



Ecological Metadata Language (EML) and Morpho

Matthew B. Jones
jones@nceas.ucsb.edu

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



<http://knb.ecoinformatics.org>

KNB and SEEK Products

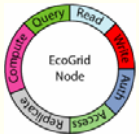
- Ecological Metadata Language (EML)



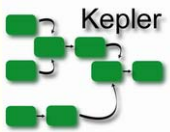
- Morpho
 - metadata and data management software



- Metacat
 - distributed metadata/data system
 - registries: KNB, UCNRS, OBFS, NCEAS, PISCO, LTSS



- EcoGrid
 - integrating distinct data systems and networks



- Kepler
 - grid-enabled scientific workflows

Ecological Metadata Language

- Metadata: a means to manage ecological data
 - Metadata is not an end, it's a means
- Many existing metadata standards
 - NBII Biological Data Profile (BDP), Dublin Core Element Set, ISO Geographic Metadata, GCMD Directory Interchange Format (DIF)
- EML initiative started in 1997
 - Common language for describing, archiving, and transporting data



- Ecological Metadata Language
 - What is it? Documentation about data, aka metadata
 - Which data? Ecologically relevant data
 - Biodiversity surveys, hydrology, atmospheric chemistry, spatial data, behavioral experiments, ...
 - But really, any scientific data

SWDB

Aug 29, 2004



Why do we need metadata?

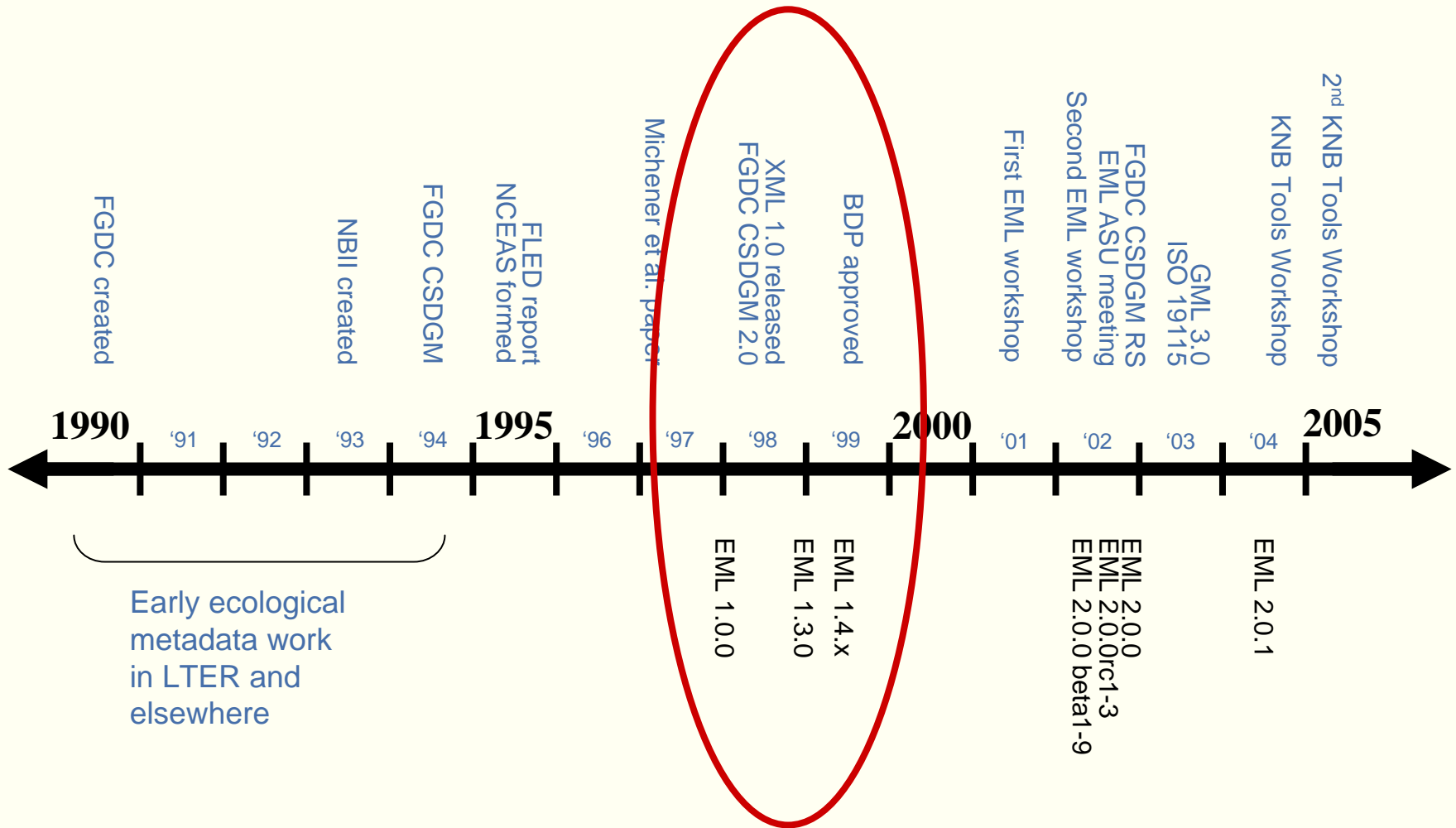
- Ecologically relevant data are:
 - Heterogeneous
 - Dispersed
 - From many disciplines
- Metadata: a means to manage ecological data
 - Accommodate heterogeneity and dispersion
 - There is no universal data model for ecology
- Data discovery
- Data interpretation
- Enable advanced analytical applications



Related metadata standards

- Dublin Core Element Set
 - Corresponds roughly to eml-resource
- Content Standard for Digital Geospatial Metadata (CSDGM)
 - Federal Geographic Data Committee (FGDC)
 - Corresponds to eml-spatialRaster, eml-spatialVector, eml-spatialReference
 - Overlaps in other modules (eml-resource)
- Biological Data Profile (BDP) of the CSDGM
 - Biological Data Working Group of the FGDC
 - Shares structure for taxonomicCoverage, geologicAge, and ascii table structures
- ISO 19115 Geographic information: Metadata
 - Incorporated in the eml-spatial* modules
 - Eml-party derived from ISO 19115
- Darwin Core
 - Partially overlaps with eml-coverage
- Geography Markup Language (GML)

Abridged History of EML

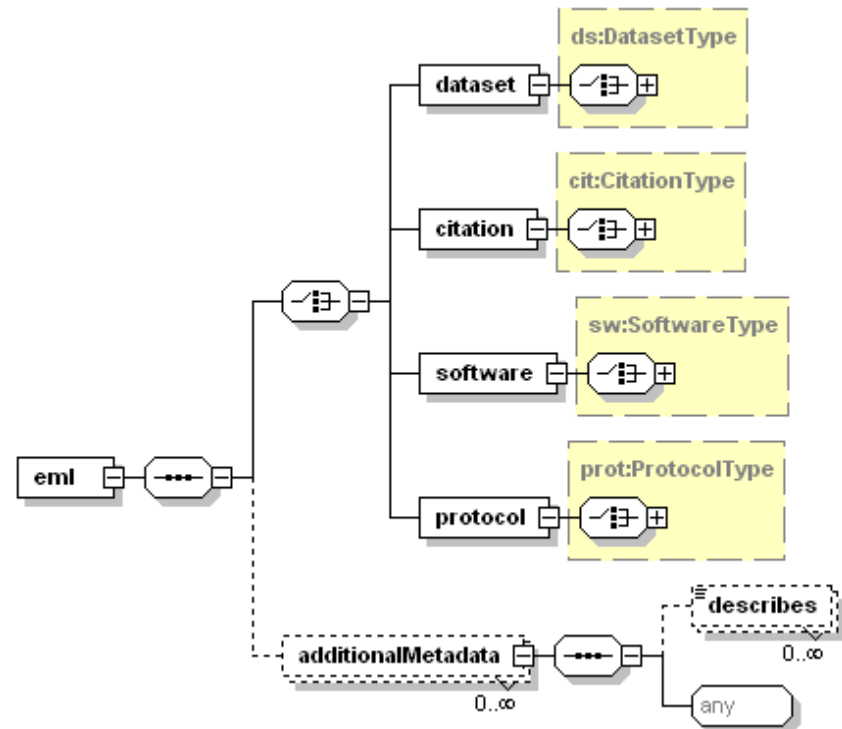


Conversion among standards

- Extensive overlap among all of these standards
- EML represents a superset of:
 - BDP
 - CSDGM
 - ISO 19115 (I think)
 - Dublin Core
- Can convert from EML to those
- Practically, have conversion script for:
 - EML → BDP
 - Uses XSLT so can be used in a variety of software

What does EML document?

- Scientific data sets
 - Tabular and relational data
 - Spatial images and GIS data
- Processing Software
- Literature citations
- Scientific Protocols
- Easily extensible
 - AdditionalMetadata



SWDB

Aug 29, 2004

Information covered by EML

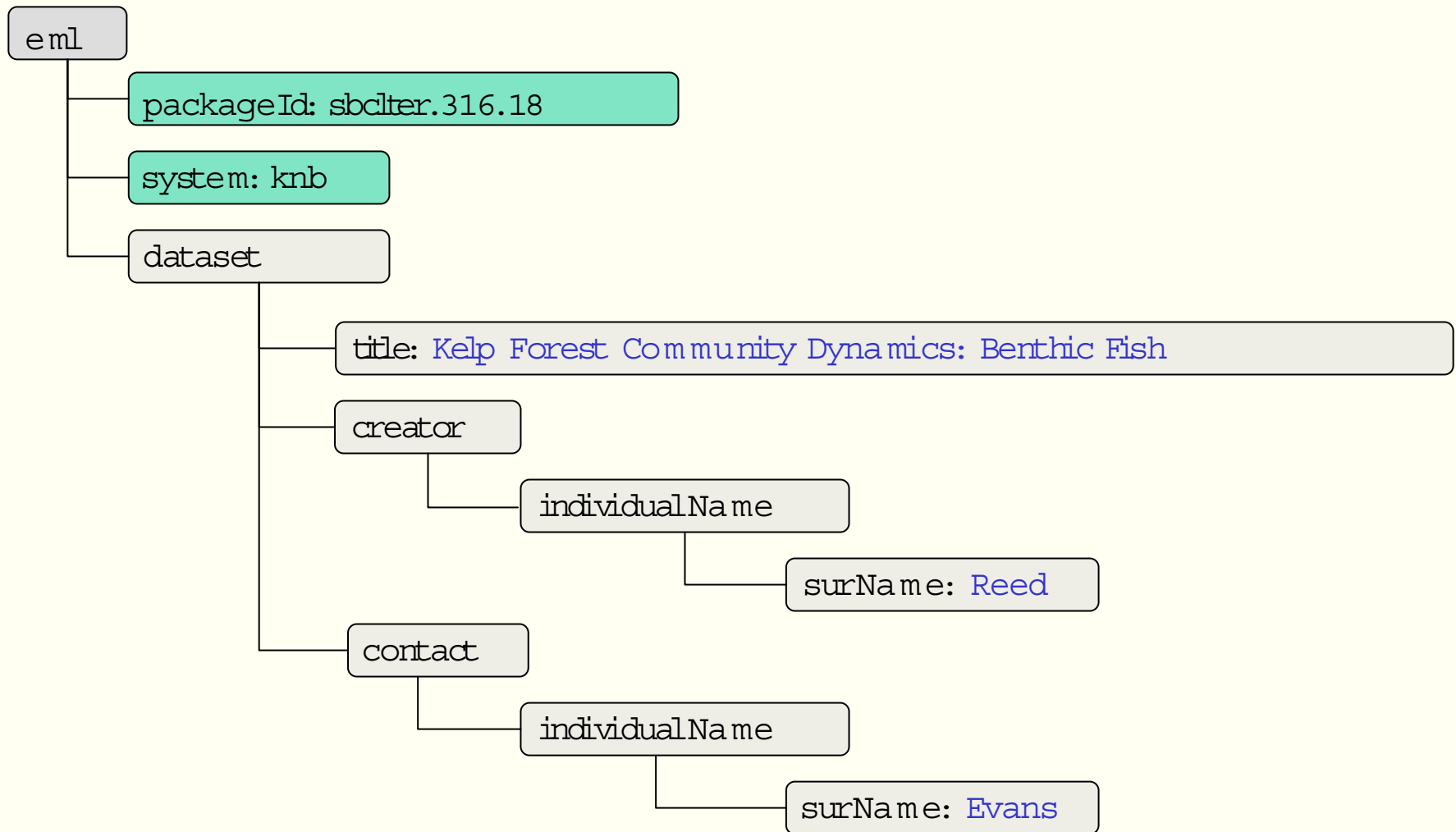
- Discovery information
 - Creator, Title, Abstract, Keyword, etc.
- Ownership and Citation information
- Intellectual property information
- Coverage
 - Geographic, temporal, and taxonomic extent
- Protocols and methods

Used in
registries
and catalogs

- Logical and physical data structure
 - Data semantics via unit definitions and typing

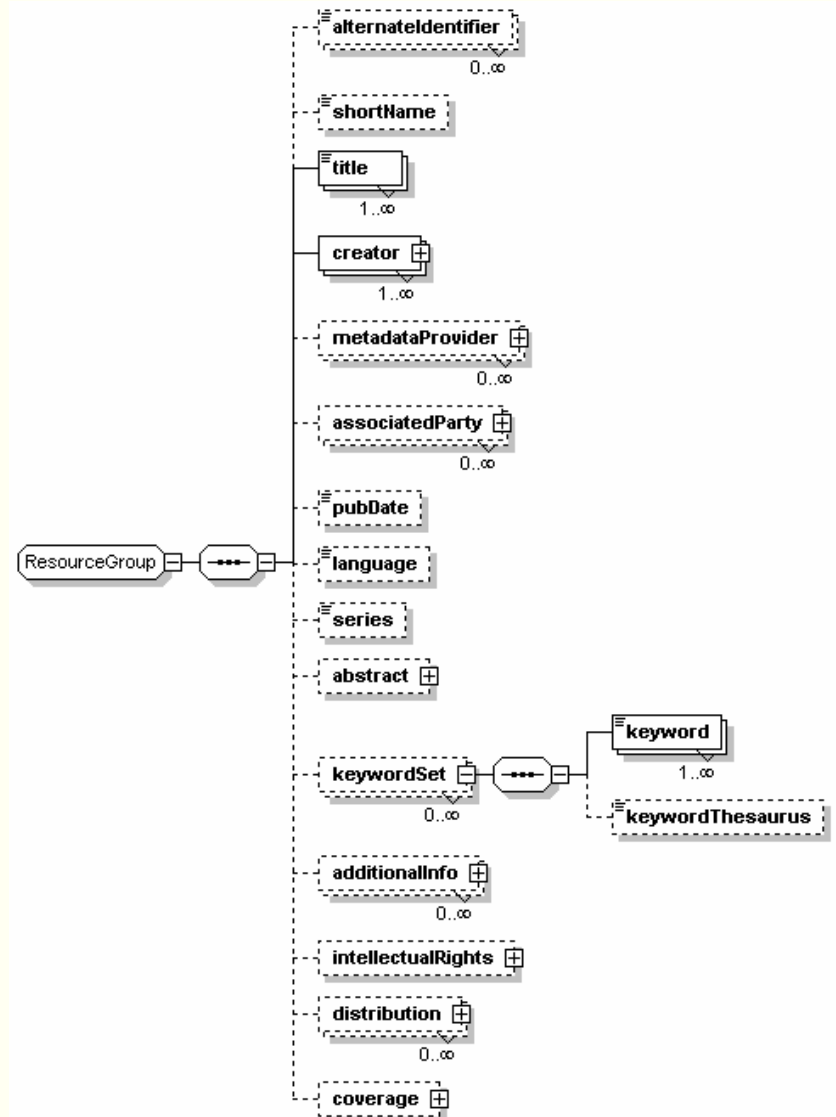
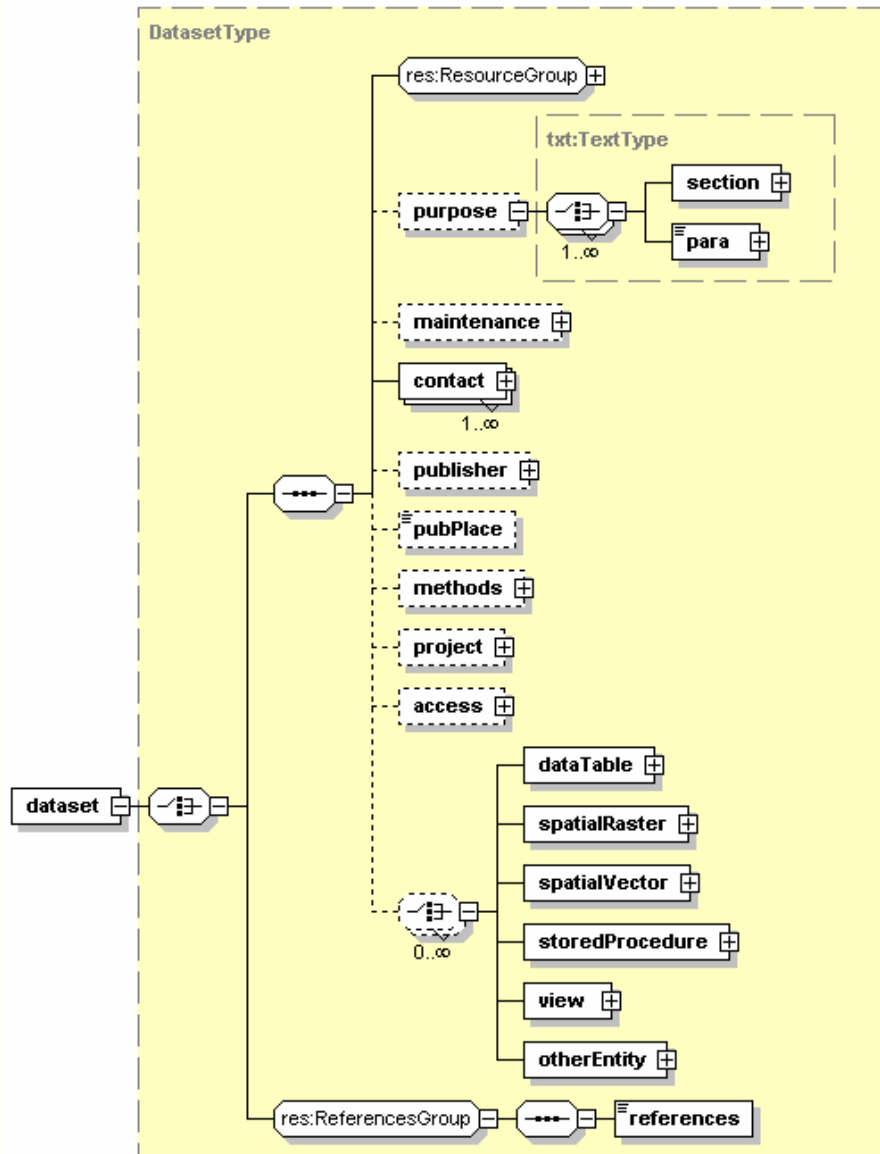
Used in
Analytical
applications

The simplest EML example



*barely enough information for 'identifying' the dataset

Identification: resource elements



EML Strengths

- EML is:
 - Designed for interoperability and portability
 - Supports validation and machine processing
- Modular
 - 24 modules for flexibility and reusability
- Extensible
 - Simple to add metadata from other standards or site-specific metadata
- Designed to support scientific analysis
 - Metadata details needed for analysis and models

SWDB

Aug 29, 2004

EML uses open standards

- Uses Extensible Markup Language (XML)
 - Documents are cross-platform
 - EML documents support international languages
 - Excellent exchange language
- EML documents are human readable text
 - Therefore, archive friendly
- EML documents are fine-grained
 - Simpler conversion to other standards
 - More difficult to convert from coarse-grained standards like Dublin Core

An Example EML Document

```
<?xml version="1.0"?>
<eml:eml packageId="piscoUCSB.5.20" system="knb"
xmlns:eml="eml://ecoinformatics.org/eml-2.0.0">
<dataset>
<shortName>Alegria Temperatures</shortName>
<title>PISCO: Intertidal Temperature Data:
Alegria, California: 1996-1997</title>
<creator id="C. Blanchette">
<individualName>
<givenName>Carol</givenName>
<surName>Blanchette</surName>
</individualName>
<organizationName>PISCO</organizationName>
<address>
<deliveryPoint>UCSB Marine Science
Institute</deliveryPoint>
<city>Santa Barbara</city>
<administrativeArea>CA</administrativeArea>
<postalCode>93106</postalCode>
</address>
</creator>
<abstract>
<para>These temperature data were collected
at Alegria Beach, California, and were ...
</para>
</abstract>
<keywordSet>
<keyword>OceanographicSensorData</keyword>
<keyword>Thermistor</keyword>
<keywordThesaurus>
PISCO Categories
</keywordThesaurus>
</keywordSet>
<intellectualRights><para>Please contact the
authors for permission to use these data.
Please also acknowledge the authors in any
publications.</para>
</intellectualRights>
<contact>
<references>C. Blanchette</references>
</contact>
</dataset>
</eml:eml>
```

Transform

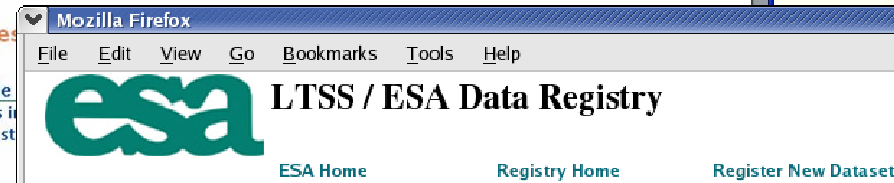


The screenshot shows a Mozilla browser window with a 'DATA CATALOG' search results page. The page has a blue header with 'DATA CATALOG' and 'Search' in yellow. A sidebar on the left contains a logo and links for 'data catalog', 'login', 'search', and 'insert'. The main content area displays the following information:

- Metadata Identifier:** piscoUCSB.5.20
- Short Name:** Alegria Temperatures
- Title:** PISCO: Intertidal Temperature Data: Alegria, California: 1996-1997
- Individual:** Dr. Carol Blanchette
- Organization:** PISCO
- Address:** UCSB Marine Science Institute, Santa Barbara, CA 93106
- Abstract:** These temperature data were collected at Alegria Beach, California, and were part of an array of intertidal temperature sensors that extend from Piedras Blancas, on the California Central Coast, to Port Hueneme, on the California South Coast. Sensors are set to a frequency of 20 minutes per measurement, and the units are collected approximately once every couple of months (up to ~6 months). The data are downloaded and saved as BoxCar(tm) .DTF files, and are then converted to ASCII Comma Separated Values files for archive purposes. Please see the methods section for detailed processing descriptions.
- Keywords:**
 - intertidal
 - temperature

Data Registries

- Registries
 - UC NRS
 - OBFS
 - ESA LTSS
 - Spec Net
 - NCEAS
- Use metacat
- Web-based metadata entry



16 data packages found

Title	Contacts	Organization
» Cover of Plant Species in California Coastal Terrace Grasslands.	Stromberg Stromberg Mark Stromberg	Hastings Natural History Reservation

ID: nrs. 8.4

» Plant Species in Carmel Valley Grasslands.	berkeley Mark Stromberg berkeley Stromberg Stromberg	University of California Natural Reserve System
--	--	---

6.3

c interactions between plant distributions and generated disturbances: trajectories of restored	Seabloom Seabloom	Sedgwick Reserve
--	----------------------	------------------



29 data packages found

Title
» Amphibian and reptile species for the Pymatuning area

ID: obfs.334.2

» A survey of the aquatic vascular plants of seven natural
lakes in Northwestern Pennsylvania

ID: obfs.338.2

» Bird species for the Pymatuning area

ID: obfs.328.2

» Browns Creek, stream and groundwater temperature

ID: obfs.336.1



Organization of Biological Field Stations Data Registry

OBFS Home Registry Home Register New Dataset Search for Data

176 data packages found

Title	Contacts	Organization	Keywords	Actions
Florida sandhill crane territory monitoring dataset for Disney Wilderness Preserve, Florida	Sandy Woiak Woiak The Nature Conservancy	Organization of Biological Field Stations	Florida Wildlife monitoring Productivity Florida sandhill crane Survival Population dynamics Endangered Species	View Edit Delete
ID: obfs.114.3				
Florida scrub-jay territory monitoring	Sandy Woiak	Organization of	Population dynamics	View

Morpho – Managing metadata and data

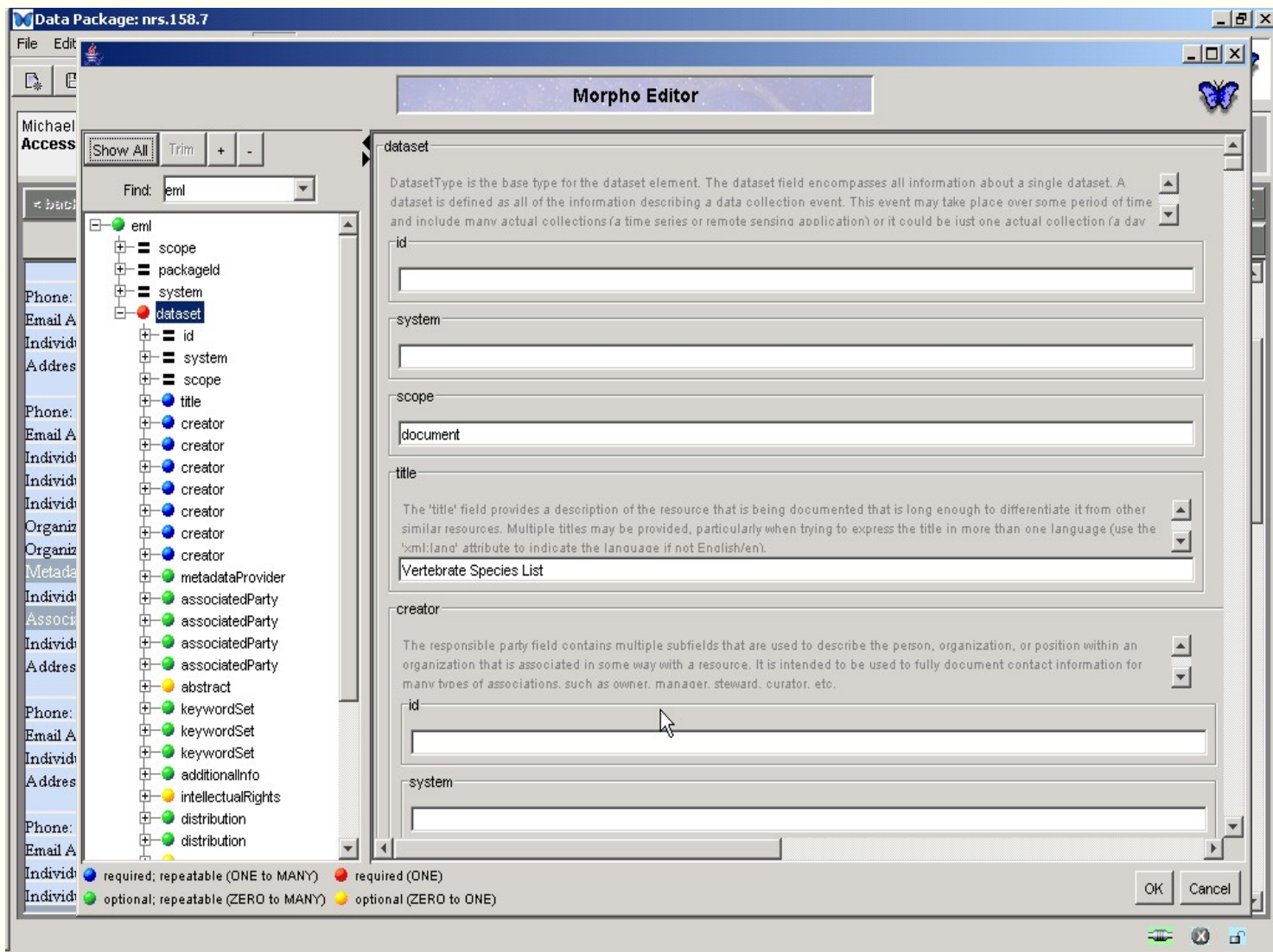
- Create and manage data and metadata using wizards
- Search and locate data
- Share data with colleagues
- Usable in field station environments
 - Can't assume a network connection is available always
- Can be run on Windows, Linux and Mac OS
- Is open source (GNU GPL)



What can Morpho do?

- Create EML data packages
 - Open and edit data packages, set access control
 - Import data into a data package
 - Save the EML and data
 - locally and to metacat servers
- Search for EML
 - locally and on metacat servers
- Export data packages

Morpho: Open and edit data packages



Morpho: Create data packages

- Data Package Wizard

The screenshot displays the Morpho Data Package Wizard interface with several overlapping windows:

- New Data Package Wizard - Welcome to the Data Package Wizard:** The main window, showing the "Define Temporal Coverage:" section. It includes radio buttons for "Single Point in Time" and "Range of Date/Time" (selected). Below, there are fields for "Enter starting date:" and a calendar for February 2005.
- New Data Package Wizard - Title and Abstract:** A window for entering the title and abstract of the data package.
- Define Access:** A window for selecting users and defining access levels. It features a table of users and a list of access levels.
- Named Regions:** A window for selecting named regions from a list.

Define Access: Select a user or group from the list below:

Name	Email / Description / Distinguished Name
Access Tree	
SDSC	
OBFS	
UCNRS	
A Tester	atester@ucnrs.org, mottrott@nceas.ucsb.edu
Alexander Glazer	aglazer@ucnrs.org, alexander.glazer@ucop.edu
Alicia Flammia	aflammia@ucnrs.org, alicia_flammia@hotmail.com
Allan Muth	amuth@ucnrs.org, deepcanyon@mindspring.com
Andrew Brooks	abrooks@ucnrs.org, brooks@lifesci.ucsb.edu
Arnulfo Lozoya	alozya@ucnrs.org, lozoya@gte.net

Description of access levels:

- Read: Able to view data package.
- Read & Write: Able to view and modify data package.
- Read, Write & Change Permissions: Able to view and modify datapackage, and modify access permissions.
- All: Able to do everything (this is the same as Read, Write & Change Permissions)

Named Regions:

- Lake Erie Center
- Lake Itasca Forestry and Biological Station
- Lake Louise Field Station
- Lake Michigan Biological Station
- Landels-Hill Big Creek Reserve
- Landels-Hill Big Creek Reserve UCNRS
- Las Cuevas Forest Research Station

Morpho: Import data

- New Data Table Wizard

The image shows two overlapping windows from the Morpho software. The background window is titled 'New Datatable Wizard' and has a tab labeled 'Data File Information:'. It contains a section 'File Format' with the instruction 'Enter some information about your data file.' Below this, it asks 'What is the format of your data?' and offers three radio button options: 'Simple delimited text format (uses one or more delimiters throughout the file)', 'Complex text format (delimited fields, fixed width fields, and mixtures of text and numbers)', and 'Non-text or proprietary formatted file that is externally defined (e.g. "Microsoft Excel")'. The first option is selected. Below the options, it says 'Simple delimited text format (uses one or more delimiters throughout the file)'. Then, it asks 'Data Attributes are arranged in:' with two radio button options: 'Columns' (selected) and 'Rows'. Below that, it says 'Define one or more delimiters used to indicate the ends of fields:' and lists five checkboxes: 'tab', 'comma', 'space', 'semicolon', and 'other'. The 'other' checkbox is selected, and there is a text input field next to it. The foreground window is titled 'Define Attribute or Column:'. It has a 'Name:' field with the text 'Length' and a description 'Name of the attribute as it appears in the data file'. Below that is a 'Label:' field with the description 'A more readable label for the attribute'. Then is a 'Definition:' field with the description 'Define the contents of the attribute (or column) precisely, so that a data user could interpret the attribute accurately. e.g. "spden" is the number of individuals of all macro invertebrate species found in the plot'. Below the definition field are five radio button options for 'Category': 'Unordered: unordered categories or text (statistically **nominal**) e.g. Male, Female' (selected), 'Ordered: ordered categories (statistically **ordinal**) e.g. Low, High', 'Relative: values from a scale with equidistant points (statistically **interval**) e.g. 12.2 meters', 'Absolute: measurement scale with a meaningful zero point (statistically **ratio**) e.g. 273 Kelvin', and 'Date-Time: date or time values from the Gregorian calendar e.g. 2002-10-24'. There is a 'Help' button next to the 'Category:' label. Below the category options is a section for 'Unordered' with a 'Choose:' dropdown menu set to 'Enumerated values (belong to predefined list)' and a description 'Describe any codes that are used as values of the attribute.' Below that is a 'Location:' dropdown menu set to 'Codes are defined here'. Then is a table with two columns: 'Code' and 'Definition'. Below the table is a 'Definitions:' label. At the bottom of the 'Unordered' section is a checkbox 'Attribute contains free-text in addition to those values listed above'. To the right of the table are 'Add' and 'Delete' buttons. At the bottom right of the window are 'OK' and 'Cancel' buttons.

New Datatable Wizard

Data File Information:

File Format

Enter some information about your data file.

What is the format of your data?

- ☒ Simple delimited text format (uses one or more delimiters throughout the file)
- ☐ Complex text format (delimited fields, fixed width fields, and mixtures of text and numbers)
- ☐ Non-text or proprietary formatted file that is externally defined (e.g. "Microsoft Excel")

Simple delimited text format (uses one or more delimiters throughout the file)

Data Attributes are arranged in:

- ☒ Columns
- ☐ Rows

Define one or more delimiters used to indicate the ends of fields:

- ☐ tab
- ☐ comma
- ☐ space
- ☐ semicolon
- ☒ other

Delimiter(s)

Define Attribute or Column:

Name: Name of the attribute as it appears in the data file

Label: A more readable label for the attribute

Definition: Define the contents of the attribute (or column) precisely, so that a data user could interpret the attribute accurately.
e.g. "spden" is the number of individuals of all macro invertebrate species found in the plot

Category:

- ☒ Unordered: unordered categories or text (statistically **nominal**) e.g. Male, Female
- ☐ Ordered: ordered categories (statistically **ordinal**) e.g. Low, High
- ☐ Relative: values from a scale with equidistant points (statistically **interval**) e.g. 12.2 meters
- ☐ Absolute: measurement scale with a meaningful zero point (statistically **ratio**) e.g. 273 Kelvin
- ☐ Date-Time: date or time values from the Gregorian calendar e.g. 2002-10-24

Help

Unordered

Choose: Describe any codes that are used as values of the attribute.

Location:

Code	Definition
<input type="text"/>	

Definitions:

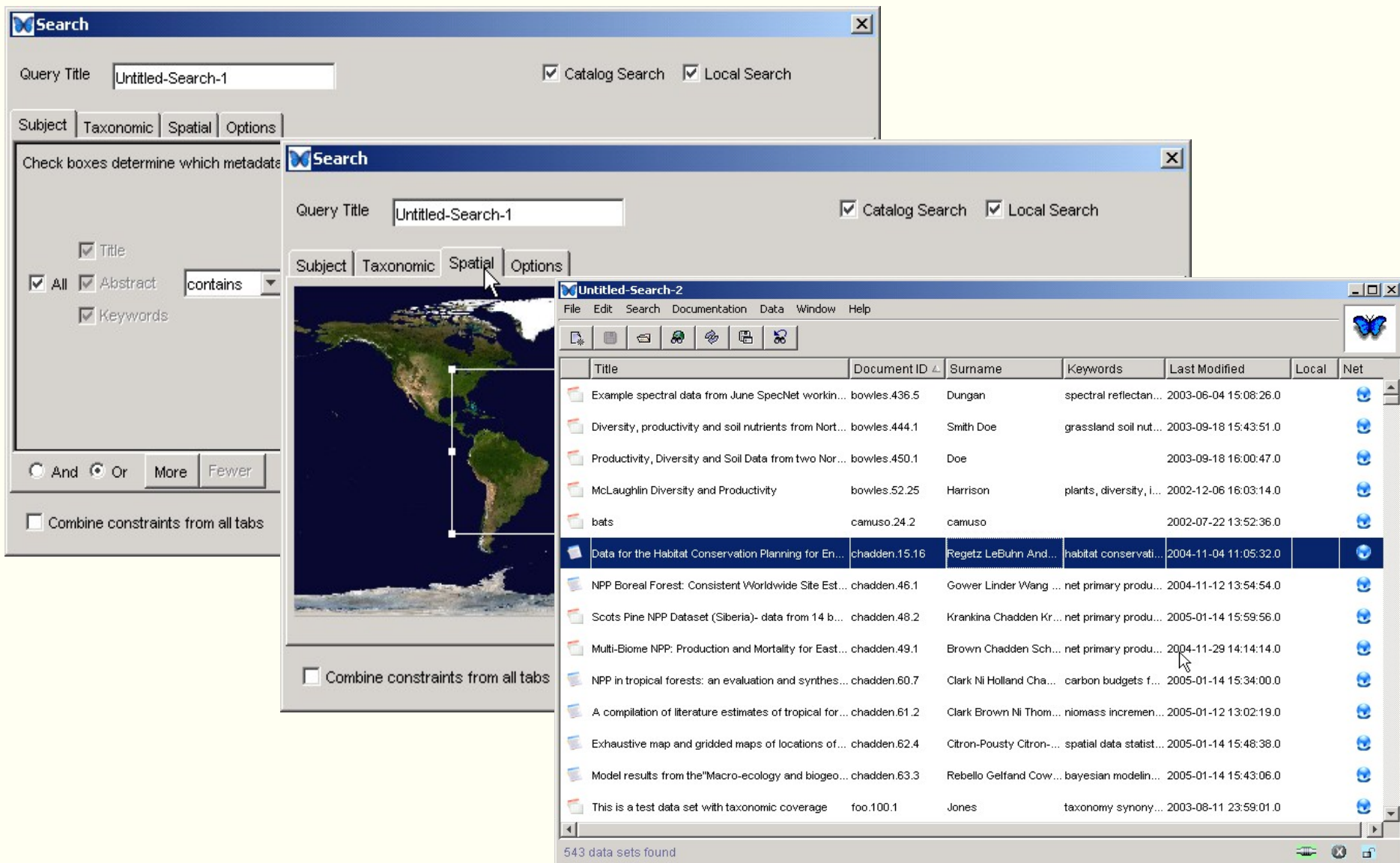
☐ Attribute contains free-text in addition to those values listed above

OK Cancel

Morpho: defining attributes

- Natural language description of each attribute
- Standardized measurement scales
 - Precisely define attribute domain
- Standardized measurement units
 - EML has unit dictionary with 140 pre-defined units
 - Relationship and conversion to base SI unit provided
 - Uses STMML to quantitatively define custom units

Morpho: Search for data packages



Search

Query Title:

☒ Catalog Search ☒ Local Search

Subject | Taxonomic | Spatial | Options

Check boxes determine which metadata

☒ Title
☒ All ☒ Abstract
☒ Keywords

☐ And ☒ Or

☐ Combine constraints from all tabs

Search

Query Title:

☒ Catalog Search ☒ Local Search

Subject | Taxonomic | Spatial | Options

Untitled-Search-2

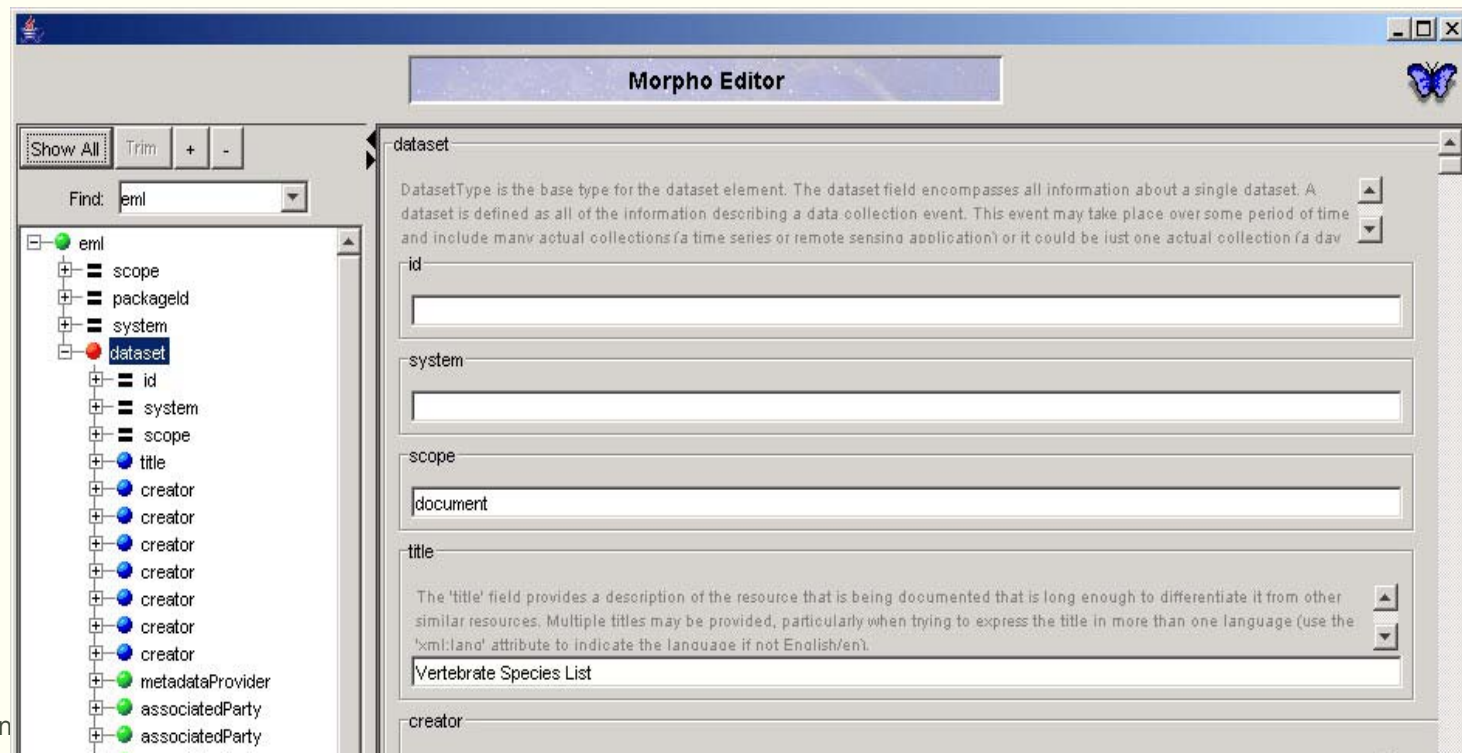
File Edit Search Documentation Data Window Help

Title	Document ID	Surname	Keywords	Last Modified	Local	Net
Example spectral data from June SpecNet workin...	bowles.436.5	Dungan	spectral reflectan...	2003-06-04 15:08:26.0		
Diversity, productivity and soil nutrients from Nort...	bowles.444.1	Smith Doe	grassland soil nut...	2003-09-18 15:43:51.0		
Productivity, Diversity and Soil Data from two Nor...	bowles.450.1	Doe		2003-09-18 16:00:47.0		
McLaughlin Diversity and Productivity	bowles.52.25	Harrison	plants, diversity, i...	2002-12-06 16:03:14.0		
bats	camuso.24.2	camuso		2002-07-22 13:52:36.0		
Data for the Habitat Conservation Planning for En...	chadden.15.16	Regetz LeBuhn And...	habitat conservati...	2004-11-04 11:05:32.0		
NPP Boreal Forest: Consistent Worldwide Site Est...	chadden.46.1	Gower Linder Wang ...	net primary produ...	2004-11-12 13:54:54.0		
Scots Pine NPP Dataset (Siberia)- data from 14 b...	chadden.48.2	Krankina Chadden Kr...	net primary produ...	2005-01-14 15:59:56.0		
Multi-Biome NPP: Production and Mortality for East...	chadden.49.1	Brown Chadden Sch...	net primary produ...	2004-11-29 14:14:14.0		
NPP in tropical forests: an evaluation and synthes...	chadden.60.7	Clark Ni Holland Cha...	carbon budgets f...	2005-01-14 15:34:00.0		
A compilation of literature estimates of tropical for...	chadden.61.2	Clark Brown Ni Thom...	biomass incremen...	2005-01-12 13:02:19.0		
Exhaustive map and gridded maps of locations of...	chadden.62.4	Citron-Pousty Citron...	spatial data statist...	2005-01-14 15:48:38.0		
Model results from the "Macro-ecology and biogeo...	chadden.63.3	Rebello Gelfand Cow...	bayesian modelin...	2005-01-14 15:43:06.0		
This is a test data set with taxonomic coverage	foo.100.1	Jones	taxonomy synony...	2003-08-11 23:59:01.0		

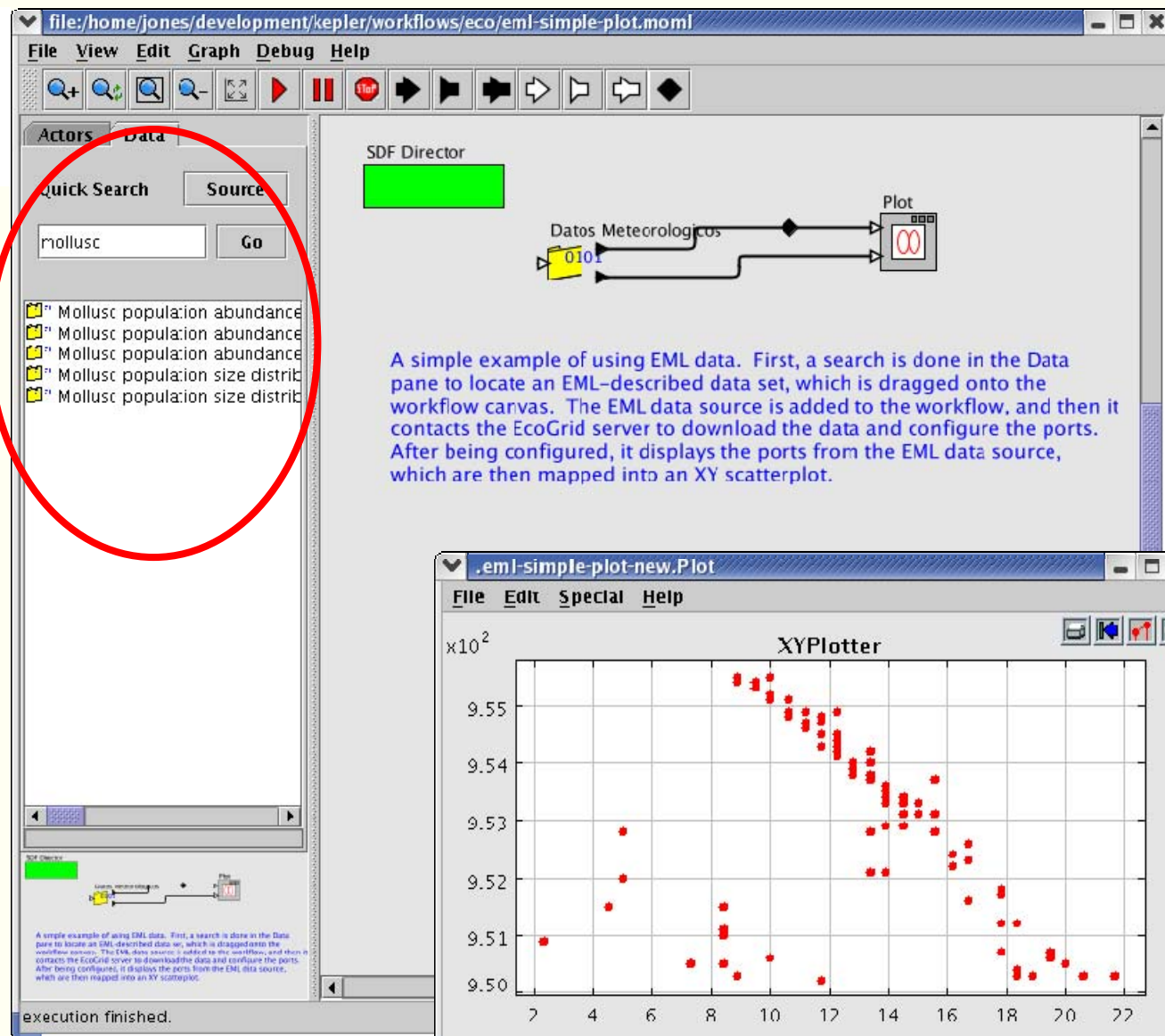
543 data sets found

Morpho extensibility

- Originally designed to support multiple metadata standards
 - Later customized for ease-of-use for EML
 - Could be extended to support editing BDP or other standards
- Could add conversion capabilities to import/export BDP



EML-driven analyses in Kepler



EML Metadata Display in Kepler

file:/home/jones/.kepler/cache/html/knb-lter-gce.109.5

File View Help

Data Set Description

Identifier: knb-lter-gce.109.5
Catalog System: knb
Alternate Identifier: INV-GCEM-0305a1 1 1
Title: **Mollusc population abundance monitoring: Fall 2000 mid-marsh and creekbank infaunal and epifaunal mollusc abundance collections from GCE marsh, monitoring sites 1-10**

Data Set Owner(s):

Organization: **Georgia Coastal Ecosystems LTER Project**
Address: Dept. of Marine Sciences,
University of Georgia,
Athens, Georgia 30602-3636 USA
Email Address: gcelter@uga.edu
Web Address: <http://gce-lter.marsci.uga.edu/lter/>

Individual: **Dr. Dale Bishop**
Organization: **University of Georgia**
Address: Dept. of Marine Sciences,
University of Georgia,
Athens, Georgia 30602-3636 USA
Email Address: tdbish@uga.edu
Web Address: <http://lmer.marsci.uga.edu/bios/bishop.html>

Metadata Provider(s):

Organization: **Georgia Coastal Ecosystems LTER Project**
Address: Dept. of Marine Sciences,
University of Georgia,
Athens, Georgia 30602-3636 USA
Email Address: gcelter@uga.edu
Web Address: <http://gce-lter.marsci.uga.edu/lter/>

Associated Party:

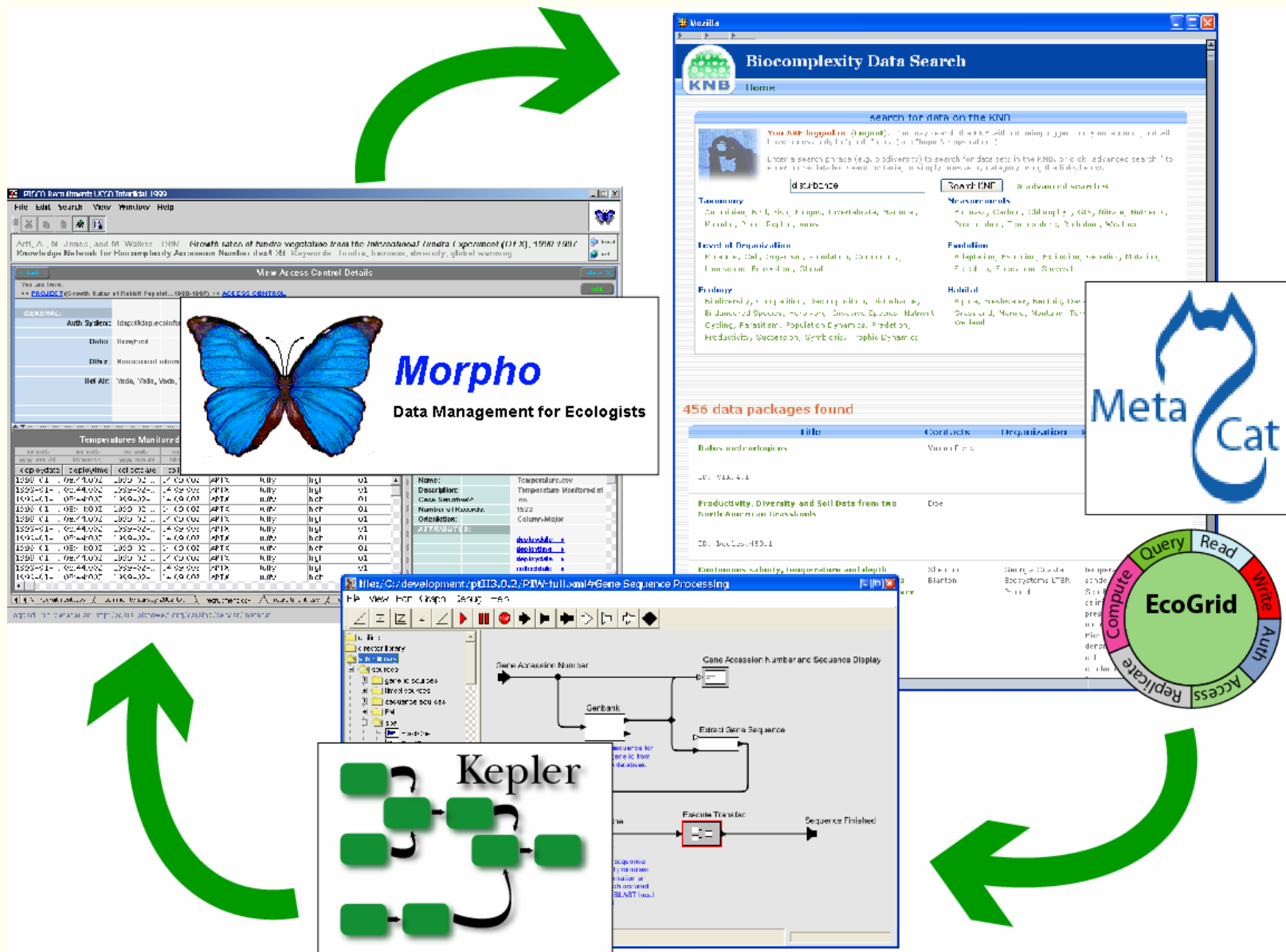
Individual: **Dr. Merryl Alber**
Organization: **University of Georgia**
Email Address: malber@uga.edu

Individual: **Mr. Kenneth Helm**
Organization: **University of Georgia Marine Institute**
Email Address: khelm@darientel.net

Abstract:

This data set is the Fall 2000 estimate of infaunal and epifaunal mollusc abundance at the GCE-LTER marsh sites used for population monitoring. Species abundance was determined by hand-collecting all the infaunal and epifaunal molluscs from within quadrats in mid-marsh and creekbank zones (n = 4 quadrats per zone) at all sites. The molluscs were returned to the lab, fixed in formalin and preserved in ethanol, counted and measured (size data is reported separately). The counts were converted to number per square meter.

Metadata-driven analysis cycle



Acknowledgements

This material is based upon work supported by:

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

Collaborators: NCEAS (UC Santa Barbara), University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research), University of Vermont, University of North Carolina, Napier University, Arizona State University, UC Davis

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.

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