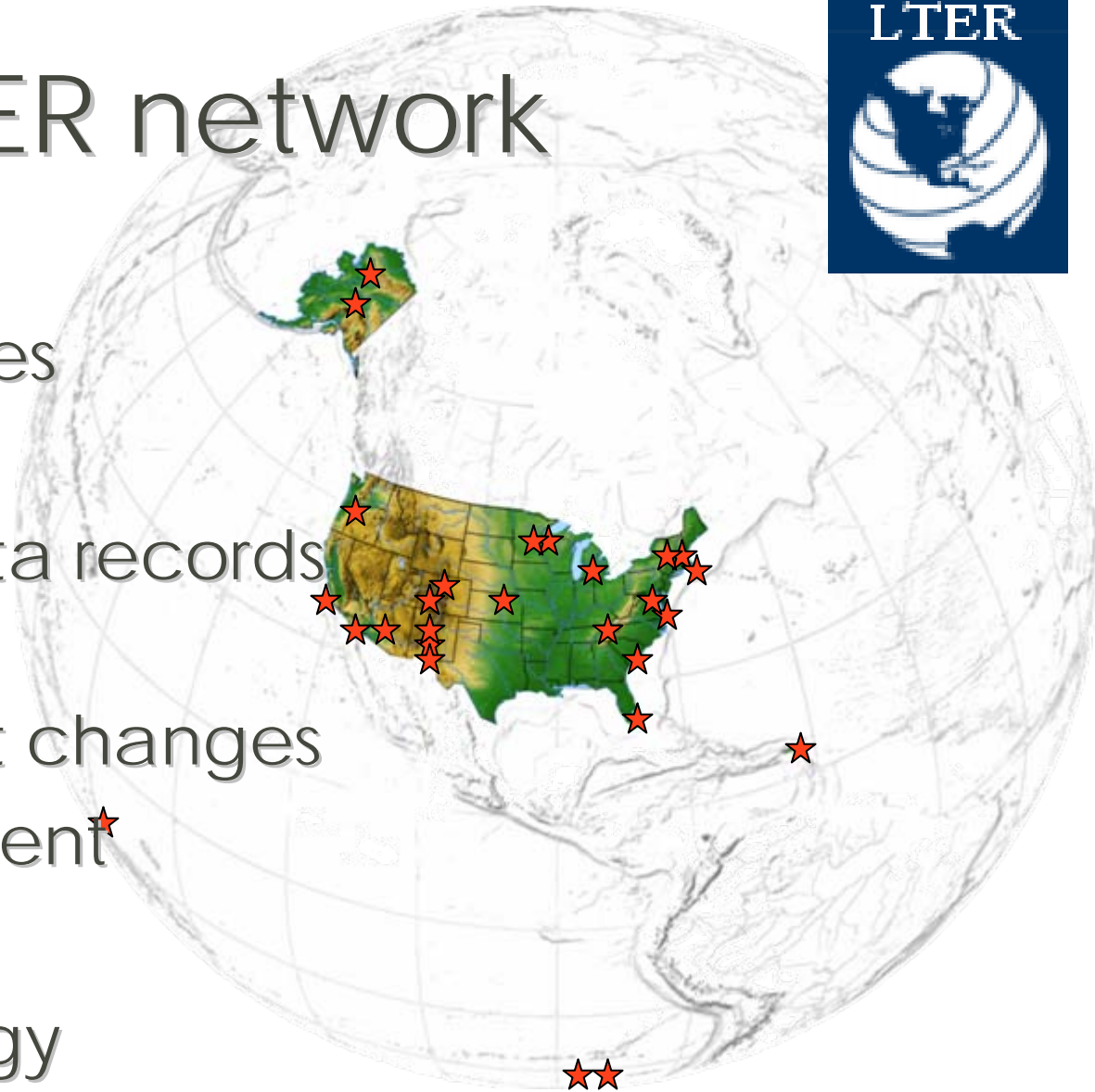


26 autonomous sites

20 yr+ worth of data records

Relatively frequent changes  
in data management

Evolving technology





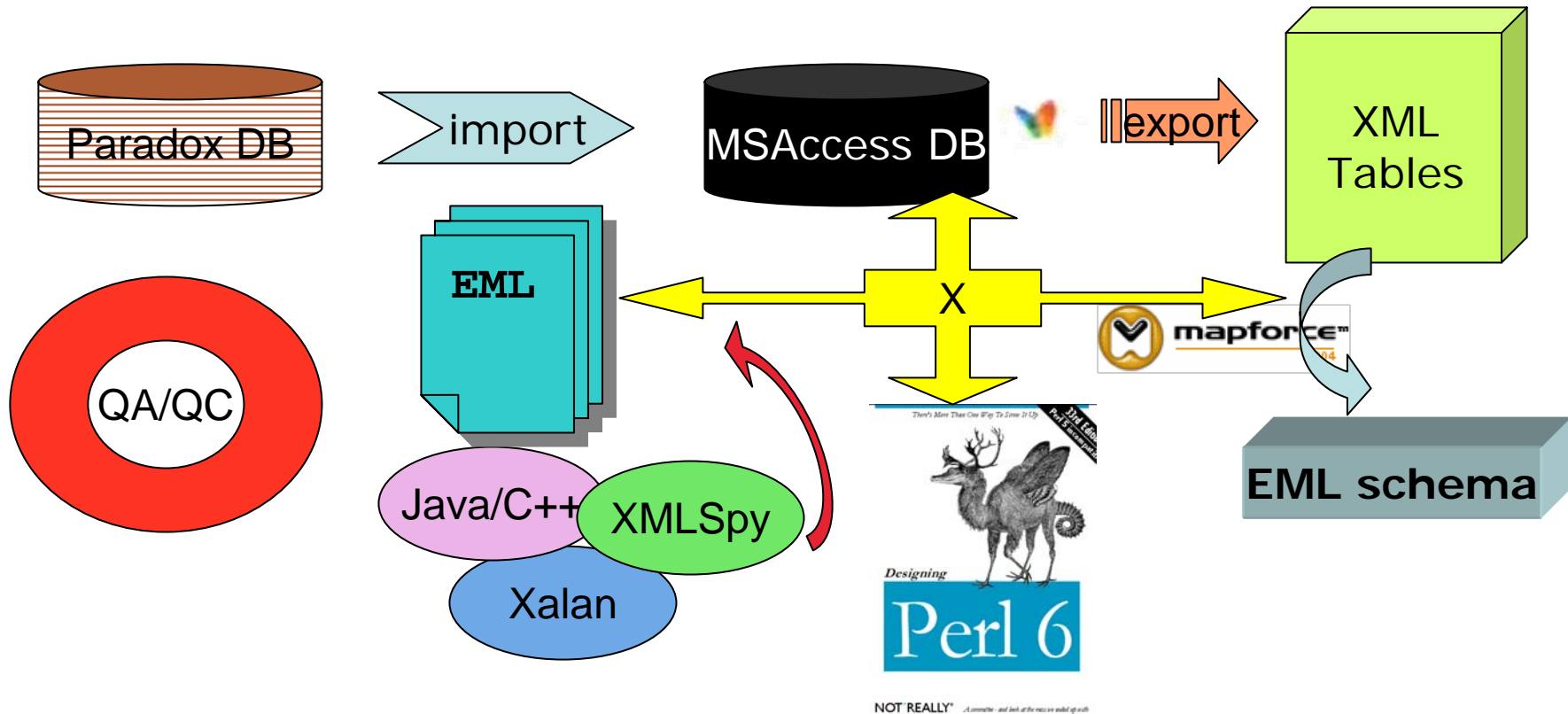


# Case study – Luquillo





# Case study – Luquillo





# Talk Outline

Metadata  
standardization  
(EML)



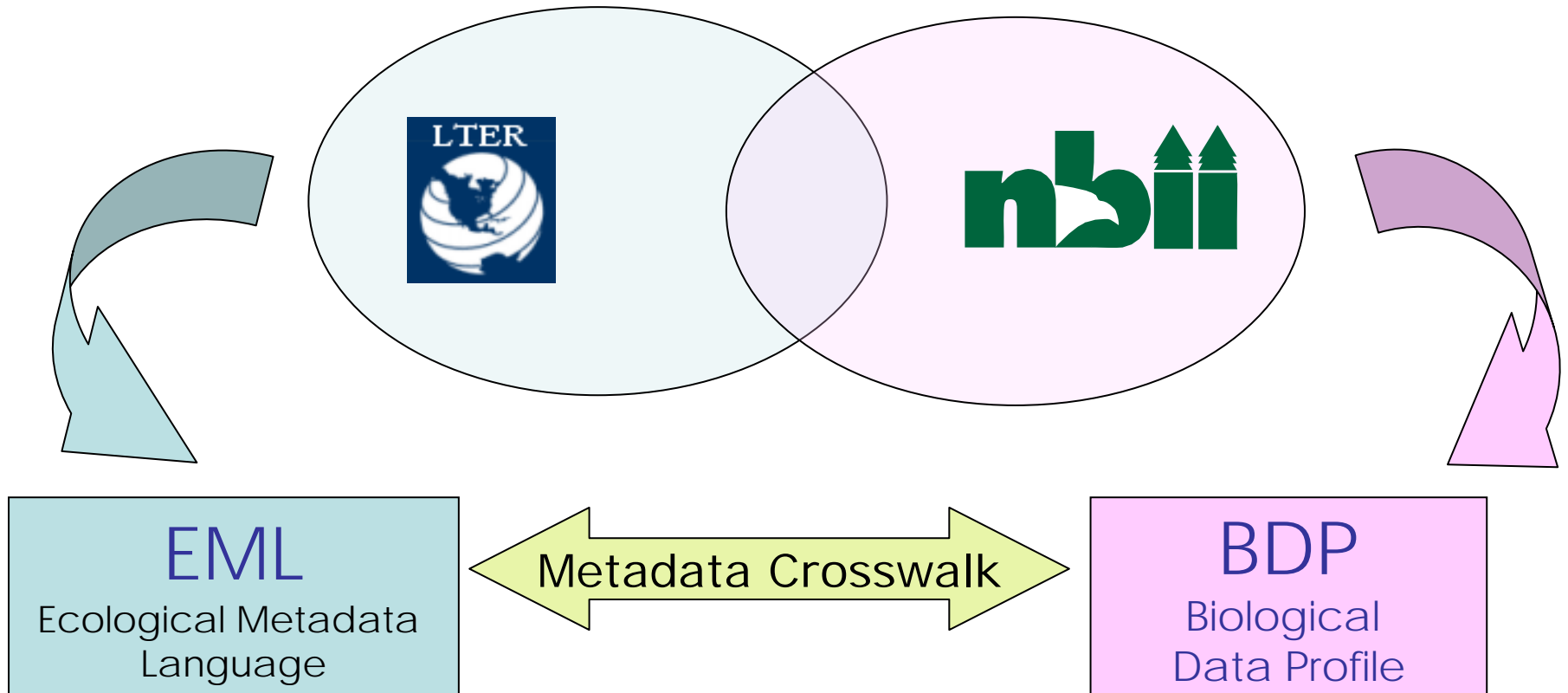
**EML Conversion to &  
from the BDP standard**

Dissemination  
of metadata





# Metadata Crosswalks





# Metadata Crosswalk

Yeah, well, but HOW??

Web application at

[http://fire.Iternet.edu/~isangil/Web\\_translation\\_application/EmlToBdp/eml2bdp.php](http://fire.Iternet.edu/~isangil/Web_translation_application/EmlToBdp/eml2bdp.php)

XSLT code available for free !

Web service to be released soon\*

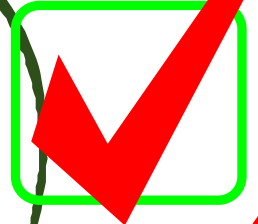
\*Contact [isangil@Iternet.edu](mailto:isangil@Iternet.edu)





# Talk Outline

Metadata  
standardization  
(EML)



EML Conversion to & from  
the BDP standard



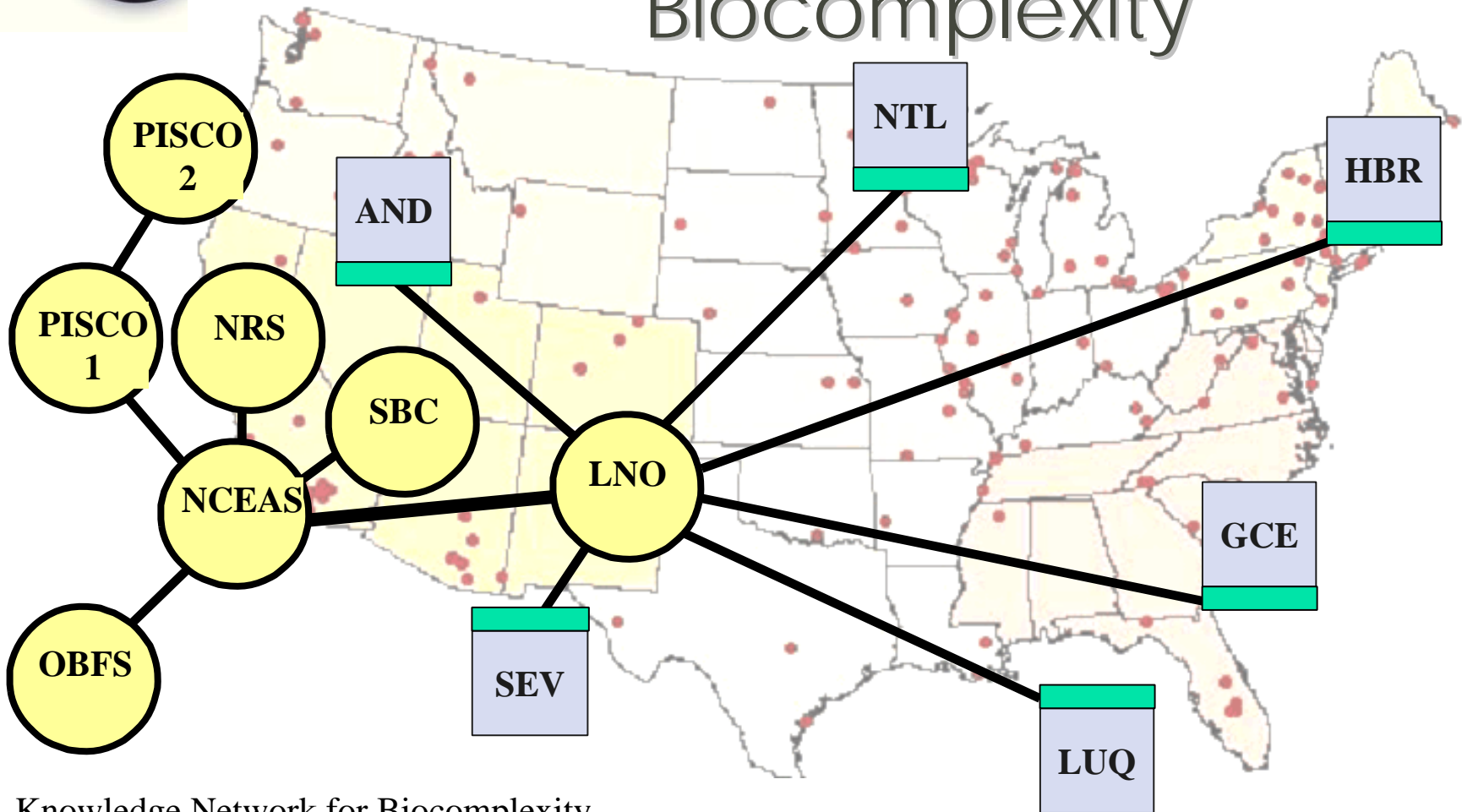
Dissemination  
of metadata



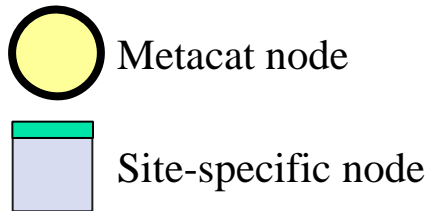




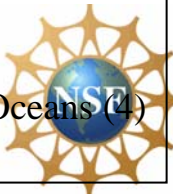
# Knowledge Network for Biocomplexity



Knowledge Network for Biocomplexity

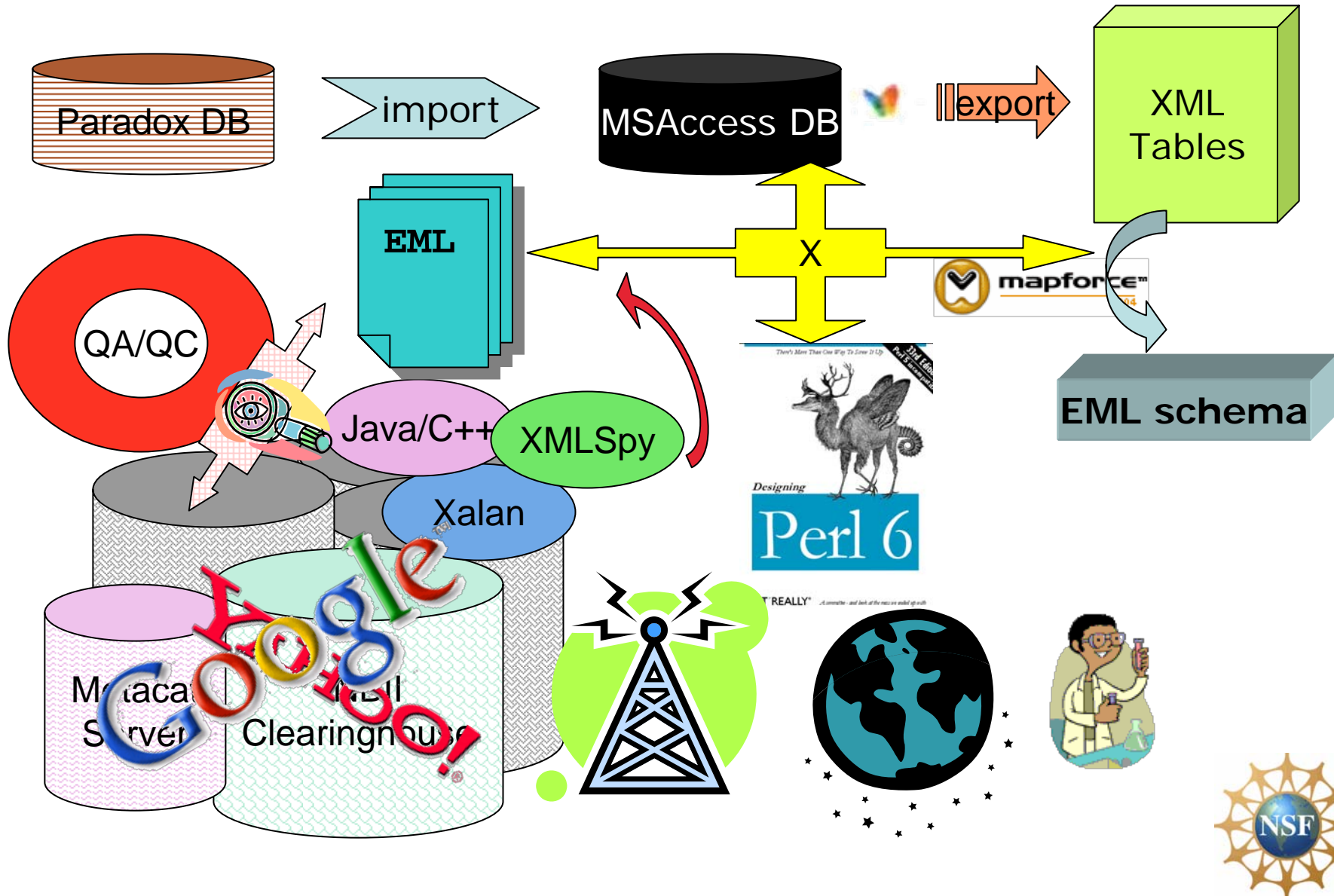


LTER Network (27)  
 Organization of Biological Field Stations (180+)  
 UC Natural Reserve System (36)  
 Partnership for Interdisciplinary Studies of Coastal Oceans (4)  
 Multi-agency Rocky Intertidal Network (60)





# Case study – Luquillo







# The US Long Term Ecological Research Network

## Data Set Description

Identifier: knb-lter-vcr.23.2  
Catalog System: VCR  
Alternate Identifier: VCR97012  
Title: **1992-93 Parramore Permanent Plot Baseline Data : Shrub Data**  
Publication Date: 1995-01-01

### Data Set Owner(s):

Individual: **David Richardson**  
Address: Department of Environmental Sciences, Clark Hall , University of Virginia,  
Charlottesville, VA 22903 USA  
Phone: (434) 924-3263 (voice)  
Phone: (434) 982-2137 (facsimile)  
Email Address: [dlr2n@virginia.edu](mailto:dlr2n@virginia.edu)  
Individual: **John Porter**  
Address: UVA, Department of Environmental Sciences, 291 McCormick Road, P.O. Box 400123,  
Charlottesville, VA 22903-4123 USA  
Phone: 434-924-8999 (voice)  
Phone: 434-982-2137 (facsimile)  
Email Address: [jhp7e@virginia.edu](mailto:jhp7e@virginia.edu)  
Individual: **Johann Knutson**



# 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, Nutrient 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

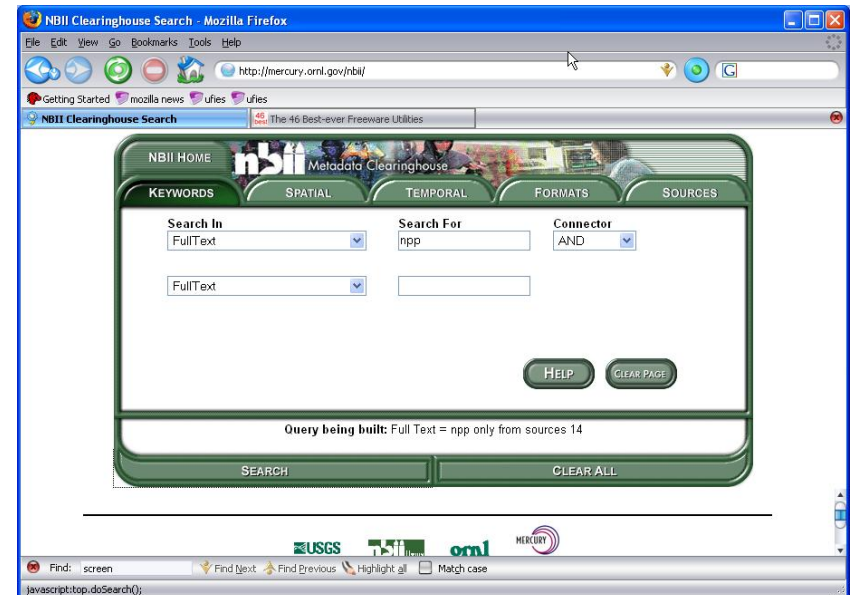
### 456 data packages found

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			



# Metadata Discovery: NBII

- URL: <http://mercury.ornl.gov/nbii>
- Searches can be as generic as you want, or really specific.
- 28 nodes, including the LTER node
- Metadata only





# Links, people contacts

[www.altova.com](http://www.altova.com) (XML Spy editor, MapForce)

[www.oxygenxml.com](http://www.oxygenxml.com) (XML editor )

[www.activestate.com](http://www.activestate.com) (Free Perl runtime compiler)

Email addresses:

[isangil@lternet.edu](mailto:isangil@lternet.edu) Inigo San Gil

[servilla@lternet.edu](mailto:servilla@lternet.edu) Mark Servilla

<http://Seek.ecoinformatics.org> : these presentations!

