



**ARCHIVE AND RETRIEVAL OF EPA STAR GRANT  
DATA AND METADATA:  
A PROTOTYPE DEVELOPED BY THE ESTUARINE  
AND GREAT LAKES STAR COASTAL INDICATORS  
RESEARCH CENTERS AND EPA EIMS**

Valerie Brady, Terry Brown, Barbara Levinson,  
Susan Eversole, Derek Lane, Gkay Bishop,  
John Sykes



**OR**

**The need to publicly archive EPA grant-  
funded research data**

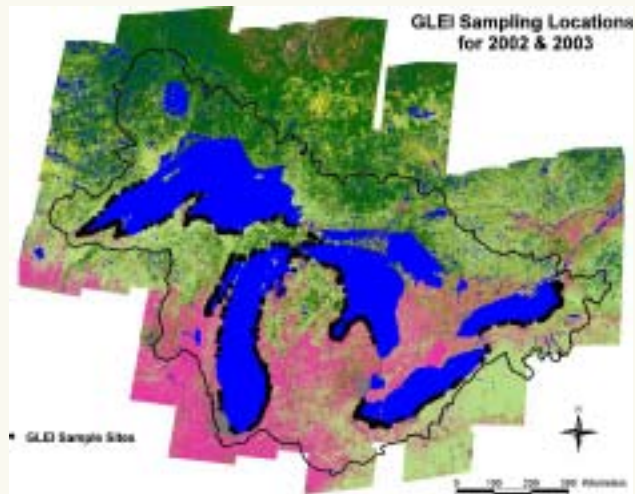
Valerie Brady, Terry Brown, Barbara Levinson,  
Susan Eversole, Derek Lane, Gkay Bishop,  
John Sykes



## Estuarine and Great Lakes Coastal Indicators Program



## Great Lakes Environmental Indicators Project (GLEI)



## Why are we the guinea pigs?



## EaGLE Data Committee Mission Statement



- ▶ Develop an information management plan to archive EaGLE data with appropriate metadata so that EPA can make it readily available
- ▶ Ensure that data usefulness outlives the EaGLE project (and does not require continued maintenance by EaGLE researchers)

## EIMS overview



## EaGLE Data Types



- ▶ Geospatial & Imagery
- ▶ Genomic
- ▶ Remote Sensing
- ▶ Biological
- ▶ Routine Monitoring

## What data must be archived?



All new data created or collected using EaGLE funds

- ▶ Field data
- ▶ Genomics experiments
- ▶ New GIS coverages
- ▶ New remote sensing data
- ▶ Other images, models

All important summary, supplemental, and explanatory information

- ▶ Journal articles
- ▶ Poster Sessions
- ▶ Presentations
- ▶ Rules governing data QC or transforms
- ▶ SOPs, protocols, experimental design documents, QA/QC documents

## Data File Formats:



### Acceptable

- ▶ Files converted into character delimited ASCII files (i.e., comma delimited .csv files)
- ▶ jpeg, jpg, tiff, gif, img, png, geo-tiff, ecw, ArcView, simple html or htm, xml, LaTeX, TeX, pdf (method files)
- ▶ Programs in programming language (must have text support).

### Unacceptable

- ▶ Excel Spreadsheets (convert to .csv)
- ▶ Presentation files such as PowerPoint (convert to .pdf)
- ▶ Word-processing files (convert to ASCII)
- ▶ Proprietary files
- ▶ RTF files (convert to ASCII)
- ▶ Special characters (Greek letters and other symbols not found in ASCII)



## EML: Standard for Ecological Metadata

- ▶ Core: Definitions and units of the columns (fields or attributes) in all data tables
- ▶ Methods, procedures, and protocols
- ▶ Research questions and hypotheses
- ▶ Site selection
- ▶ Authors, contacts, and proper citation for use
- ▶ Sampling Extent: spatial, biological, & temporal



## Metadata Creation / Data Uploading



### ▶ Metadata Entry Form (MEF)

- ▶ Generates an EML-compliant metadata file in XML format

### ▶ Automatic upload to ERSL

- ▶ Data packages stored in EIMS repository (ERSL backend)

### ▶ EaGLE Portal—intranet interface for grantees

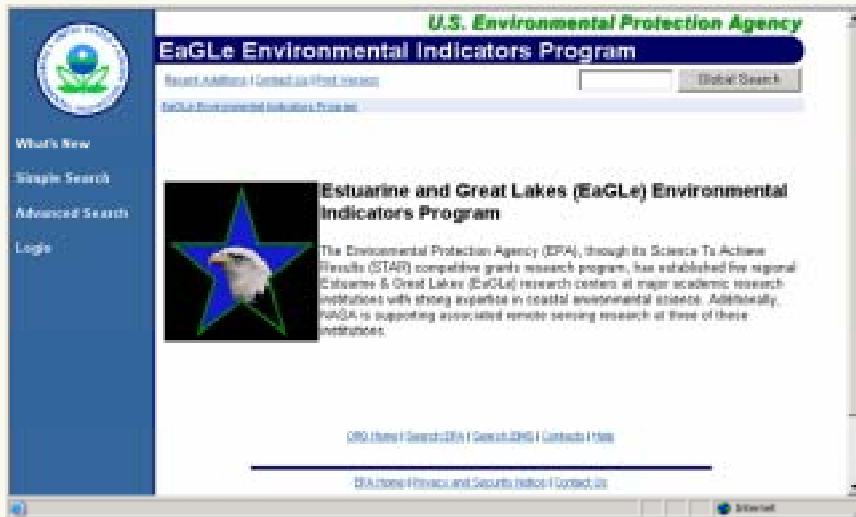
- ▶ Review, Approval, and Release Processes

### ▶ Post-Release: Search, Store and Update

- ▶ Searchable Metadata Records in one area of EIMS/ERSL
- ▶ Actual Datasets stored in EIMS/ERSL Repository



## EaGLE Prototype Home Page



[canned website](#)

## Selling the data archiving vision



Or

How to prevent researchers and technicians from running screaming from the room once they understand what it means to write good metadata.



## Why archive the EaGLE data?



- ▶ **To ensure its preservation for future generations of scientists (EaGLE legacy)**
- ▶ **To ensure it is broadly available for current scientists to use**
- ▶ **To create the broadest possible public benefit from this taxpayer-funded program**
- ▶ **To help EPA retain the data that is collected / created through its funding**
- ▶ **Because we wish that earlier researchers had archived their data for us to use**

## Data Access and Security



- ▶ **Only registered users may enter or edit a metadata record**
  - ▶ Record-level edit permissions required for input and update
- ▶ **Only registered Data Librarians can release records to a designated user base (Public, EPA Only, Group, Owner)**
- ▶ **Confidential records can be restricted to a subset of users**
  - ▶ EPA Only – accessible only to EPA registered users
  - ▶ Group – accessible only to members of a specified group of users (including system users outside the EPA firewall, if necessary)
  - ▶ Owner – accessible only by the designated owner of the EIMS record
- ▶ **Post-release: any internet user may view metadata records.**
- ▶ **Separate access controls for actual datasets**



## What Good are Metadata?

High quality metadata serve 5 purposes:

### Data Integrity Maintenance over the long term: 20-year rule

- Across expected changes in data storage technology, compression, etc.

### Tracking, searching for, and retrieving datasets

- Like a library card catalogue—where to find data, where to shelve it.

### Scientific collaboration

- Joint analysis and secondary analysis potential

### Cathedral effect

- Pooling data across regions contributes to an environmental “big picture”
- Longitudinal studies--building science efforts upon a shared data foundation.

### Economical

- Extending the shelf life of data gives taxpayers more return on investment



## How much does it cost to collect metadata?

- ▶ **Estimate the value of your research results**
  - ▶ Total amount of research grant(s) plus 15% added value
- ▶ **Divide by number of years project is funded**
- ▶ **Allocate 10% of resulting \$/efforts to metadata collection**
- ▶ **Distribute amounts evenly over years—don't stint!**
  - ▶ Collecting metadata at the beginning of a study captures important data decisions and research design elements
- ▶ **Use metadata collection as an *ad hoc* method of data quality control during each year of the study.**



## How much time is this going to take?

- ▶ **Between 8 and 40 hours per data group**
  - ▶ All similar data bundled together—not a per dataset cost!
  - ▶ More complex datasets take more time
  - ▶ Loading or linking to pre-written material can save time
- ▶ **Training for use of Metadata Entry Form**
  - ▶ One-time 3-hour training session
  - ▶ Minimum 3 hours hands-on practice
  - ▶ Availability of live “help” during first solo MEF work



## Getting in Gear:

- ▶ **Feb. 1, 2004: Begin metadata creation.**
- ▶ **Summer 2004: Begin EaGLE data uploading.**
- ▶ **Jan. 2005: EaGLE metadata completed.**
- ▶ **End of no-cost extensions (early 2006): Most of EaGLE datasets archived but password-protected.**
- ▶ **Jan. 2008: Most of EaGLE data released to public**

## Generations of Research



For a true confluence of research efforts, clarity in metadata is the key

## M) EaGLE Prototype Global Search



**U.S. Environmental Protection Agency**  
**EaGLE Environmental Indicators Program**

Research Address | Contact Us | What's New

Search:  All Data

[EaGLE Environmental Indicators Program](#)

**Estuarine and Great Lakes (Ea) Indicators Program**

The Environmental Protection Agency (EPA), through its Science To Achieve Results (STAR) competitive grant research program, has established five regional Estuarine & Great Lakes (EaGLE) research centers at major academic research institutions with strong expertise in coastal environmental science. Additionally, EPA is supporting associated interdisciplinary research at three of these institutions.

[EaGLE Home](#) | [Search EIS](#) | [Search EMS](#) | [Contact Us](#) | [Help](#)

[EPA Home](#) | [Privacy and Security Notice](#) | [Contact Us](#)



## N) EaGLE Prototype Search Results

**U.S. Environmental Protection Agency**  
**EaGLE Environmental Indicators Program**

Search Address / Control List / Print Screen  Global Search

U.S. Environmental Indicators Program > Search Results

### Search Results

# Entry ID Classes 2 Records Found

1) [ent.1.1] [ATLANTIC COAST ENVIRONMENTAL INDICATORS CONSORTIUM STUDY AREA BOUNDARY, November 17, 2000](#)  
 The ACE study area boundary is the outer boundary of drainage basins contained in the Atlantic Slope region.

2) [ent.1.2] [Final Year - ATLANTIC COAST ENVIRONMENTAL INDICATORS CONSORTIUM LACE INCLUDE NEW RIVER WATER QUALITY, November 17, 2000](#)  
 Biological, chemical and physical water quality data was collected from the Newse River Estuary, NC during the project funding period (February 28, 2001 through February 26, 2006) as part of an ongoing collaborative monitoring effort (1994 - Present).

0/0 Home | Search EPA | Search EMS

[Site Home](#) | [Contact and Security Notice](#) | [Contact Us](#)

Last Updated on Monday, November 24, 2008

## O) EaGLE Metadata Report

**U.S. Environmental Protection Agency**  
**EaGLE Environmental Indicators Program**

Search Address / Control List / Print Screen  Global Search

U.S. Environmental Indicators Program > Metadata Report

### Metadata Report

Entry ID: ent.1.2  
 Title: Atlantic Coast Environmental Indicators Consortium (ACE) - New River Water Quality  
 Version: 1.0

**Abstract** ← **Header**

**Abstract:** Final year - biological water quality data was collected from the Newse River Estuary, NC during the project funding period (February 28, 2001 through February 26, 2006) as part of an ongoing collaborative monitoring effort (1994 - Present) to characterize the environmental conditions of the Newse River. Biweekly water sampling and in situ measurements were performed at fixed sampling stations. Biological, chemical and physical parameters measured include: Temperature, salinity, specific conductivity, dissolved oxygen, pH, in situ fluorescence, turbidity, bacterioplankton, dissolved organic and inorganic carbon, dissolved inorganic nutrients (nitrate/nitrite, ammonium, phosphate, and silicate acid), total dissolved nitrogen, particulate nitrogen and carbon, Chlorophyll a, and phytoplankton pigment concentrations. Calculated parameters: Secchi (calculated from phytoplankton specific absorption [PSA] from [C]M), dissolved inorganic nitrogen (nitrate/nitrite plus urea) [total dissolved nitrogen minus dissolved inorganic nitrogen], algae, cryptophytes, cyanobacteria, diatoms and dinoflagellates. Benthic pigments using ChemTax, the statistical program that calculates relative and absolute concentrations of these commonly observed phytoplankton classes in the Newse River.

**Keywords:** Biological, chemical and physical water quality data was collected from the Newse River Estuary, NC during the project funding period (February 28, 2001 through February 26, 2006) as part of an ongoing collaborative monitoring effort (1994 - Present) to characterize the environmental conditions of the Newse River. Biweekly water sampling and in situ measurements were performed at fixed sampling stations. Biological, chemical and physical parameters measured include: Temperature, salinity, specific conductivity, dissolved oxygen, pH, in situ fluorescence, turbidity, bacterioplankton, dissolved organic and inorganic carbon, dissolved inorganic nutrients (nitrate/nitrite, ammonium, phosphate, and silicate acid), total dissolved nitrogen, particulate nitrogen and carbon, Chlorophyll a, and phytoplankton pigment concentrations. Calculated parameters: Secchi (calculated from phytoplankton specific absorption [PSA] from [C]M), dissolved inorganic nitrogen (nitrate/nitrite plus urea) [total dissolved nitrogen minus dissolved inorganic nitrogen], algae, cryptophytes, cyanobacteria, diatoms and dinoflagellates. Benthic pigments using ChemTax, the statistical program that calculates relative and absolute concentrations of these commonly observed phytoplankton classes in the Newse River.

**Administrative Details** ← **Link to top of page**



## P) EaGLE Metadata Report *(continued)*

The screenshot displays the 'Administration' section of the EaGLE Metadata Report. The left sidebar contains navigation links for Administration, Contacts, Access Information, and Data Elements. The main content area shows the following details:

- State:** REVIEWED
- IMS Partner:** EaGLE Star Credits
- Collection:** GENERAL
- Contacts:**
  - Name:** Dr. Hans Paerl
  - Role:** AUTHOR
  - Primary Phone:** 262-736-8841
  - Primary Email:** hans\_paerl@umw.edu
- Access Information:**
  - Availability:** PUBLIC
- Data Elements:**
  - Date Element:** Date
    - Description:** The date of bi-weekly water sample collection, filtering, and in situ measurements.
  - Date Element:** Station
    - Description:** The name of the fixed sampling station. (Station names increase in number from 0 (the most upstream station) to 100 (the most downstream station at the mouth of the Housa River).)



## Q) EaGLE Metadata Report *(continued)*

The screenshot displays the 'Data Elements' section of the EaGLE Metadata Report. The left sidebar contains navigation links for Administration, Contacts, Access Information, and Data Elements. The main content area shows the following details:

- Date Element:** Date
  - Description:** The date of bi-weekly water sample collection, filtering, and in situ measurements.
- Date Element:** Station
  - Description:** The name of the fixed sampling station. (Station names increase in number from 0 (the most upstream station) to 100 (the most downstream station at the mouth of the Housa River).)
- Date Element:** Distance (km)
  - Description:** The distance of the sampling station from Steeds Ferry Bridge (Station 0).
- Date Element:** Lat (lat)
  - Description:** The latitude coordinate of the sampling station.
- Date Element:** Lon (lon)
  - Description:** The longitude coordinate of the sampling station.
- Date Element:** Depth (m) (S or B)
  - Description:** The depth level from which the water sample was collected and from which the in situ measurements were made. S refers to a surface sample or in situ measurement taken at a water depth of approximately 0.2 meters. B refers to a bottom sample or in situ measurement taken at a water depth of approximately 0.5 meters from the sediment layer.
- Date Element:** YSI Time (minutes)
  - Description:** The exact time when the in situ measurements were made using the YSI 6600 sonde. These times approximate water sampling time.



## R) EaGLE Metadata Report *(continued)*

**Description:** Parameters, and Classification (all in American L.) as generated from 'Chem' vs. 'Should equal' total data generated from FPLC analysis.

**Data Element**  
**Description:** Comments for Parameter.

**Data Element**  
**Description:** Lists the parameter for which comments are included for values in that row. If there are comments for more than one parameter, there will be multiple columns of this column type.

**Data Element**  
**Description:** Comments.

**Data Element**  
**Description:** This column includes comments on the specific parameter value indicated on the previous column.

**Geographic Area** | 0 |

**Keywords** | Null |

**Methods** | 0 |

**Objectives** | 0 |

**References** | 0 |

**Time Frame** | 1 |

**Sampling Start Date:** 2001-02-28

**Sampling End Date:** 2001-02-28

[Data Element](#) | [Description](#) | [Search](#) | [Help](#) | [Feedback](#) | [Home](#)

[Data Element](#) | [Description](#) | [Search](#) | [Help](#) | [Feedback](#) | [Home](#)



## S) EaGLE Prototype Simple Search

**U.S. Environmental Protection Agency**  
**EaGLE Environmental Indicators Program**

Search |  |

[Search](#) | [Advanced Search](#) | [Simple Search](#)

### Simple Search

Restrict matches by title, author, keyword, abstract or project. Searches are case-insensitive. Searches in "All" match any of the fields.

Title:

Author:

Abstract:

Keywords:

Project:

OR  All

results to  records per page

**...and click Search**

**Enter selection criteria...**

## T) EaGLE Prototype Advanced Search



U.S. Environmental Protection Agency  
EaGLE Environmental Indicators Program

Home | Address | Contact Us | About Us | Global Search

U.S. Environmental Indicators Program > Advanced Search

### Advanced Search

What's New  
Simple Search  
Advanced Search  
Login

Title:

Author:

Abstract:

Keywords:

Project:

Geographic Extent Description:

Geographic Extent Boundings:  W  N  E  S

Done Internet Explorer

## U) EaGLE Prototype Advanced Search *(continued)*



Geographic Extent Description:

Geographic Extent Boundings:  W  N  E  S

Funding Source:

Temporal Coverage: Start Date  End Date  Enter dates in YYYY-MM-DD format

Methods:

Responsible Organization:

OR All

Limit results to  records per page

Search Reset

Done Internet Explorer

[Go Back](#)