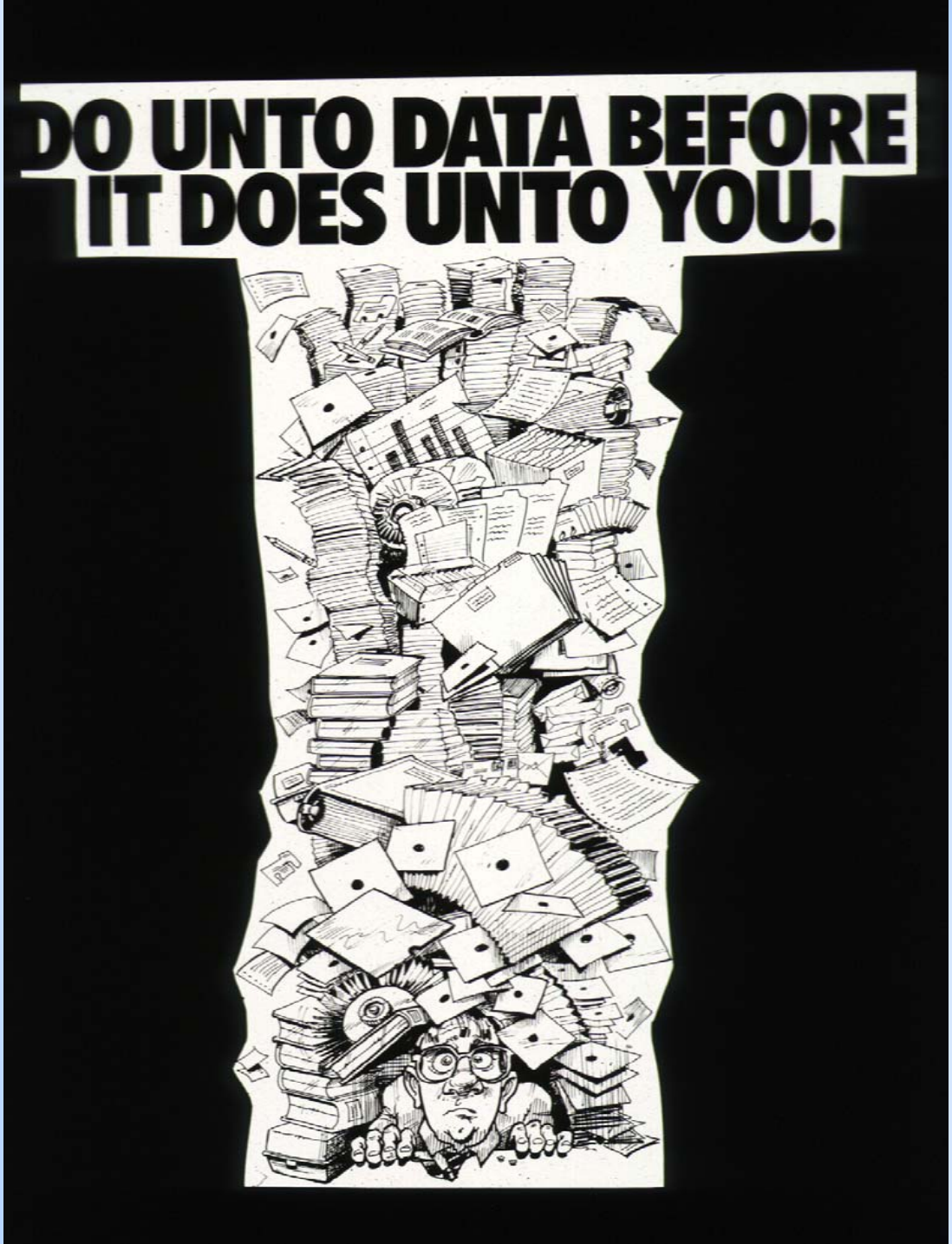
Ecological Metadata

John Porter
University of Virginia



Slide from
Susan Stafford

The Ecoinformatics Challenge:

- Can we make information available to ecologists:
  - In ways they can **locate** the information they need?
  - With information in forms they can readily **use**?
- How can we assure that the information is current and accurate?

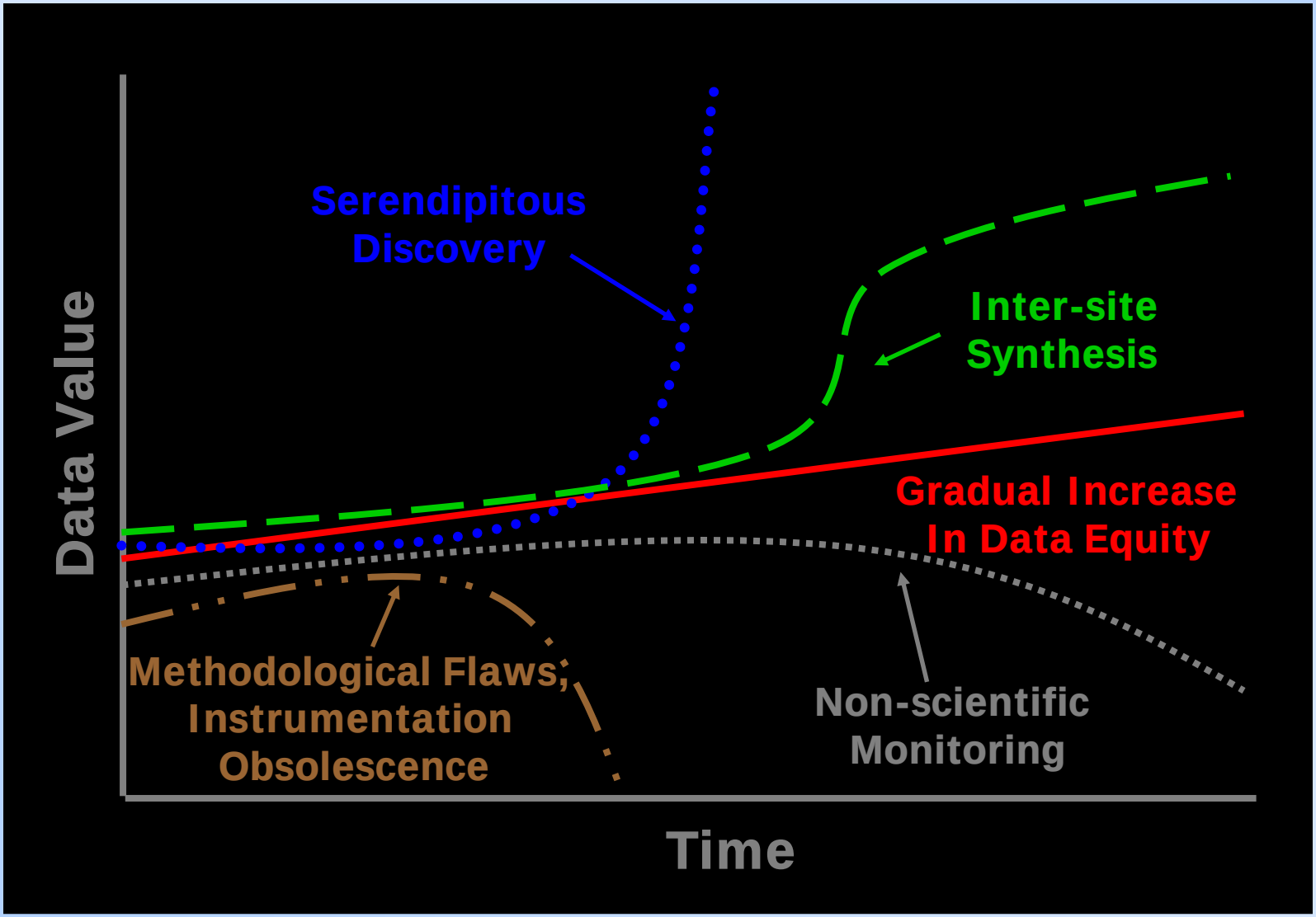
Why Metadata?

Metadata is needed to reach our goals:

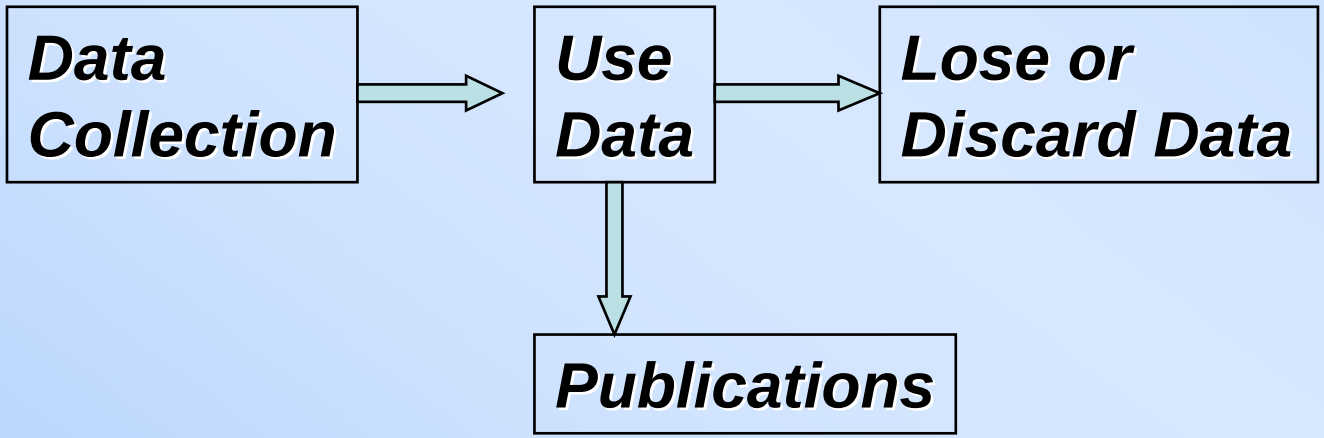
- Enabling new scientific approaches by making data available
  - Regional and Global Analyses
  - Multi-disciplinary syntheses
  - Multi-factorial analyses
- Creating data that is still useful 20-years or more into the future

Slide from William Michener & James Brunt

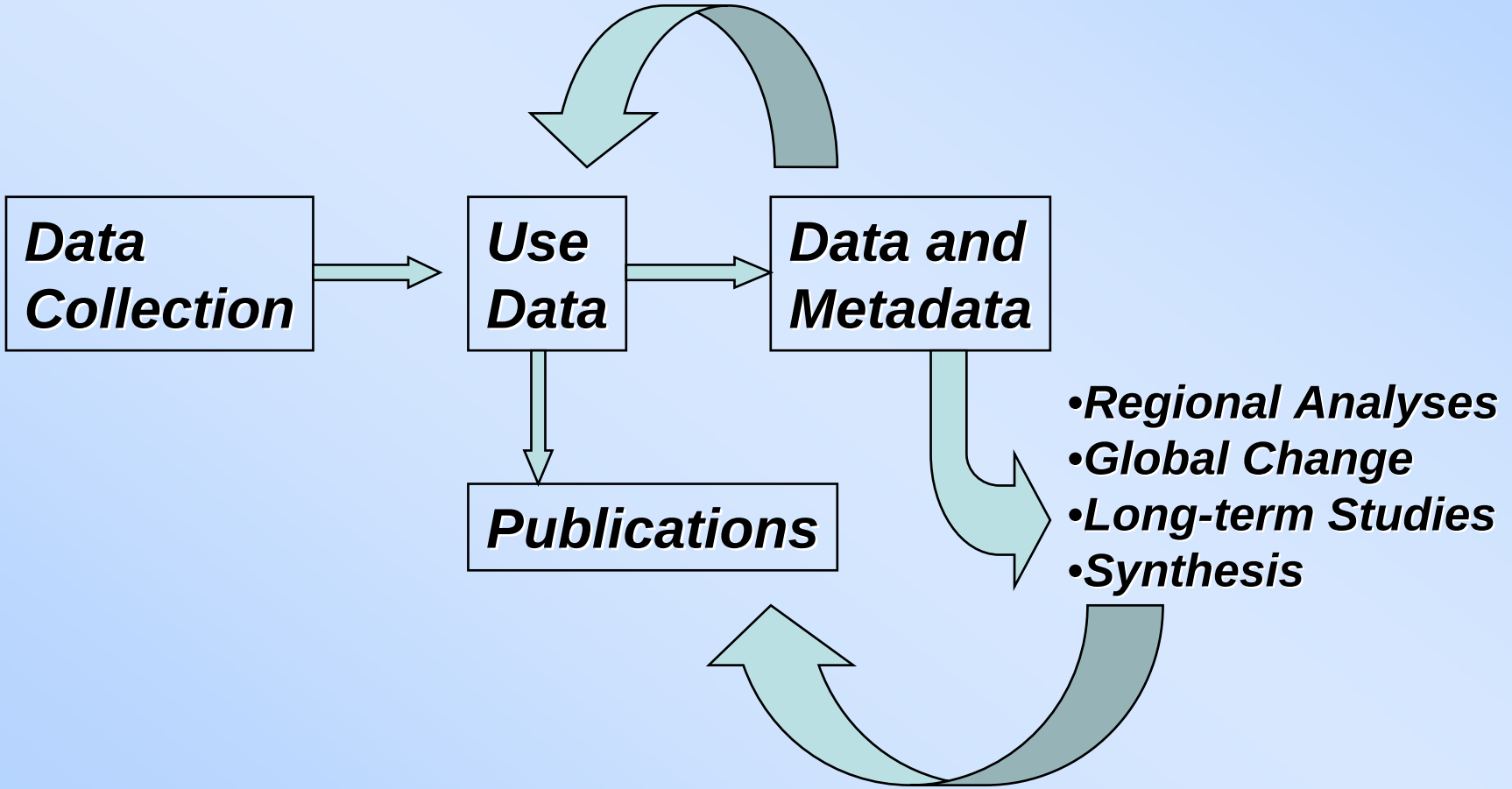
- ✓ Increasing value of data over time



Evolution of Data Sharing - Traditional Model



Evolution of Data Sharing – A New Model



The Critical Roles of Metadata

- **Data Discovery**
  - Scientists need to be able to identify important data sets
- **Data Retrieval**
  - Scientists need know how and where to access data
- **Data Use**
  - Scientists need to know enough details about how the how the data were collected and stored
- **Data Archiving**
  - Ecological data can grow more valuable with time, but only if the critical information required to retrieve and interpret the data remains available

Metadata Uses

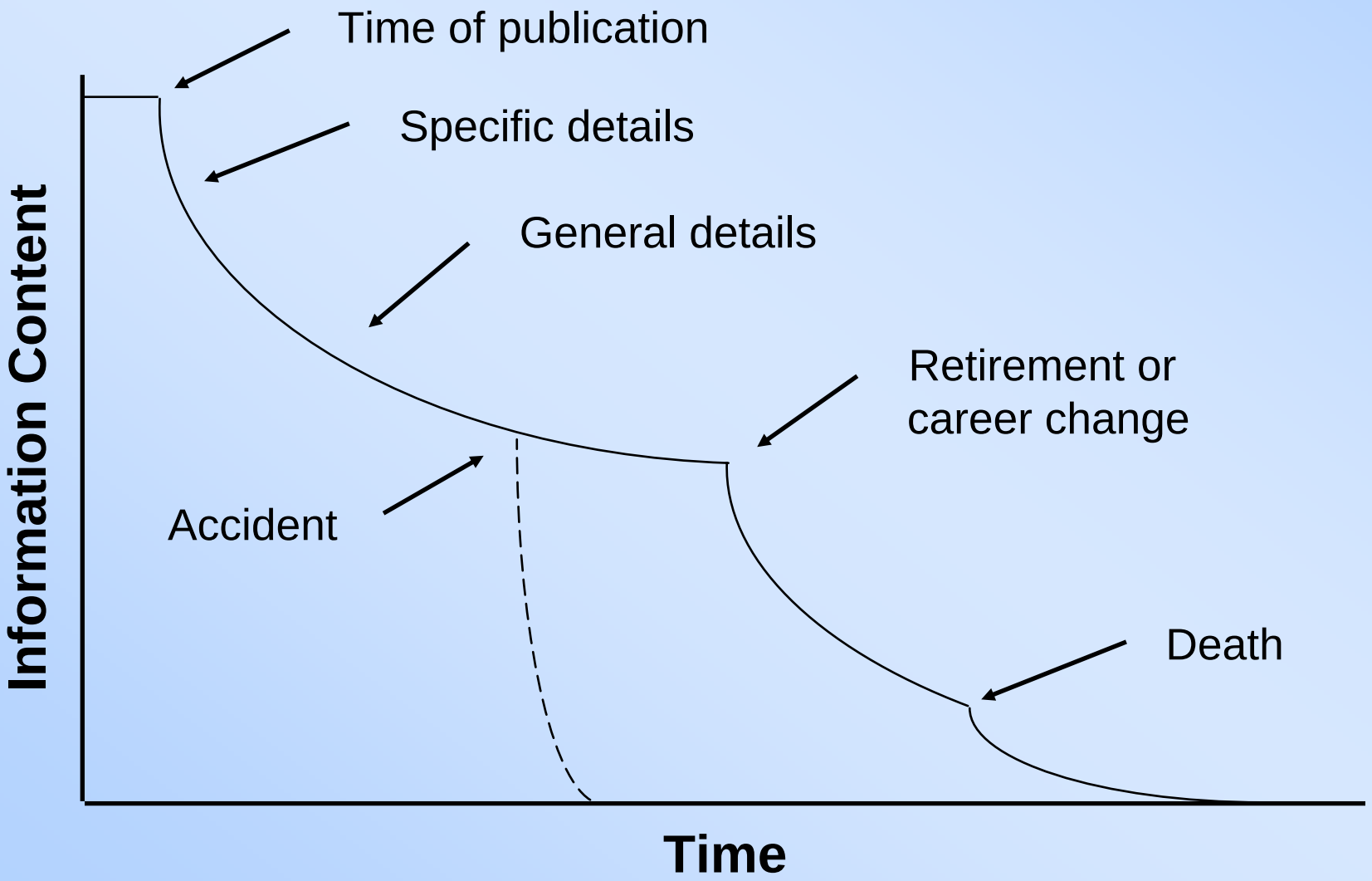
Metadata descriptor classes	Levels of secondary data utilization and associated metadata content		
	Level I: exchange with expert colleague	Level II: searchable and third party data reuse	Level III: publishable and auditable
I. Data set descriptors	X	X	X
II. Research origin descriptors		X	X
III. Data set status and accessibility		X	X
IV. Data structural descriptors	X	X	X
V. Supplemental descriptors			X

Michener et al. 1997 – NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES, Ecological Applications Vol. 7, No. 1, pp. 330–342.

Challenges

- Entropy is not easily conquered! The natural tendency is towards increasing disorder (2nd law of thermodynamics)
- The diversity of ecological data
  - Genome data is much simpler!
- The diversity of users and uses to which data will be put

The Challenge: Fighting Data Entropy (Michener et al. 1997)



Existing Metadata Standards

- General Standards
  - Dublin Core - designed for publications
- Spatial Metadata
  - ISO 19115 – International Standard
  - FGDC (Federal Geographic Data Committee)
- Taxonomic & Collection Data
  - Darwin Core
- Catalog
  - Directory Interchange Format (DIF)

Common Elements

- Each of the existing metadata standards includes:
  - **Who** – information on who to contact
  - **What** – a description of the available item
- Some include:
  - **Where** – geographical locations
  - **When** – dates and times
- Standards vary widely in the degree of structure from free text to XML and database schema

***The wonderful thing about standards is that there are so many of them to choose from….. -* Anonymous**

Metadata – Two examples

Benson, Barbara

Trout Lake Temperature

Water temperature data was collected hourly at Trout Lake

Trout Lake Temperature

Water temperature data was collected hourly at Trout Lake from January 1, 2005 to December 31, 2005

Collected by Barbara Benson

Both metadata documents are readable by humans, can't be processed by a computer into new forms because both the content and structure are different

Metadata – Content Standardized

Originator: Benson, Barbara

Trout Lake Temperature

Water temperature data was collected hourly at Trout Lake

Start Date: January 1, 2005

End Date: December 31, 2005

Trout Lake Temperature

Water temperature data was collected hourly at Trout Lake

Time period: January 1, 2005 to December 31, 2005

Originator: Benson, Barbara

Now both metadata documents have the same content – title, originator, etc. but it still can't be automatically processed by a computer because the structure is different

Metadata – Structure Standardized

```
<title>Trout Lake Temperature</title>
<originator>
<namelast>Benson</namelast>
<namefirst>Barbara</namefirst>
</originator>
<abstract>Water temperature data
    was collected hourly at Trout
    Lake</abstract>
<date>
<start>January 1, 2005</start>
<end>December 31, 2005</end>
</date>
```

```
<title>Trout Lake Temperature</title>
<originator>
<namelast>Benson</namelast>
<namefirst>Barbara</namefirst>
</originator>
<abstract>Water temperature data
    was collected hourly at Trout
    Lake </abstract>
<date>
<start>January 1, 2005</start>
<end>December 31, 2005</end>
</date>
```

With standardized content and structure, computers can automatically extract information from the metadata!

Why Ecological Metadata Language?

- Other existing standards do not support the development of interoperable systems
  - Lack needed structure
  - Lack needed content to support automated access to common ecological data formats (e.g., delimited text, formatted text, SQL databases)

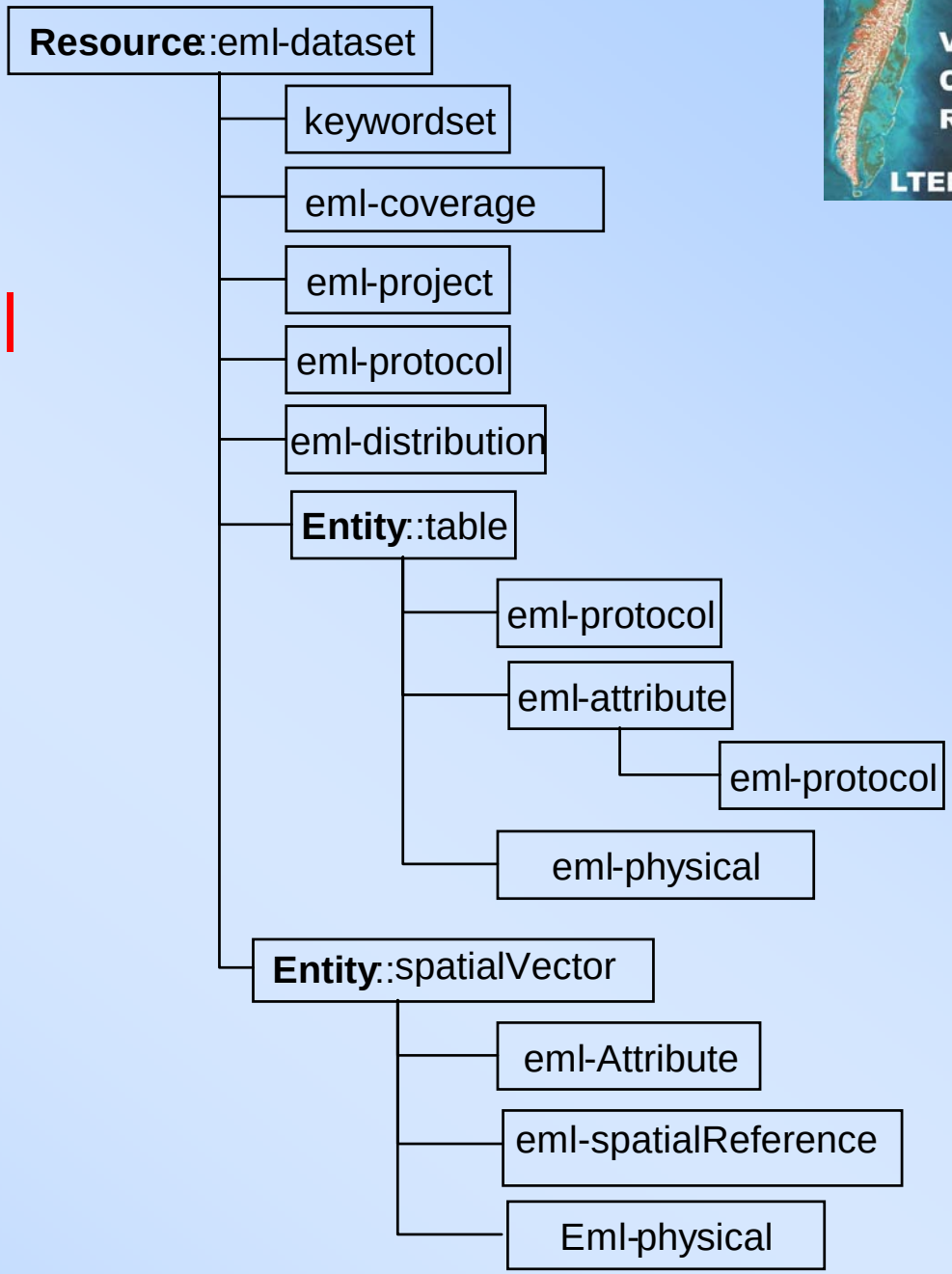
Introduction to Ecological Metadata Language (EML)

- Currently EML version 2.01 is being implemented by US LTER sites
- EML uses eXtensible Markup Language (XML) schema to define content and structure
- The EML standard is made up of a number of different modules

EML 2.01 – Major Modules

- **eml-resource** -- bibliographic info
- **eml-party** -- people and organizations
- **eml-entity** -- file/object info
- **eml-attribute** -- variable/attribute info
- **eml-access** -- access control
- **eml-distribution** -- distribution info

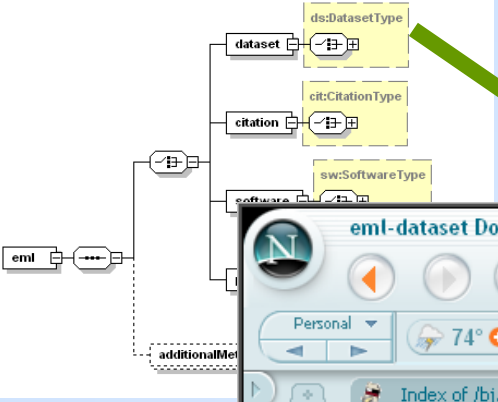
- EML -Ecological Metadata Language – different EML modules are nested within other modules



Full list of EML 2.01 Modules

- eml
- eml-access
- eml-attribute
- eml-constraint
- eml-coverage
- eml-dataset
- eml-dataTable
- eml-entity
- eml-literature
- eml-methods
- eml-party
- eml-physical
- eml-project
- eml-protocol
- eml-resource
- eml-software
- eml-spatialRaster
- eml-spatialReference
- eml-spatialVector
- eml-storedProcedure
- eml-text
- eml-view

EML Documentation describes each module and how it is used



http://www.ecoinformatics.org

eml-dataset Documentation - Netscape Browser

File Edit Vie Microphone Tools

est Practices LTER SEARCH http://knb.ecoinformatics.org/software/eml/eml-2.0.1/eml-dat GO SECURITY CENTER

Personal 74° Setup E-mail Netscape.com Inside Netscape

Index of /bj... 2005 LTER I... ESA > Meet... http:/...7.html ecoinformat... eml-dataset Documentation

Content of this field:

Description of this field:

DatasetType is the base type for the dataset element. The dataset field encompasses all information about a single dataset. A dataset is defined as all of the information describing a data collection event. This event may take place over some period of time and include many actual collections (a time series or remote sensing application) or it could be just one actual collection (a day in the field).

Elements:	Use:	How many:
A choice of (		
A sequence of (		
res:ResourceGroup		
purpose	optional	
maintenance	optional	
contact	required	unbounded
publisher	optional	
pubPlace	optional	
methods	optional	
project	optional	
access	optional	
A choice of (		
dataTable	required	
OR		
spatialRaster	required	
OR		
spatialVector	required	
OR		
storedProcedure	required	
OR		
view	required	
OR		
otherEntity	required	

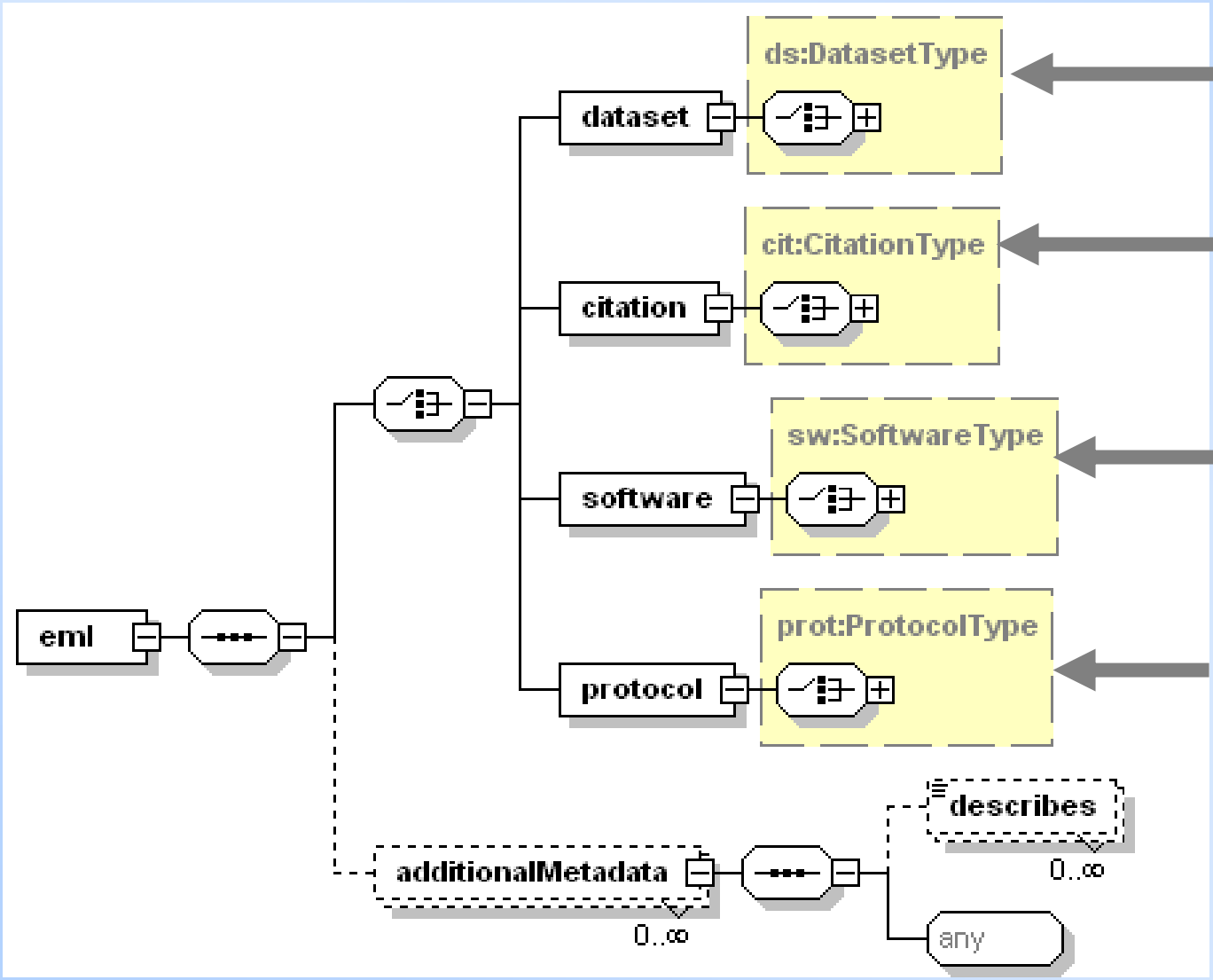
EML documents consist of nested modules each of which have OPTIONAL and REQUIRED elements.

Here is the EML specification for a dataset.

Find: EML Find Next Find Previous Highlight Match case

http://knb.ecoinformatics.org/software/eml/eml-2.0.1/eml-dataset.html#DatasetType

EML Diagram

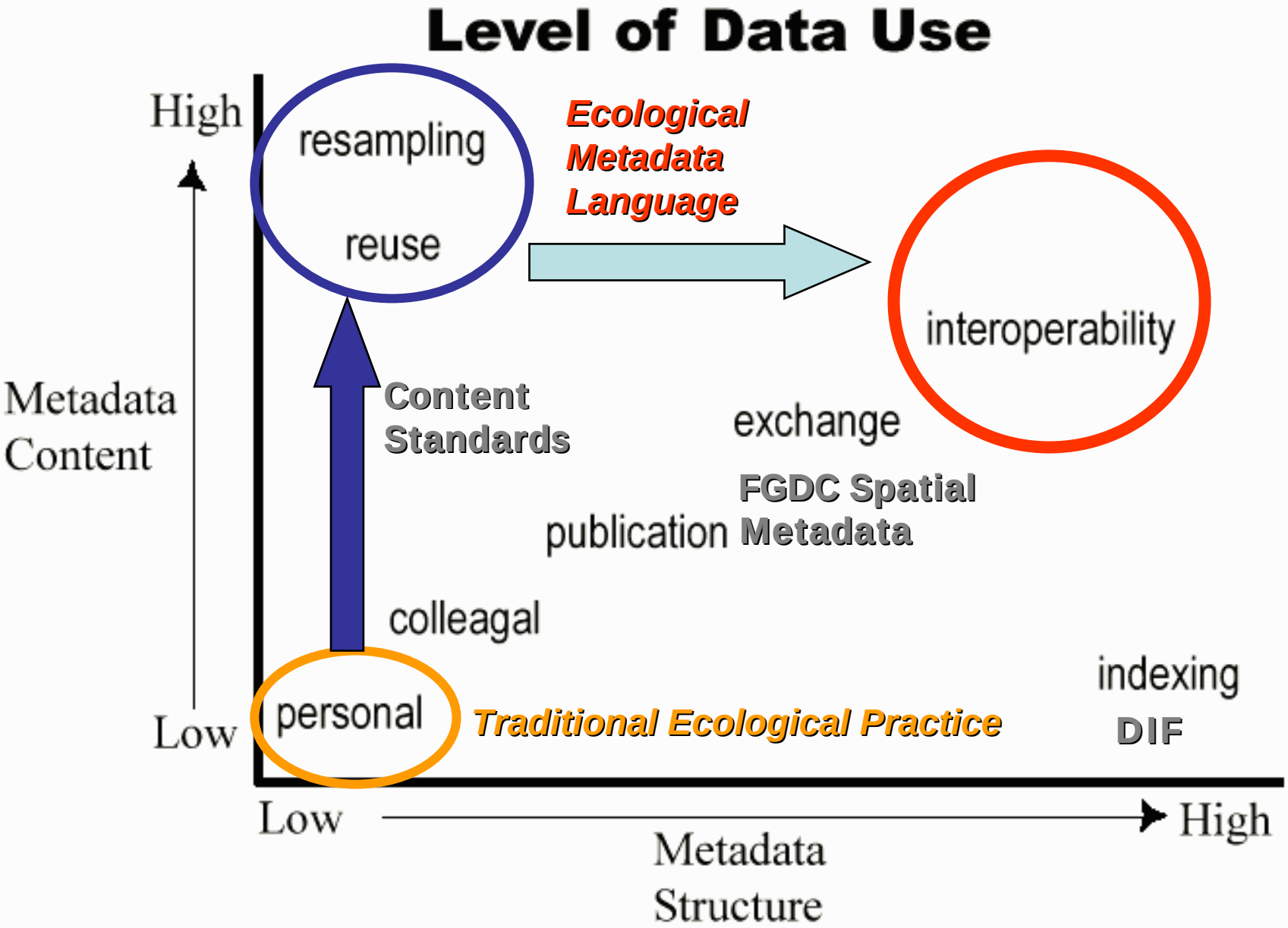


Data

Literature Citations

Software & Models

Procedures & Protocols

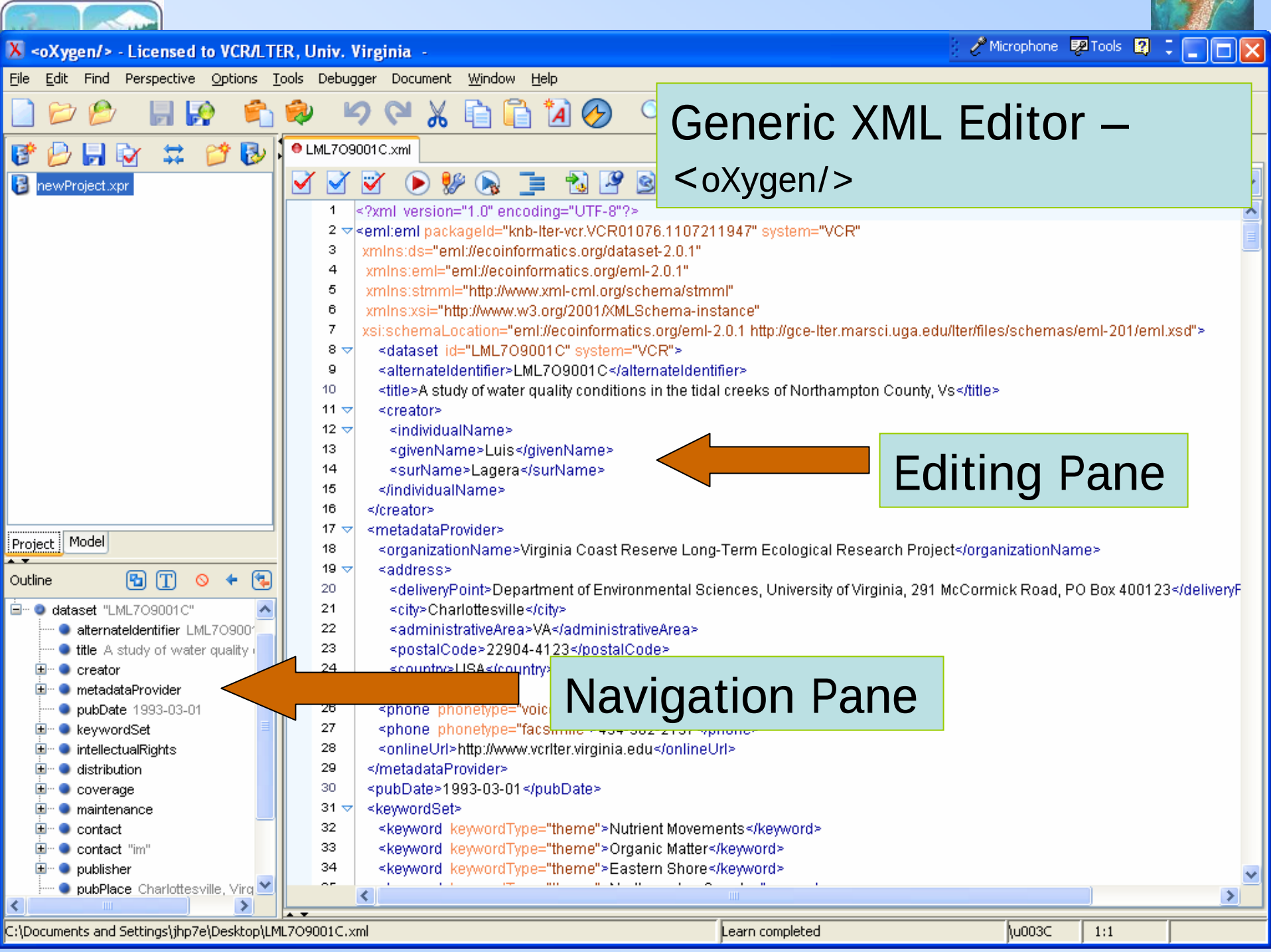


Modified from Slide by James Brunt

Producing EML

EML documents can be created in a variety of ways. These include:

- General XML Editors (e.g., oXygen, XMLspy)
- Customized XML Editors (e.g., Morpho)
- Output from metadata stored in databases
- Metadata stored in specialized spreadsheets


Generic XML Editor –
<oXygen/>
Editing Pane
Navigation Pane
<oXygen/> - Licensed to VCR/LTER, Univ. Virginia -
File Edit Find Perspective Options Tools Debugger Document Window Help
LML7O9001C.xml
newProject.xpr
Project Model
Outline
dataset "LML7O9001C"
alternateIdentifier LML7O900
title A study of water quality
creator
metadataProvider
pubDate 1993-03-01
keywordSet
intellectualRights
distribution
coverage
maintenance
contact
contact "im"
publisher
pubPlace Charlottesville, Virg
1 <?xml version="1.0" encoding="UTF-8"?>
2 <eml:eml packageId="knb-lter-vcr.VCR01076.1107211947" system="VCR"
3 xmlns:ds="eml://ecoinformatics.org/dataset-2.0.1"
4 xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
5 xmlns:stmml="http://www.xml-cml.org/schema/stmml"
6 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
7 xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 http://gce-lter.marsci.uga.edu/lter/files/schemas/eml-201/eml.xsd">
8 <dataset id="LML7O9001C" system="VCR">
9 <alternateIdentifier>LML7O9001C</alternateIdentifier>
10 <title>A study of water quality conditions in the tidal creeks of Northampton County, Vs</title>
11 <creator>
12 <individualName>
13 <givenName>Luis</givenName>
14 <surName>Lagera</surName>
15 </individualName>
16 </creator>
17 <metadataProvider>
18 <organizationName>Virginia Coast Reserve Long-Term Ecological Research Project</organizationName>
19 <address>
20 <deliveryPoint>Department of Environmental Sciences, University of Virginia, 291 McCormick Road, PO Box 400123</deliveryF
21 <city>Charlottesville</city>
22 <administrativeArea>VA</administrativeArea>
23 <postalCode>22904-4123</postalCode>
24 <country>USA</country>
26 <phone phonetype="voic
27 <phone phonetype="facs
28 <onlineUrl>http://www.vcrlter.virginia.edu</onlineUrl>
29 </metadataProvider>
30 <pubDate>1993-03-01</pubDate>
31 <keywordSet>
32 <keyword keywordType="theme">Nutrient Movements</keyword>
33 <keyword keywordType="theme">Organic Matter</keyword>
34 <keyword keywordType="theme">Eastern Shore</keyword>
C:\Documents and Settings\jhp7e\Desktop\LML7O9001C.xml
Learn completed
1:1

xls_eml_01_SAMPLE.xls

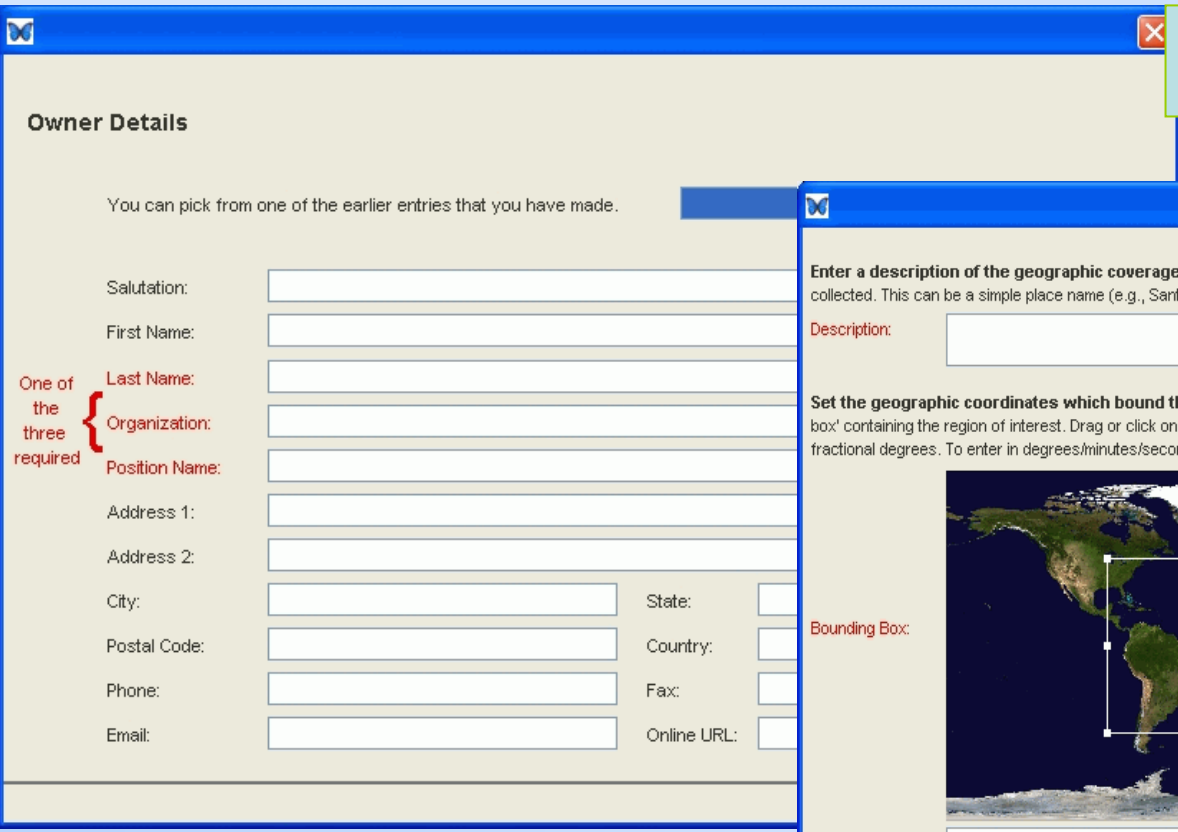
	B	C	D
17	**I. LTER Dataset Information**		
18	LTER site acronym	FCE	
19	Metacat package ID	LT_ND_Grahl_001.1	
20	Dataset LTER Identification Number	LT_ND_Grahl_001	
21	Dataset Title	Taylor Slough Water Quality Data for 2000-2001	
22	Dataset Creator Salutation	Dr.	
23	Dataset Creator First Name	Daniel	
24	Dataset Creator Last Name	Childers	
25	Dataset Creator Organization Name		Florida Coastal Everglades LTER Prograr
26	Dataset Creator Position Name		
27	Dataset Creator Mail Street Address	SERC|OE 148|Florida International University|University Park	ECS|253|Florida International University|U
28	Dataset Creator Mail City	Miami	Miami
29	Dataset Creator Mail State	FL	FL
30	Dataset Creator Mail Zip Code	33199	33199
31	Dataset Creator Mail Country	USA	USA
32	Dataset Creator Voice Telephone	305-348-3101	305-348-6054
33	Dataset Creator Facsimile Telephone	305-348-1986	305-348-4096
34	Dataset Creator Electronic Mail Address	childers@fiu.edu	fcelter@fiu.edu
35	Dataset Creator URL	http://www.fiu.edu/~ecosyst/	http://fcelter.fiu.edu/
36	**+-Dataset Creator**		
37	Dataset Abstract	Water quality samples are being collected using ISCO	
38	Dataset Keywords	inorganic nutrients	water quality
39	Dataset KeywordThesaurus		
40	Dataset Geographic Description	Shark River Slough Transect within Everglades National Park	
41	Dataset West Bounding Coordinate	-81.07794623	
42	Dataset East Bounding Coordinate	-80.72742805	
43	Dataset North Bounding Coordinate	25.76145171	
44	Dataset South Bounding Coordinate	25.36462994	
45	Dataset Beginning Temporal Coverage Date	2000-10-31	
46	Dataset Ending Temporal Coverage Date	2001-11-19	
47	Dataset Taxon Rank Name		
48	Dataset Taxon Rank Value		
49	Dataset Common Taxon Names		
50	**+-Dataset Coverage (Geographic, Temporal, Taxonomic)**		
51	Dataset Intellectual Rights		
52	Dataset Download URL		
53	Dataset Offline Medium Name		
54	Dataset Offline Medium Density		
55	Dataset Offline Medium Density Units		
56	Dataset Offline Medium Volume		
57	Dataset Offline Medium Format		
58	**+-Dataset Distribution**		
59	Dataset Associated Party First Name	Tim	
60	Dataset Associated Party Last Name	Grahl	

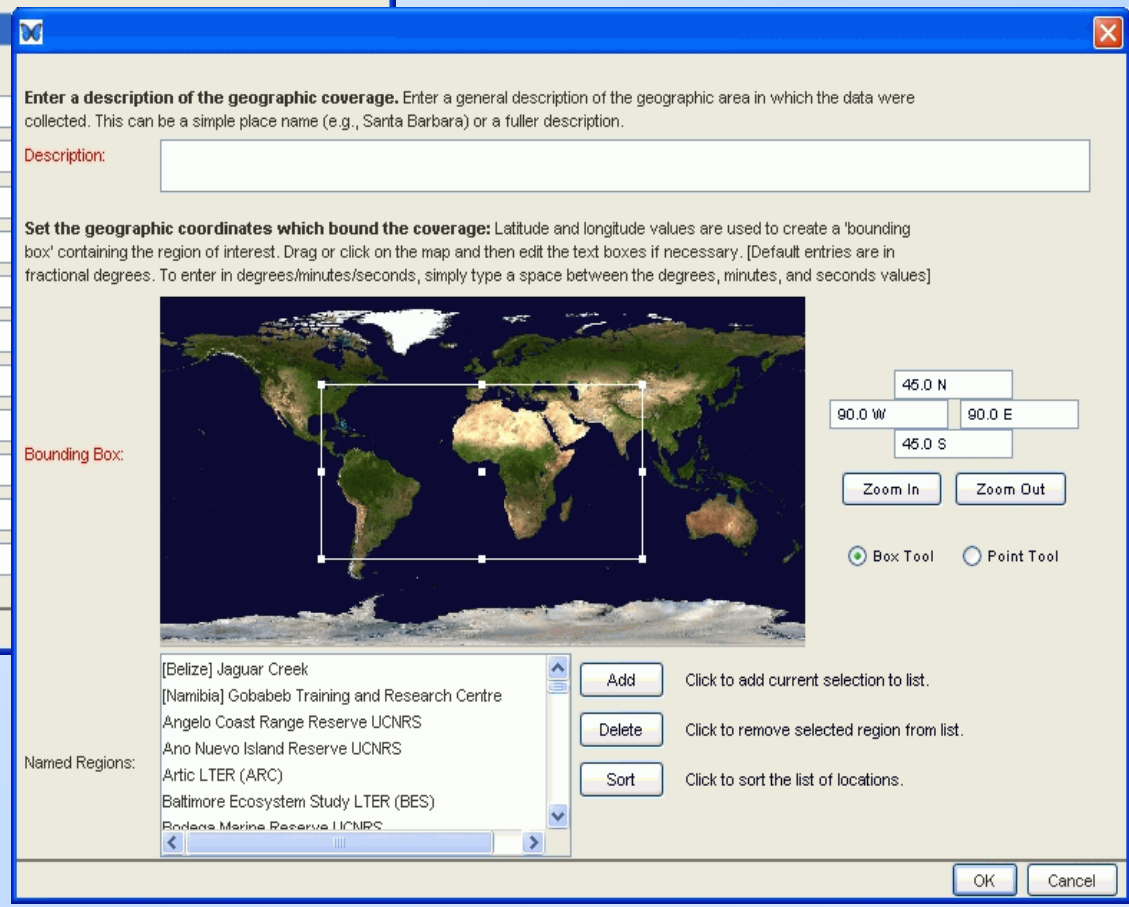
Excel spreadsheet for EML - Processed by a program to create EML document

General Metadata | MethodsCitation | MethodsProtocol | ResearchProjects | DataTable | References | IM Use Only | Units IM Use Only

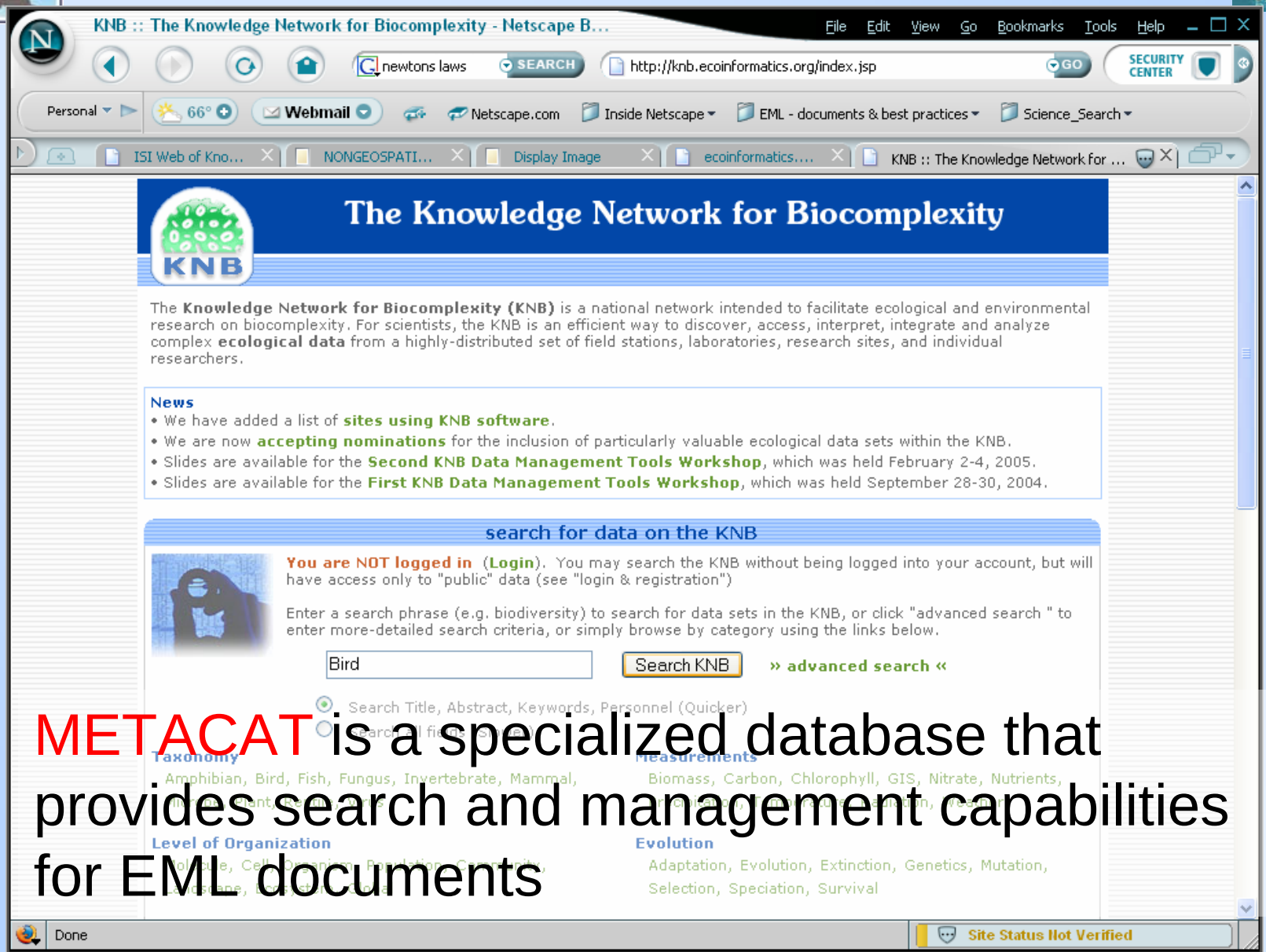
Specialized Tool - Morpho

Sample Wizard Pages





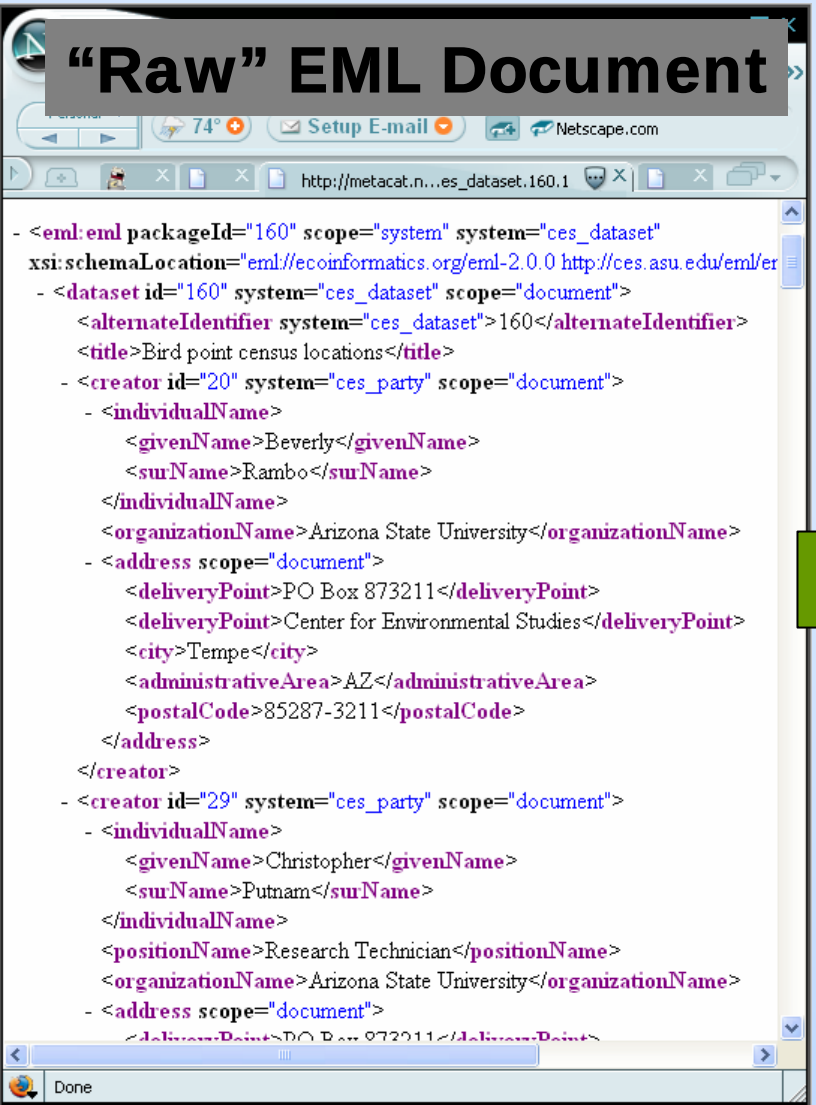
Using EML - Search



METACAT is a specialized database that provides search and management capabilities for EML documents

Using EML - Display

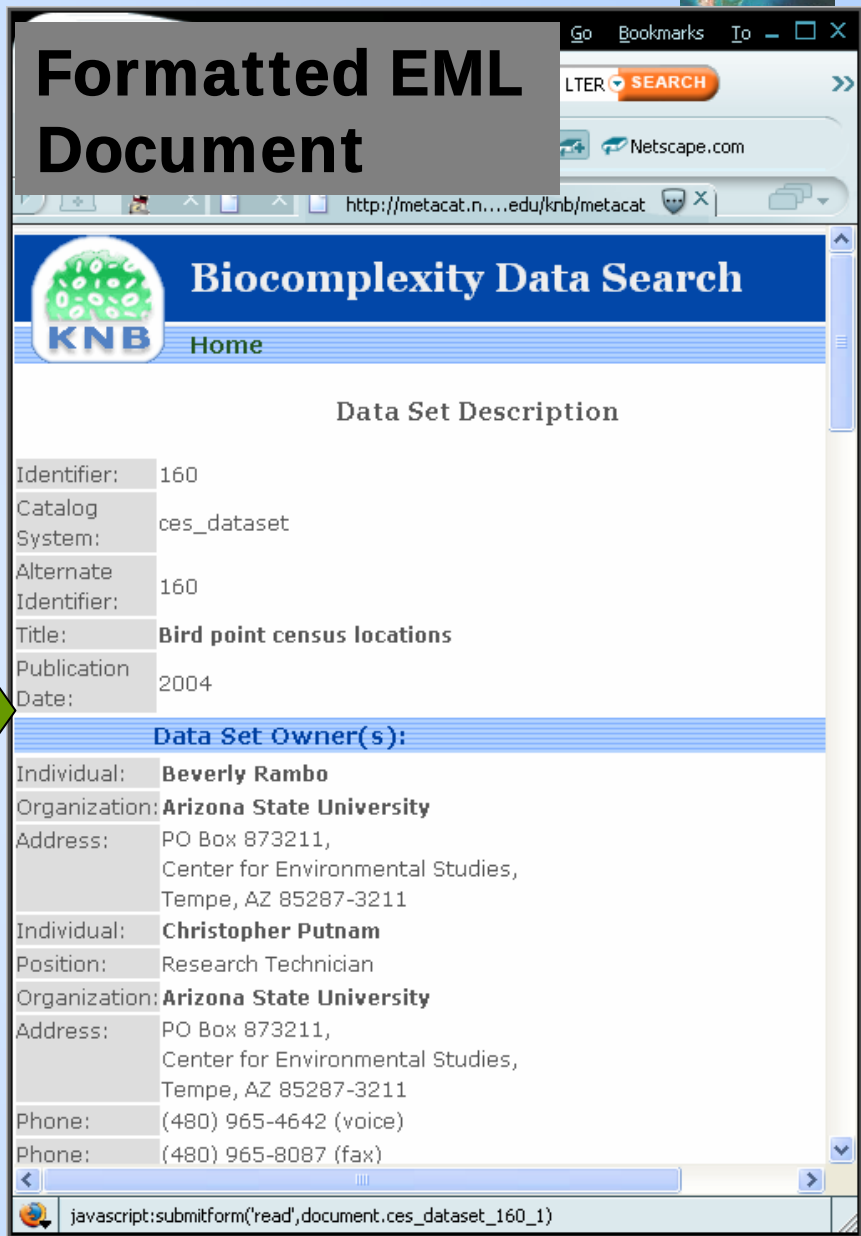
"Raw" EML Document



```
- <eml:eml packageId="160" scope="system" system="ces_dataset"
  xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.0 http://ces.asu.edu/eml/er
  - <dataset id="160" system="ces_dataset" scope="document">
      <alternateIdentifier system="ces_dataset">160</alternateIdentifier>
      <title>Bird point census locations</title>
    - <creator id="20" system="ces_party" scope="document">
      - <individualName>
          <givenName>Beverly</givenName>
          <surName>Rambo</surName>
        </individualName>
        <organizationName>Arizona State University</organizationName>
      - <address scope="document">
          <deliveryPoint>PO Box 873211</deliveryPoint>
          <deliveryPoint>Center for Environmental Studies</deliveryPoint>
          <city>Tempe</city>
          <administrativeArea>AZ</administrativeArea>
          <postalCode>85287-3211</postalCode>
        </address>
      </creator>
    - <creator id="29" system="ces_party" scope="document">
      - <individualName>
          <givenName>Christopher</givenName>
          <surName>Putnam</surName>
        </individualName>
        <positionName>Research Technician</positionName>
        <organizationName>Arizona State University</organizationName>
      - <address scope="document">
```

XSLT Translation

Formatted EML Document



Biocomplexity Data Search

KNB — Home

Data Set Description

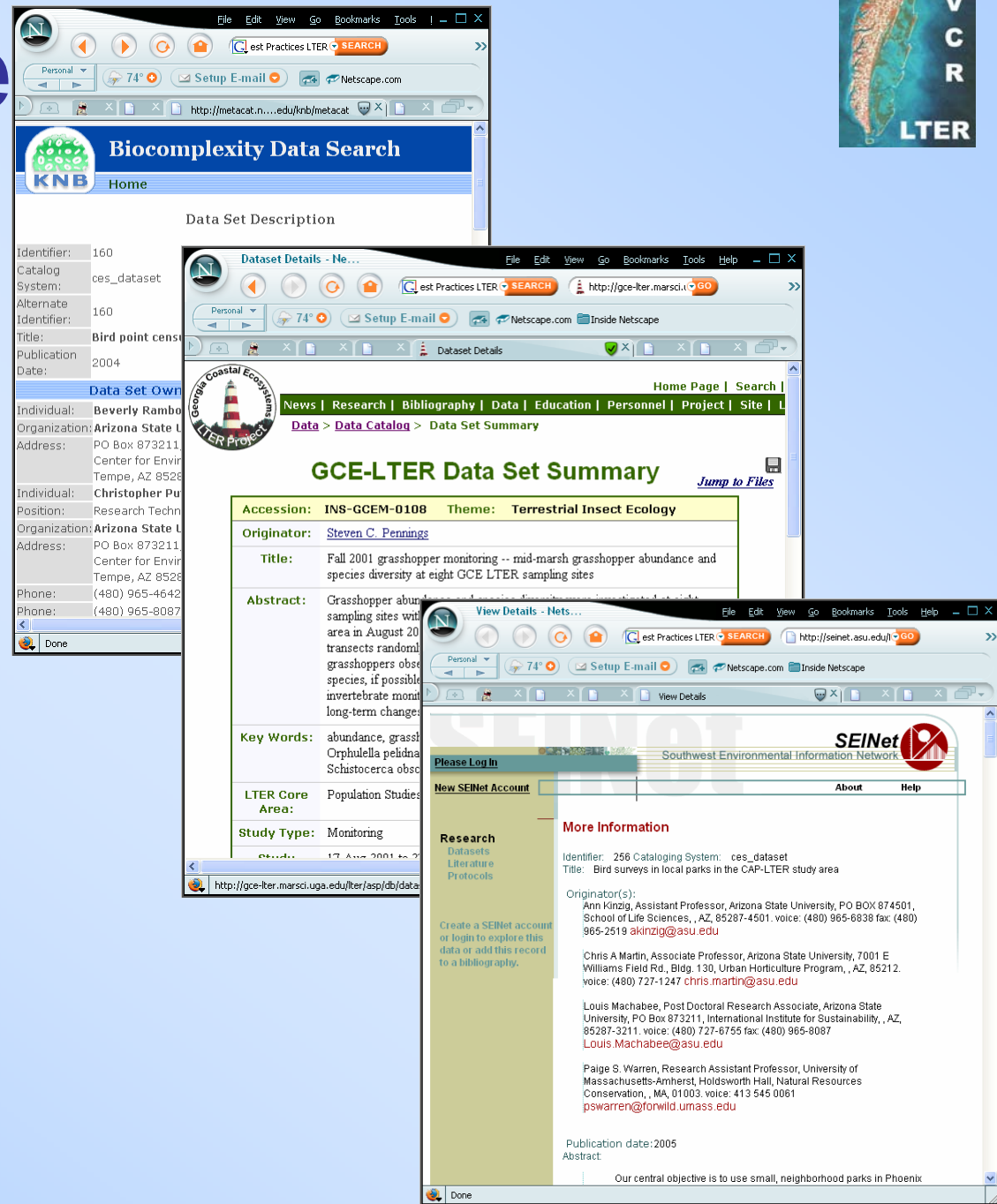
Identifier:	160
Catalog System:	ces_dataset
Alternate Identifier:	160
Title:	**Bird point census locations**
Publication Date:	2004

Data Set Owner(s):

Individual:	**Beverly Rambo**
Organization:	**Arizona State University**
Address:	PO Box 873211, Center for Environmental Studies, Tempe, AZ 85287-3211
Individual:	**Christopher Putnam**
Position:	Research Technician
Organization:	**Arizona State University**
Address:	PO Box 873211, Center for Environmental Studies, Tempe, AZ 85287-3211
Phone:	(480) 965-4642 (voice)
Phone:	(480) 965-8087 (fax)

Template

- Depending upon the XSLT template used, the same EML document can be displayed many different ways

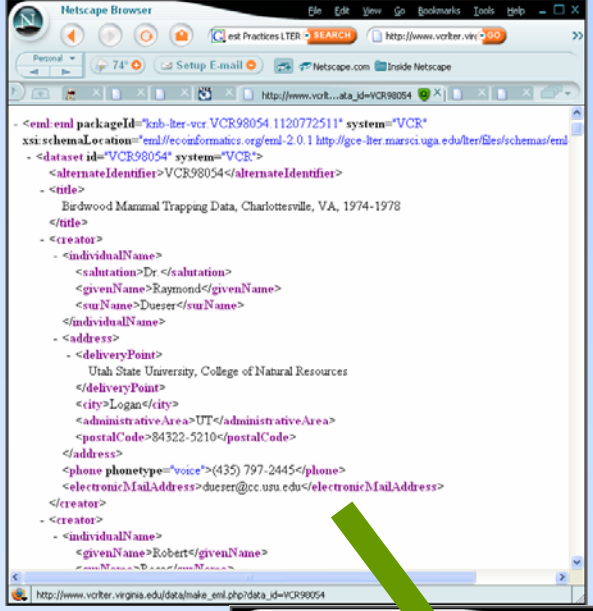


Automation

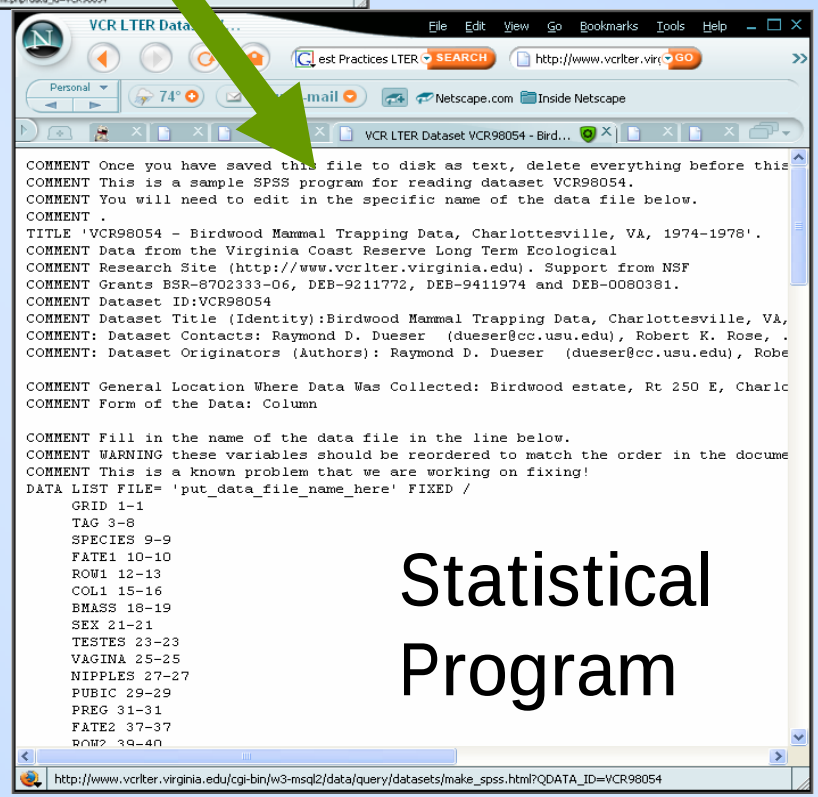
- For metadata to be used to automatically generate or configure analysis tools, it needs to contain the details required by those tools
  - Lists of attributes/variables
  - Information on storage or access formats
- Need to provide a reliable structure so that those details can be automatically extracted

Using EML

- EML transformations are not restricted to documents for viewing. They can also be used to create programs for processing the associated data



Raw EML Document



Statistical Program

Data Processing

- Researchers use a wide variety of tools to analyze data
  - Spreadsheets
  - Statistical Packages
- Many ecological datasets are available as text files
- Therefore, we need a way to facilitate the ingestion of data in text files into different tools used by researchers

We need to be able to translate an EML Document into a Statistical Program

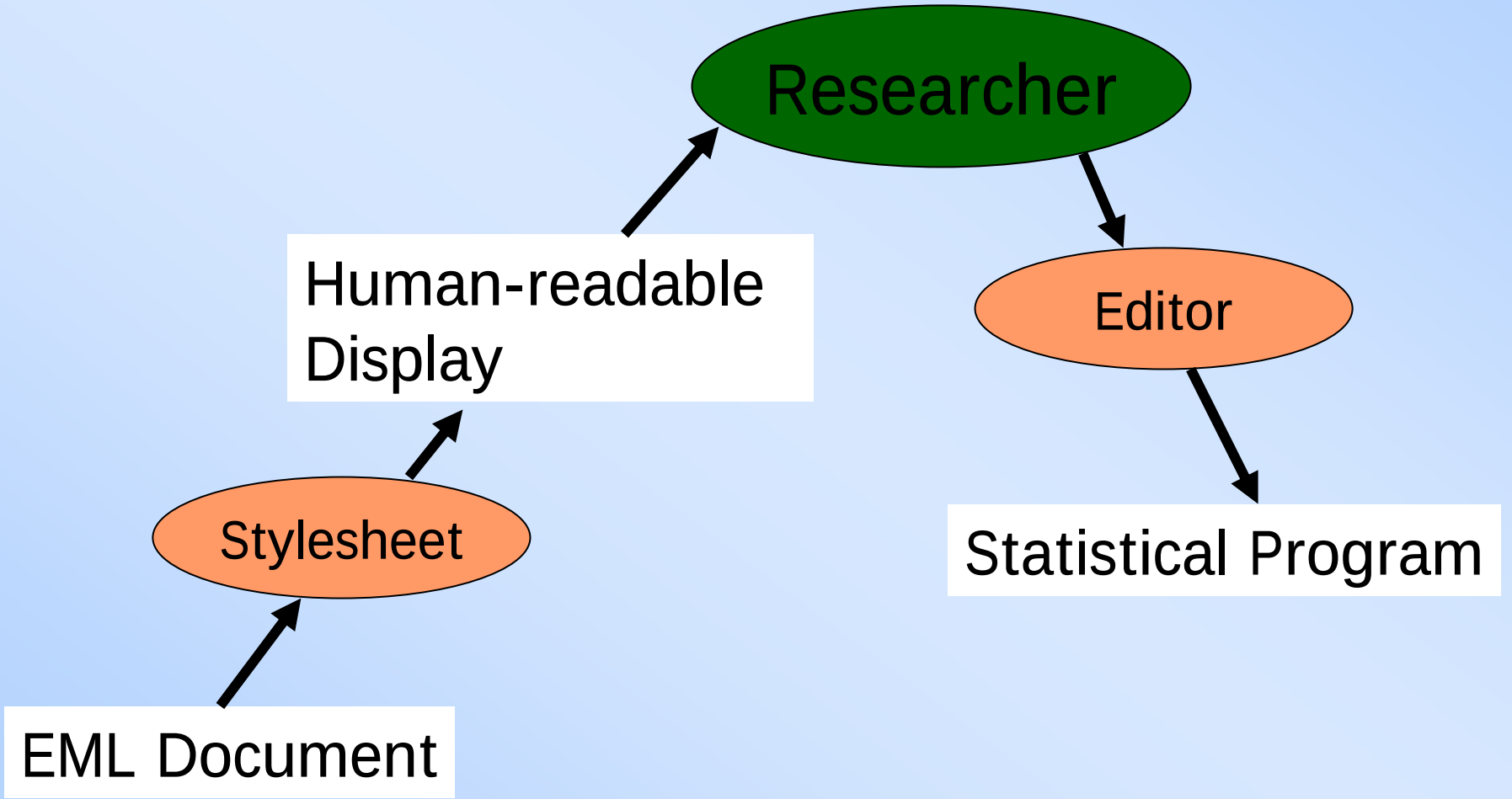
EML Document

```
<?xml version="1.0" encoding="UTF-8"?>
<eml:eml packageId="knb-lter-vcr.76.1" system="VCR"
    xmlns:ds="eml://ecoinformatics.org/dataset-2.0.1"
    xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
    xmlns:stmml="http://www.xml-cml.org/schema/stmml"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 http://gce-lter.marsci.uga.edu/lter/files/schemas/eml-201/eml.xsd">
<dataset id="76" system="VCR">
  <alternateIdentifier>VCR00073</alternateIdentifier>
  <title>Water Quality - physical data</title>
  <creator>
    <individualName>
      <salutation>Dr.</salutation>
      <givenName>Robert</givenName>
      <surName>Christian</surName>
    </individualName>
    <address>
      <deliveryPoint>East Carolina University, Department of Biology</deliveryPoint>
```

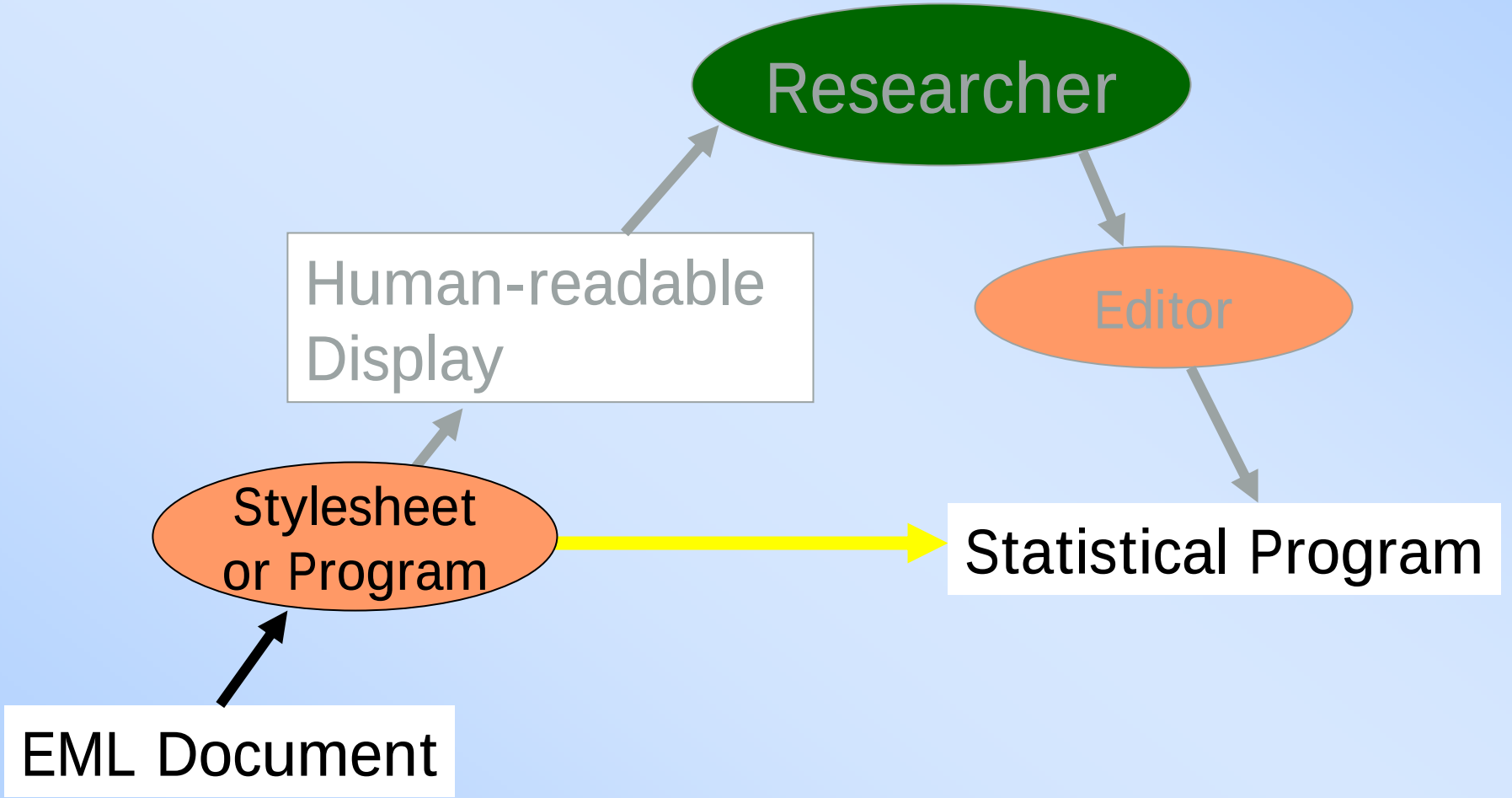
Statistical Program

```
GET DATA  /TYPE=TXT/
          / FILE="PUT-LOCAL-PATH-TO-DATA-FILE-HERE"
               /ARRANGEMENT=Delimited
                   /DELIMITERS=","
                       /QUALIFIER=""
              /VARIABLES=
              SITE A DATE A TIME F10.2 SCTTEMP F SCTSAL F
        SCTC0ND F10.2 TEMP F REFRSAL F DOTEMP F DO F SECCHI
        F DEPTH F WIND A .
execute.

VAR LABELS   SITE 'SAMPLE SITE OR STATION- ' .
VAR LABELS   DATE 'DATE SAMPLE COLLECTED- ' .
VAR LABELS   TIME 'TIME SAMPLE COLLECTED- none' .
VAR LABELS   SCTTEMP 'TEMPERATURE BY SCT- DEGREES C' .
VAR LABELS   SCTSAL 'SALINITY BY SCT- PPT' .
VAR LABELS   DOTEMP 'TEMPERATURE BY DISSOLVED OXYGEN METER-
        mg/L' .
VAR LABELS   DO 'DISSOLVED OXYGEN BY DISSOLVED OXYGEN METER-
        mg/L' .
VAR LABELS   SECCHI 'SECCHI DEPTH- cm' .
VAR LABELS   DEPTH 'WATER DEPTH- m' .
VAR LABELS   WIND 'WIND SPEED AND DIRECTION- ' .

Frequencies   variables=SITE /order=analysis.
Frequencies   variables=DATE /order=analysis.
Frequencies   variables=WIND /order=analysis.
Descriptives  variables=TIME .
Descriptives  variables=SCTTEMP .
```

The Old Way


Researcher
Human-readable Display
Editor
Stylesheet
Statistical Program
EML Document

A New Way


Researcher
Human-readable Display
Editor
Stylesheet or Program
Statistical Program
EML Document

EML structure and content allow it to be created and used in many different ways

XML Editor

or

Metadata Database

EML Document

Display Template

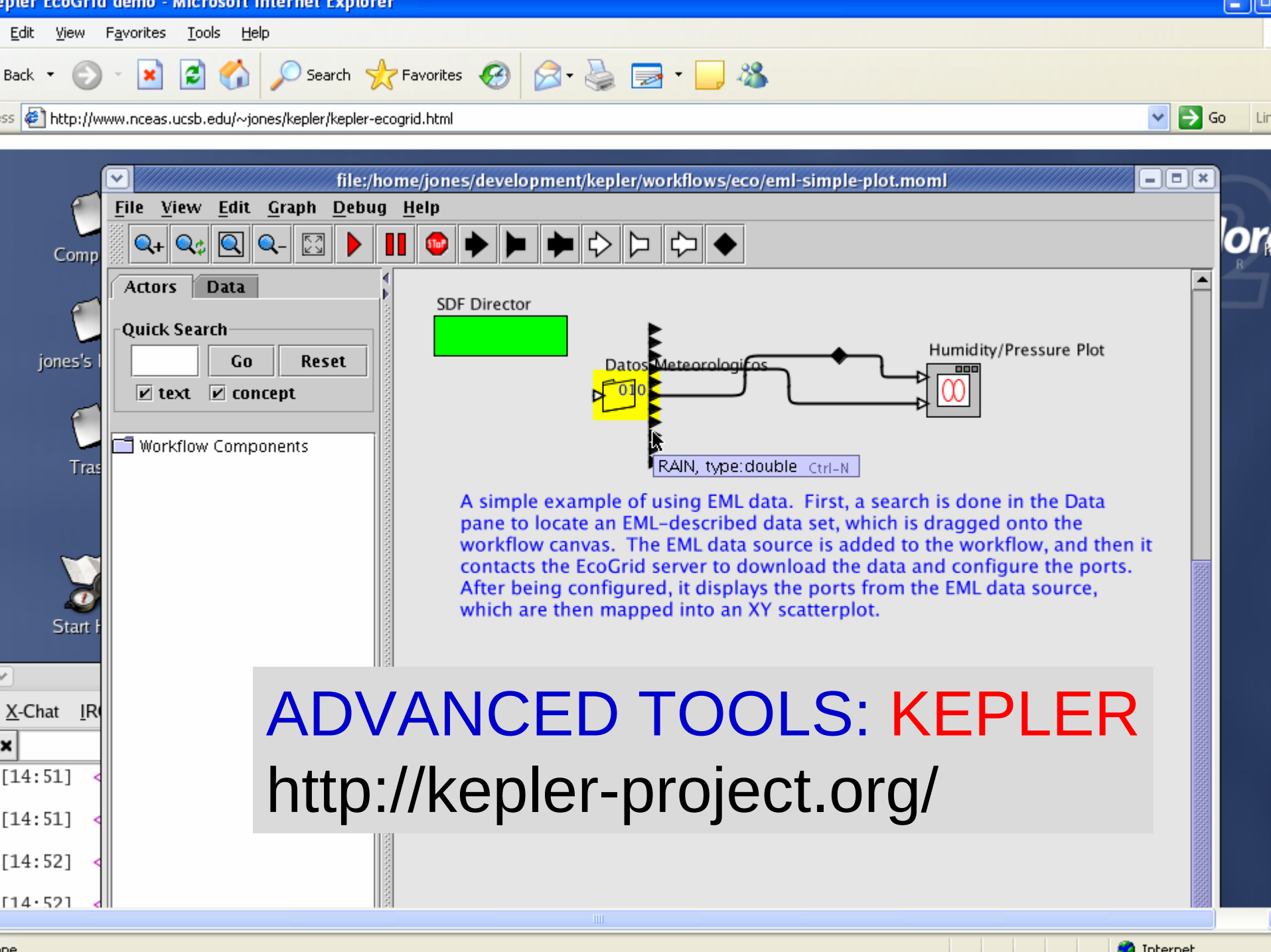
or

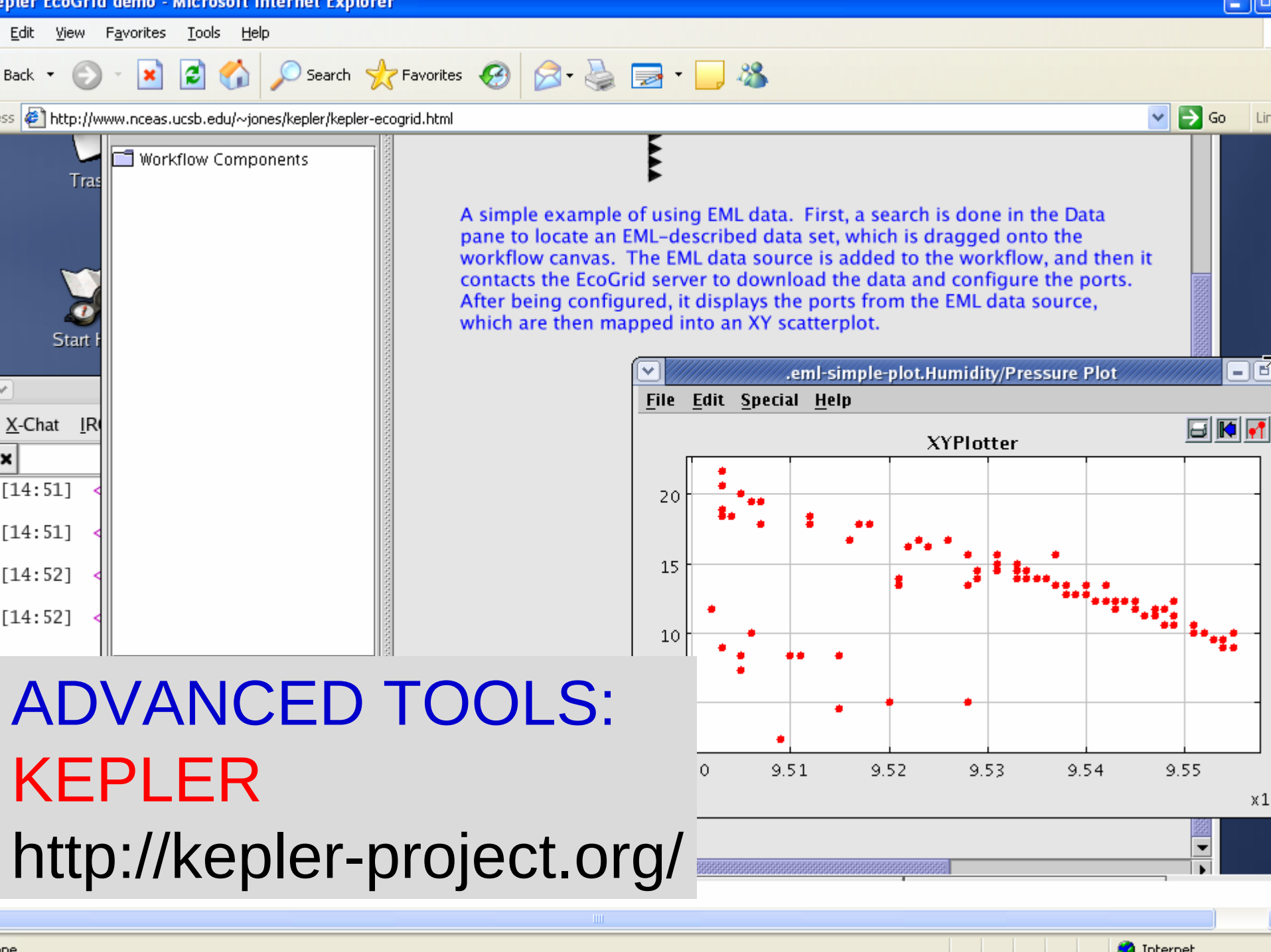
Program Template

or

Database Template

Database


Edit View Favorites Tools Help
Back Search Favorites
http://www.nceas.ucsb.edu/~jones/kepler/kepler-ecogrid.html
Go
file:/home/jones/development/kepler/workflows/eco/eml-simple-plot.moml
File View Edit Graph Debug Help
Actors Data
Quick Search
Go Reset
text concept
Workflow Components
SDF Director
Datos Meteorologicos
Humidity/Pressure Plot
RAIN, type:double Ctrl-N
A simple example of using EML data. First, a search is done in the Data pane to locate an EML-described data set, which is dragged onto the workflow canvas. The EML data source is added to the workflow, and then it contacts the EcoGrid server to download the data and configure the ports. After being configured, it displays the ports from the EML data source, which are then mapped into an XY scatterplot.
ADVANCED TOOLS: KEPLER
http://kepler-project.org/



ADVANCED TOOLS:

KEPLER

http://kepler-project.org/

Who is Using EML?

- US LTER Network
- Organization of Biological Field Stations (OBFS)
- Ecological Society of America, Long-Term Studies Section
- Cyberinfrastructure Research Projects
  - Knowledge Network for Biocomplexity (KNB)
  - Science Environment for Ecological Knowledge (SEEK)

Long-Term Studies Section

LTSS Data Registry

LTSS Home | Registry Home | Register a New Data Set | Search for Data

Data Registry Form

Use this form to submit a new data set description for inclusion in the

Please have a look at the **Guide for Completing the Data Registry Fo** e your browser's Reload/Refresh function to make sure you see the late

If you have any questions, comments or problems regarding this form, **esaadmin@nceas.ucsb.edu**.

***Denotes a required field.**

The **Long-Term Studies Section Data Registry** includes:

- Identification information
- Data set owner
- Abstract
- Keywords
- Coverage (spatial, temporal, taxonomic)
- Data Collection Methods
- Dataset Contact
- Distribution Information

BASIC INFORMATION (What's this?) ide

*First Name

*Last Name

*Data Set Title

*Organization Name

PRINCIPAL DATA SET OWNER (What's this?) Hide

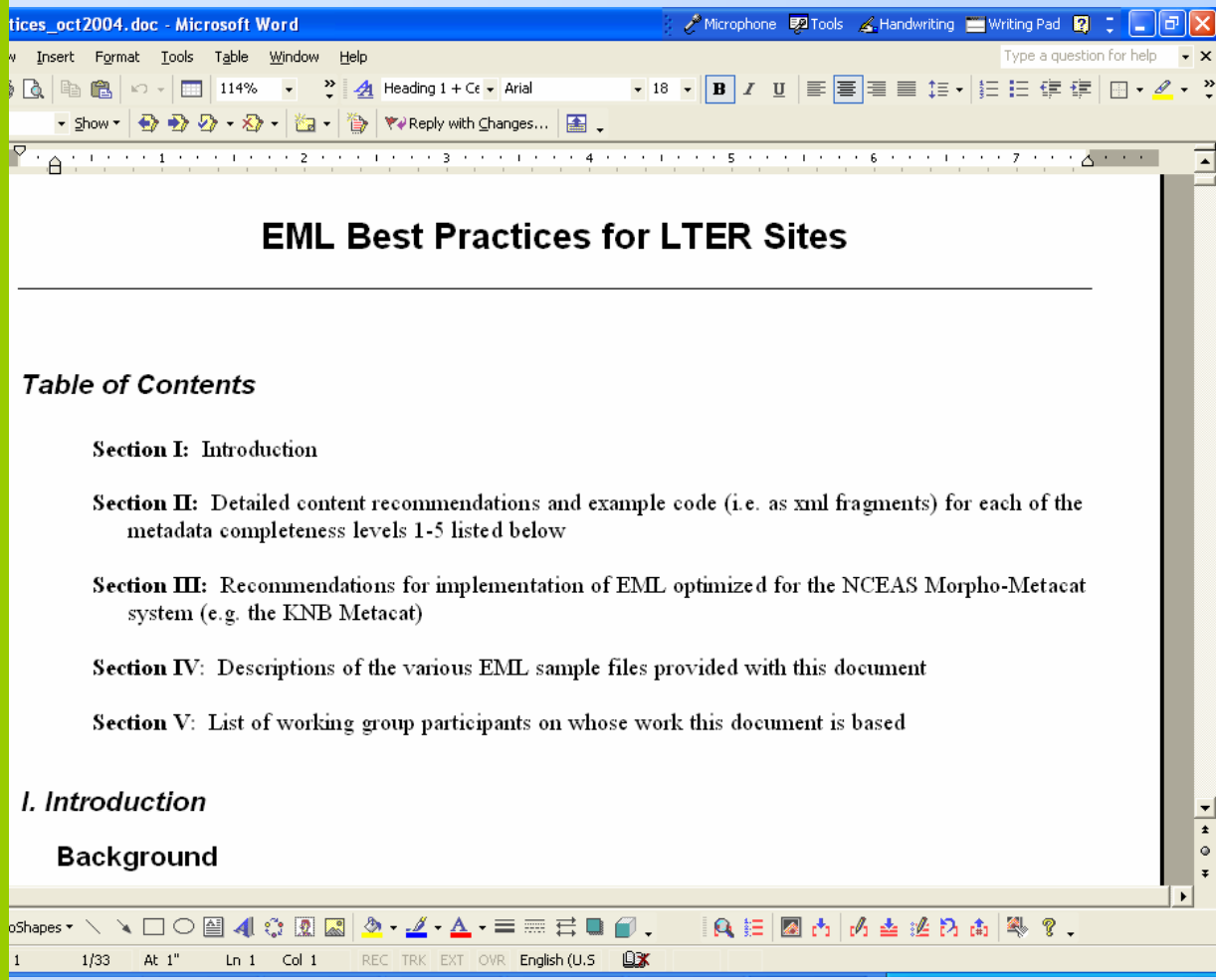
*First Name

Done

U.S. LTER

http://cvs.lternet.edu/cgi-bin/viewcvs.cgi/emlbestpractices/emlbestpractices-1.0/

- **The U.S. LTER sites have developed a "Best Practices" guide for US LTER sites**
  - **It calls for a tiered approach with sites moving from EML suitable for data identification to full-content EML metadata that supports interoperability**



Level 1 - Identification

- Minimum content for adequate data set discovery in a general cataloging system or repository
  - title
  - creator
  - contact
  - publisher
  - pubDate
  - keywords
  - abstract (recommended)
  - dataset/distribution (i.e. url for general dataset information)

Level 2 - Discovery

- Level 1 content, plus coverage information to support targeted searches, adding elements:
  - Geographic Coverage
  - Taxonomic Coverage
  - Temporal Coverage

Level 3 - Evaluation

- Level 2 content, plus data set details to enable end-user evaluation of the methodology and data entities, adding elements:
  - Intellectual Rights
  - project
  - methods
  - dataTable/entityGroup
  - dataTable/attributes

EML Best Practices for LTER Sites – Oct. 2004

Level 4 - Access

- Level 3 content plus data access details to support automated data retrieval, adding elements:
  - access
  - physical

Level 5 - Integration

- Level 4 content plus complete attribute and quality control details to support computer-assisted data integration and re-sampling, adding elements:
  - Attribute List (full descriptions)
  - Constraint
  - Quality Control

More Information on EML and EML-related Projects

http://www.ecoinformatics.org

PARTNERSHIP for BIODIVERSITY INFORMATICS

SEEK

NCEAS

SDSC

LTER
LONG TERM ECOLOGICAL RESEARCH

NATURAL HISTORY MUSEUM
BIODIVERSITY • RESEARCH • CENTER

OBFS