Cornell University
Albert R. Mann Library

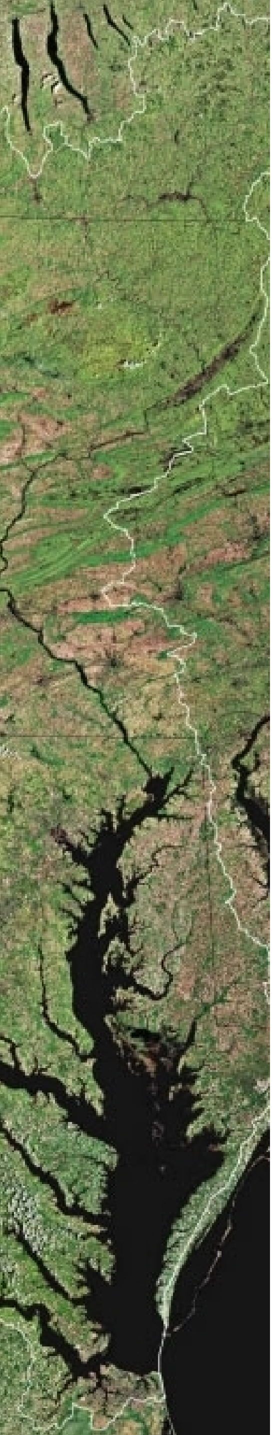
Data Curation and Distribution in Support of Cornell University’s Agricultural Ecosystems Program

Gail Steinhart
Research Data & Environmental Sciences Librarian

Brian Lowe
Metadata Programmer

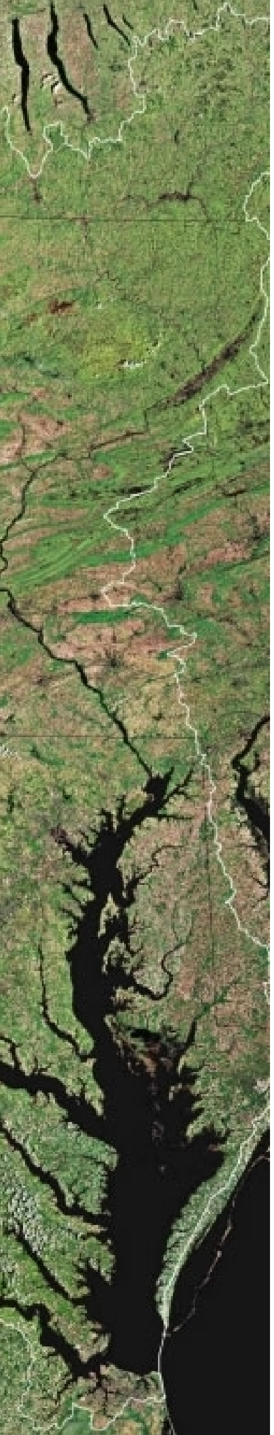
Albert R. Mann Library, Cornell University

DigCCurr - April 19, 2007

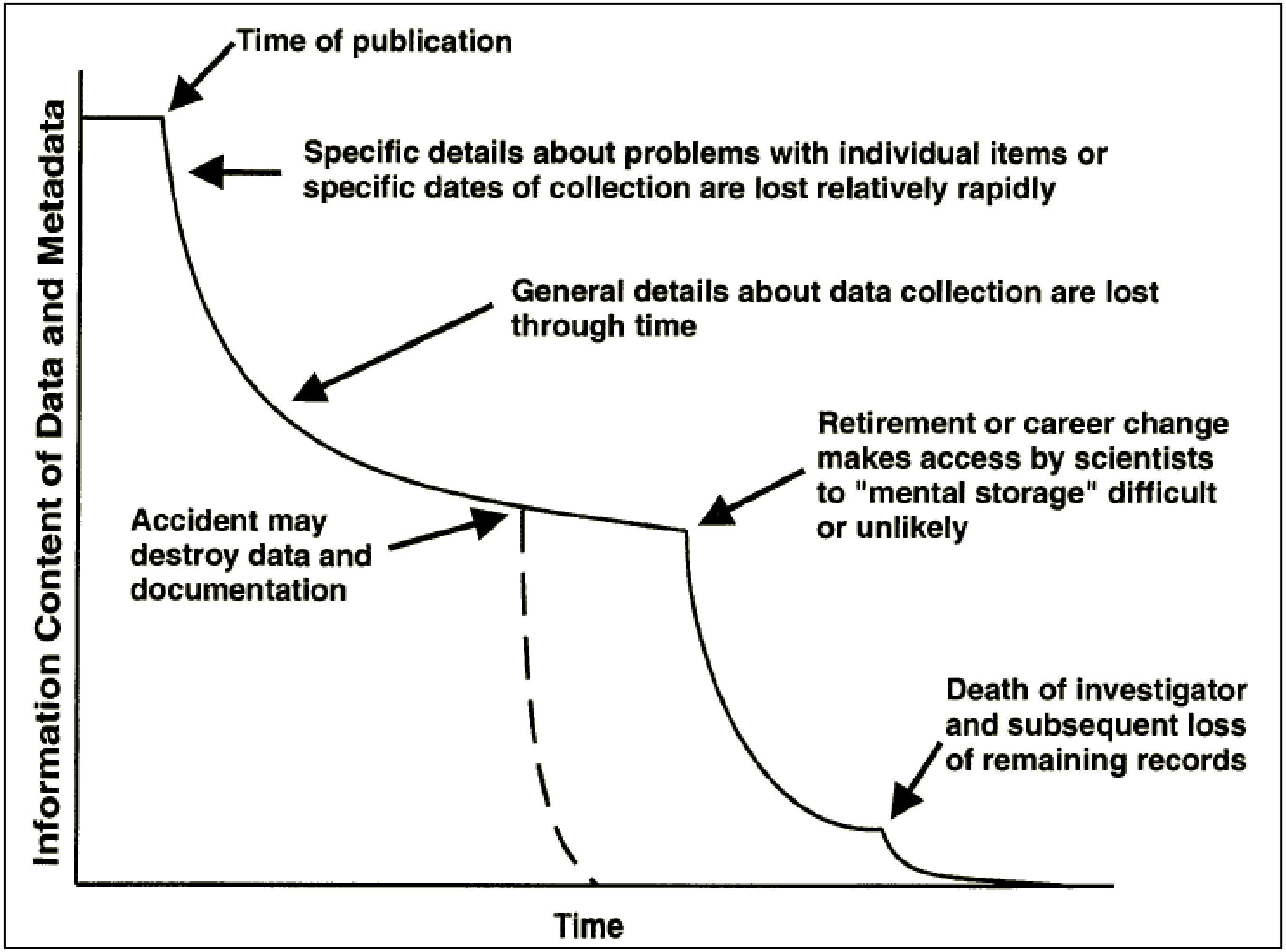


Overview

- Motivation
- Strategy
- What we’ve learned so far

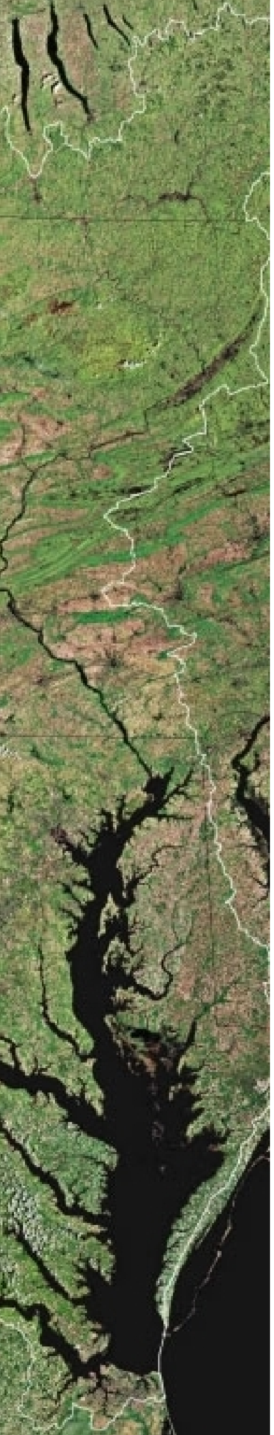


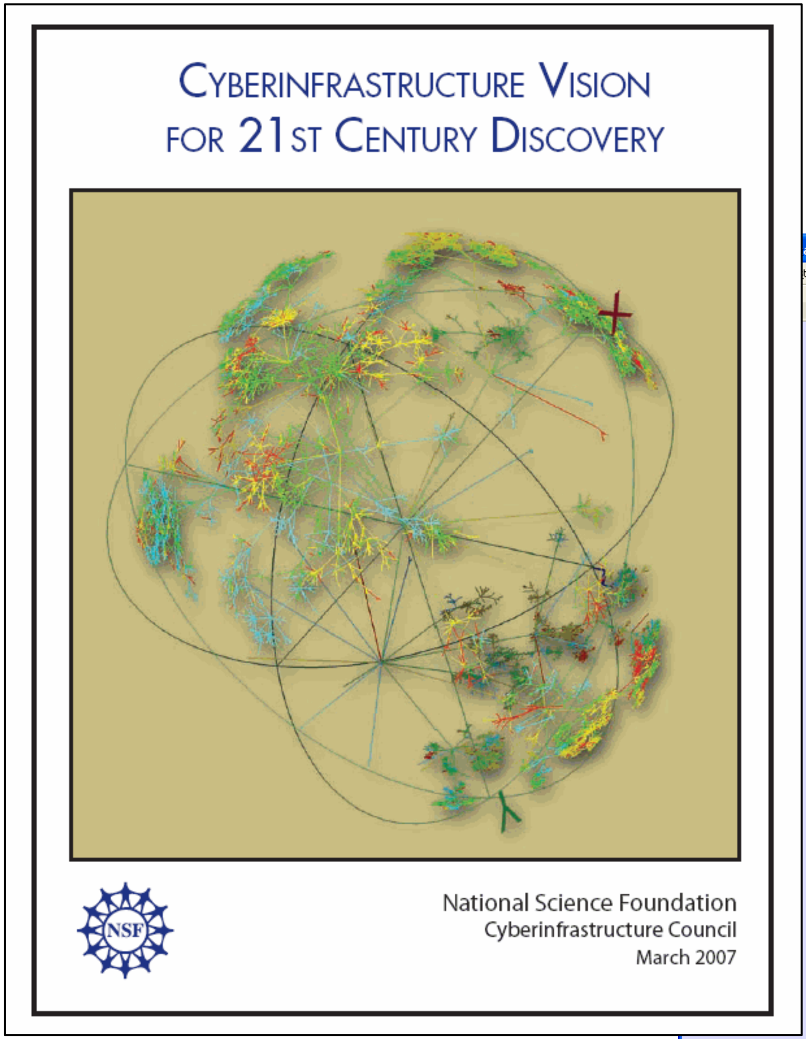
Motivation



from Michener et al., 1997

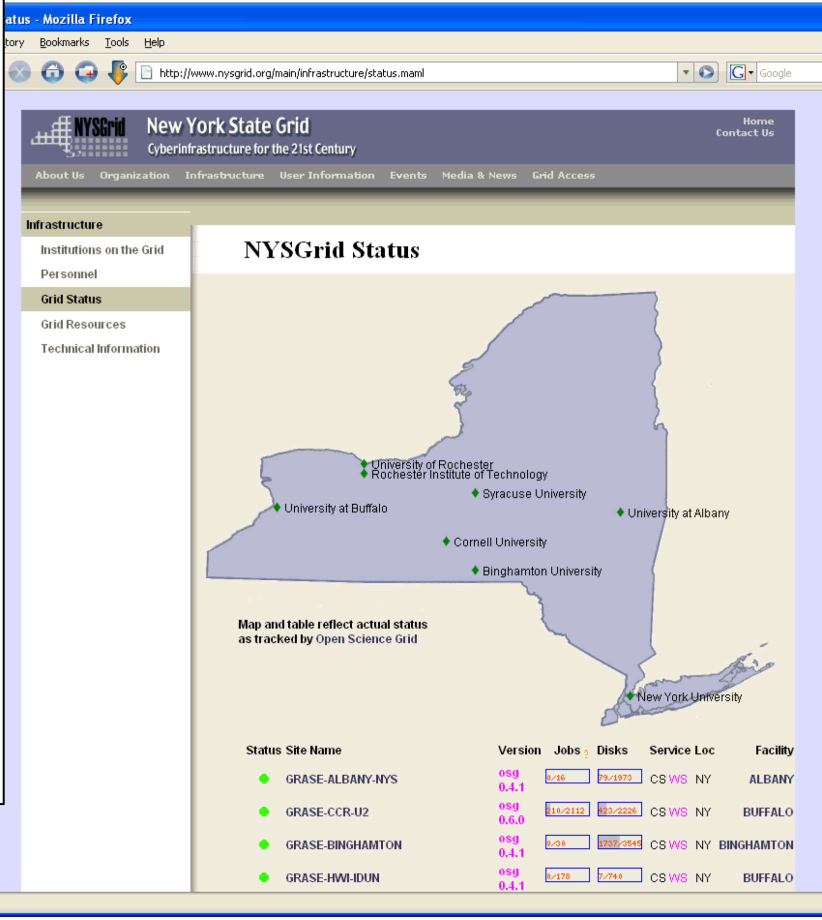
Motivation





CYBERINFRASTRUCTURE VISION FOR 21ST CENTURY DISCOVERY

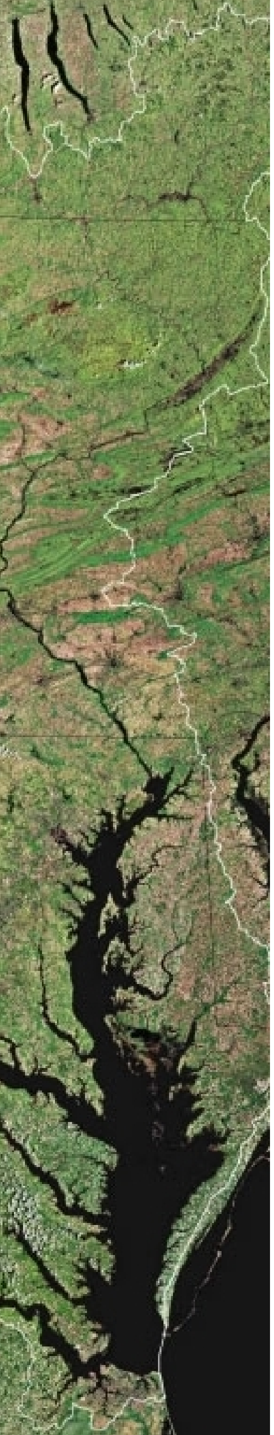
National Science Foundation
Cyberinfrastructure Council
March 2007



NYSGrid Status

Map and table reflect actual status as tracked by Open Science Grid

Status	Site Name	Version	Jobs	Disks	Service	Loc	Facility
●	GRASE-ALBANY-NYS	osg 0.4.1	0/16	79/1973	CS WS	NY	ALBANY
●	GRASE-CCR-U2	osg 0.6.0	210/2112	423/2226	CS WS	NY	BUFFALO
●	GRASE-BINGHAMTON	osg 0.4.1	0/30	1737/3545	CS WS	NY	BINGHAMTON
●	GRASE-HWI-IDUN	osg 0.4.1	0/178	7/740	CS WS	NY	BUFFALO



Definitions

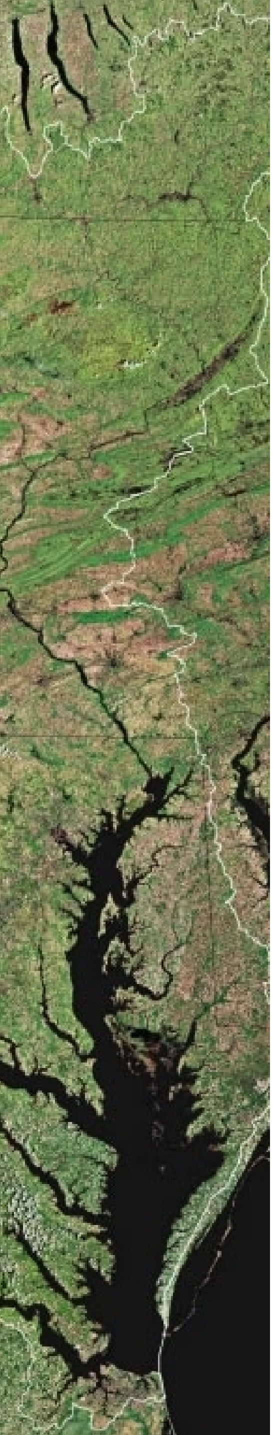
Curation

The term digital curation is used … for the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users. Implicit in this definition are the processes of digital archiving and preservation but it *also includes all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge.*

- from DCC

We

Academic and research libraries, but not alone



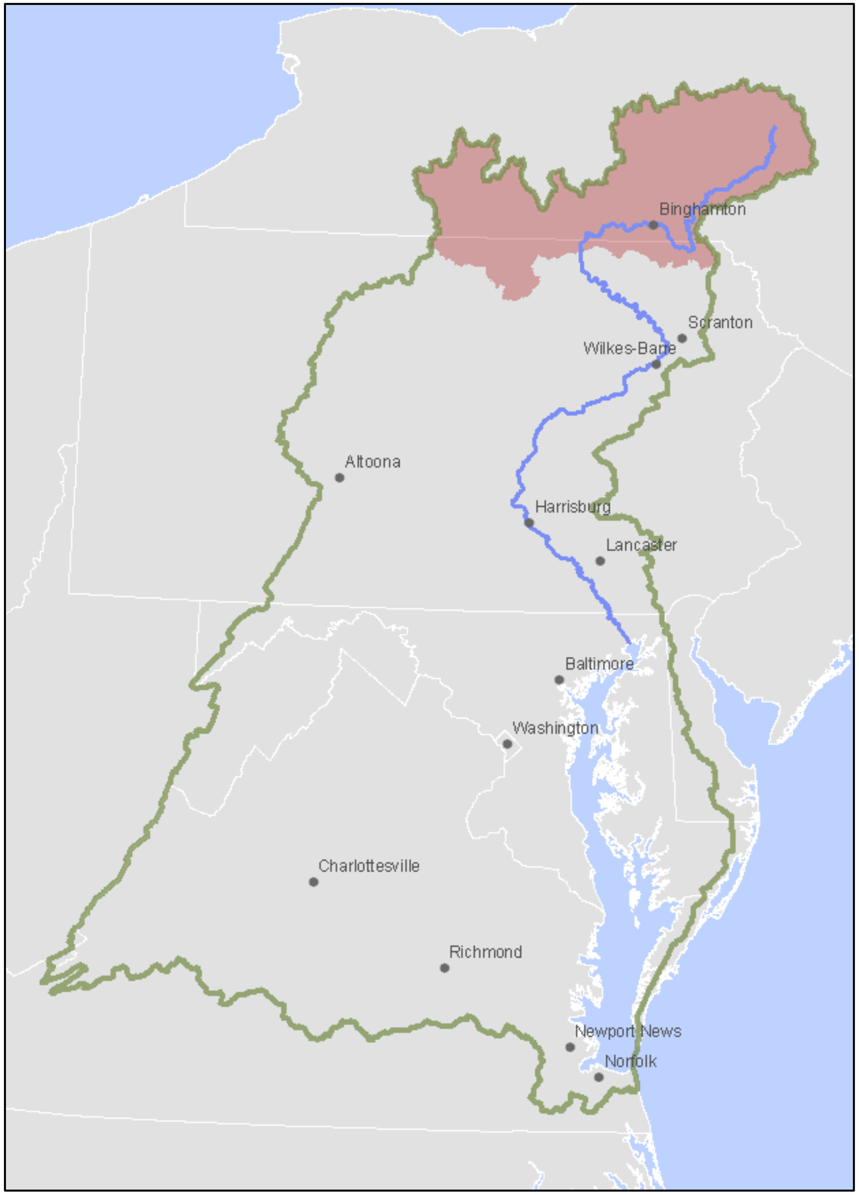
Scientific context

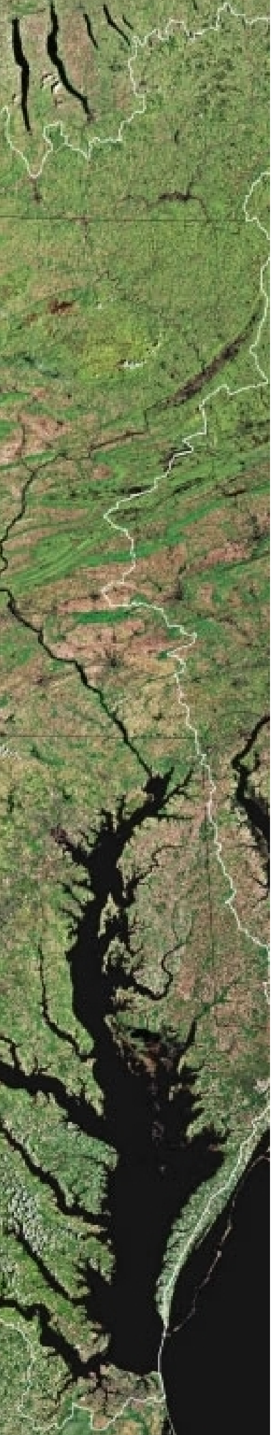
Chesapeake Bay watershed

- Largest US estuary
- Critical fishery, habitat
- Sensitive to nutrient pollution
- Chesapeake Bay Agreement of 2000

Upper Susquehanna River Basin

- Susquehanna is largest US river draining to the Atlantic; largest trib to the bay
- MOU with Chesapeake Bay Program commits NYS to water quality goals of Chesapeake Bay Agreement





Collaborators

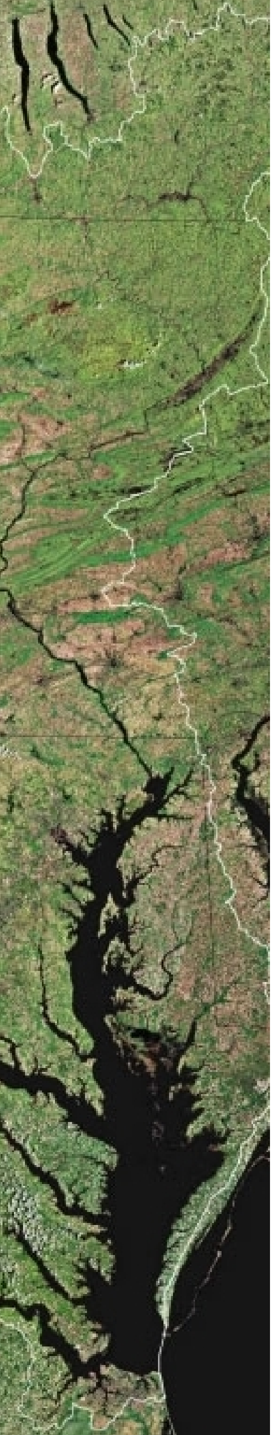
Cornell departments and units:

- Animal Science
- Biological and Environmental Engineering
- Crop and Soil Science
- Ecology and Evolutionary Biology
- Horticulture
- Natural Resources
- Mann Library

Other organizations:

- Cornell Cooperative Extension of Chemung County
- Institute of Ecosystem Studies
- Univ. Maryland Center for Environmental Science
- Univ. Nebraska-Lincoln School of Natural Resources
- Upper Susquehanna Coalition

Funding: USDA Cooperative State Research, Education, and Extension Service



Types of data

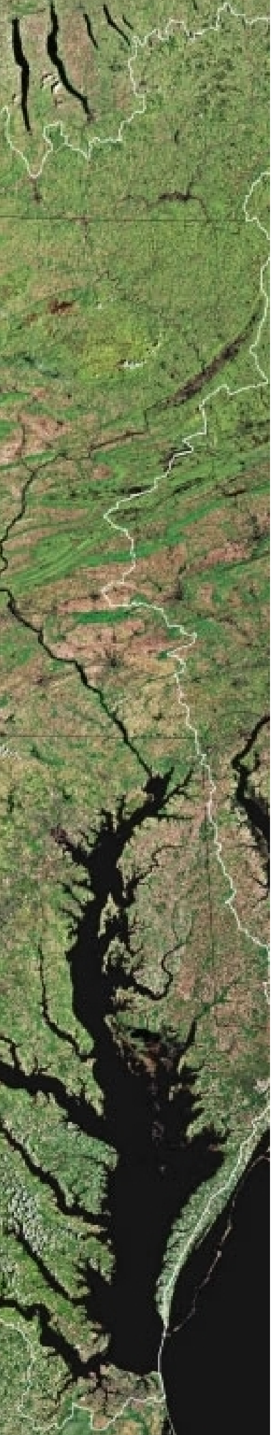
Observational:

- Atmospheric deposition of N
- Water and soil chemistry
- Hydrologic measurements
- Meteorological data
- Plant tissue chemistry
- Cs-137 in stream sediments

Experimental:

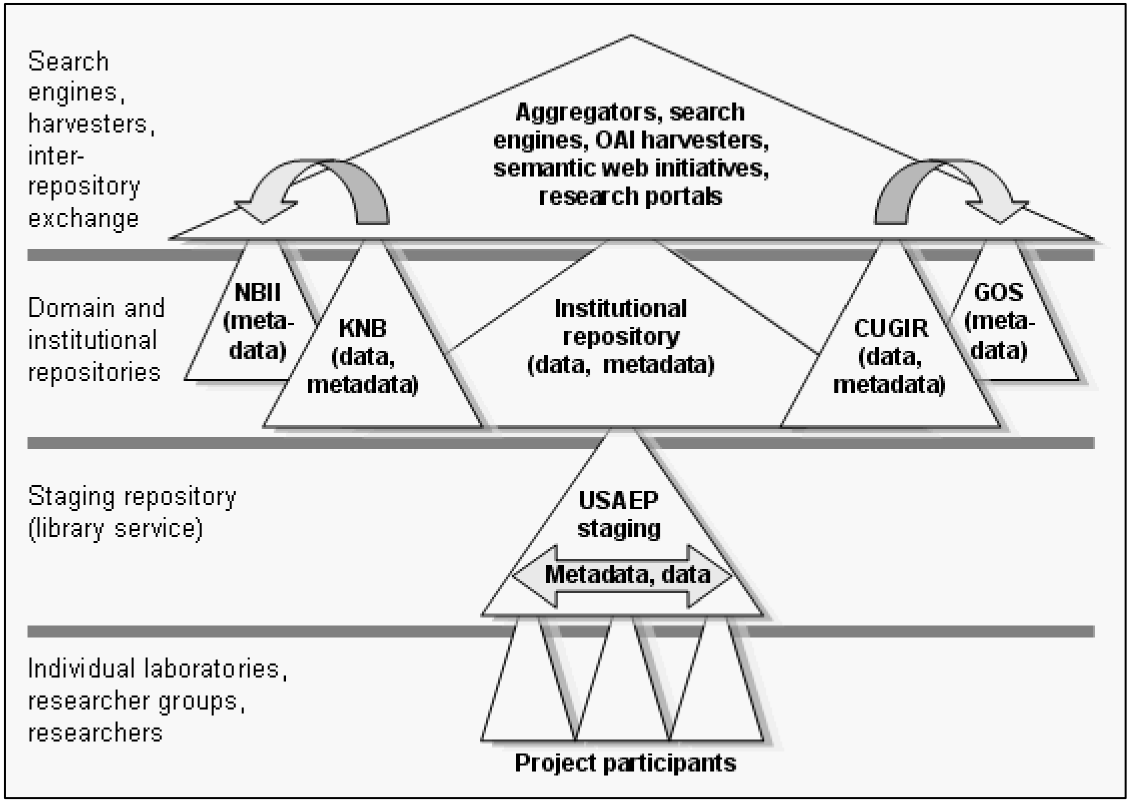
- Effects of willow char amendments to agricultural soils
- Wetland plant species responses to changes in S and P cycling
- Changes in ground water chemistry as a result of chemical amendments
- N leaching in soils under different cropping systems and snow cover manipulations
- Changes in forest and old field chemistry as a result of N fertilization

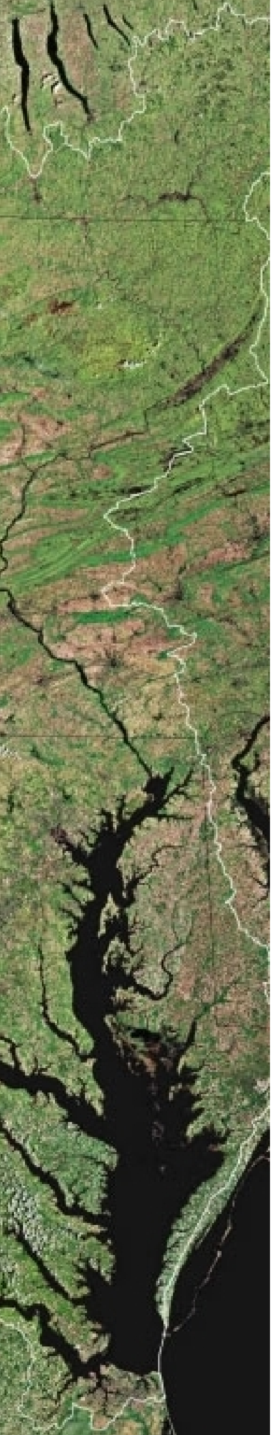
Simulation models of nutrient and sediment fluxes



Strategy

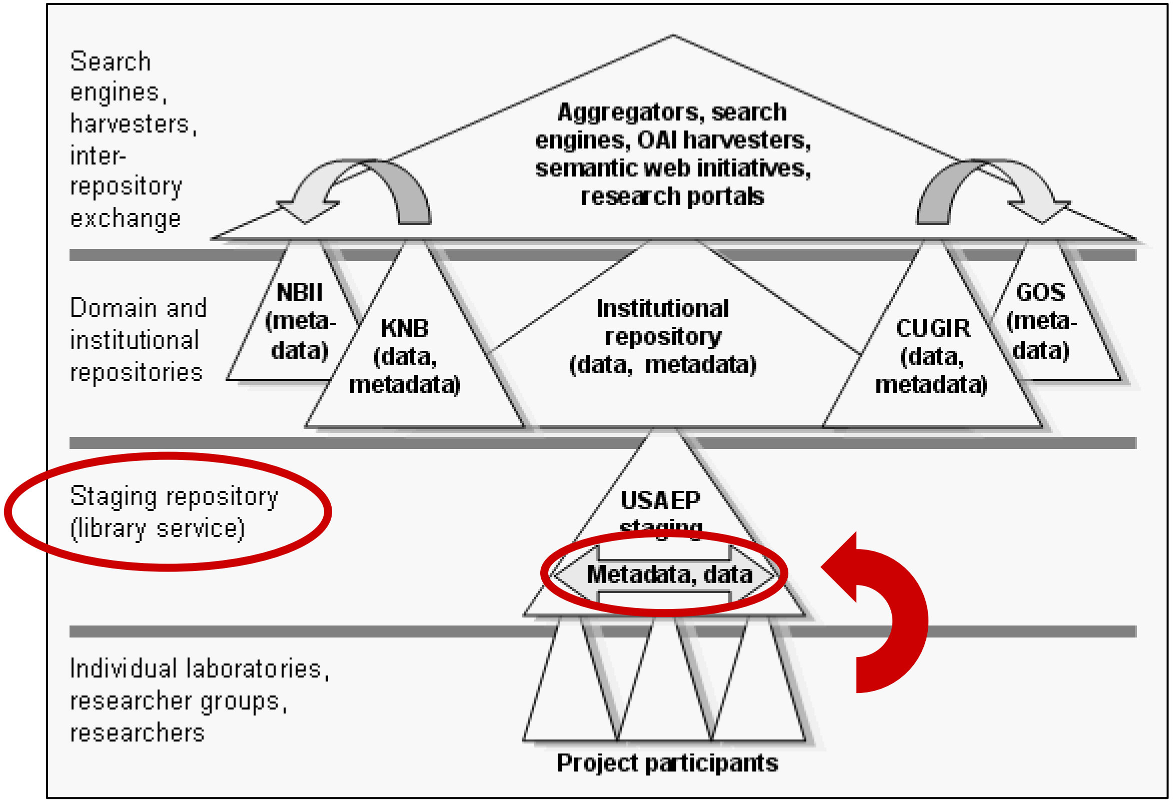
- Local support for data and metadata preparation
- Local and/or discipline-based “publication” of data, metadata


Search engines, harvesters, inter-repository exchange
Aggregators, search engines, OAI harvesters, semantic web initiatives, research portals
Domain and institutional repositories
NBII (meta-data)
KNB (data, metadata)
Institutional repository (data, metadata)
CUGIR (data, metadata)
GOS (meta-data)
Staging repository (library service)
USAEP staging
Metadata, data
Individual laboratories, researcher groups, researchers
Project participants



Strategy

- Local support for data and metadata preparation
- Local and/or discipline-based “publication” of data, metadata



Search engines, harvesters, inter-repository exchange

Aggregators, search engines, OAI harvesters, semantic web initiatives, research portals

Domain and institutional repositories

NBII (meta-data)

KNB (data, metadata)

Institutional repository (data, metadata)

CUGIR (data, metadata)

GOS (meta-data)

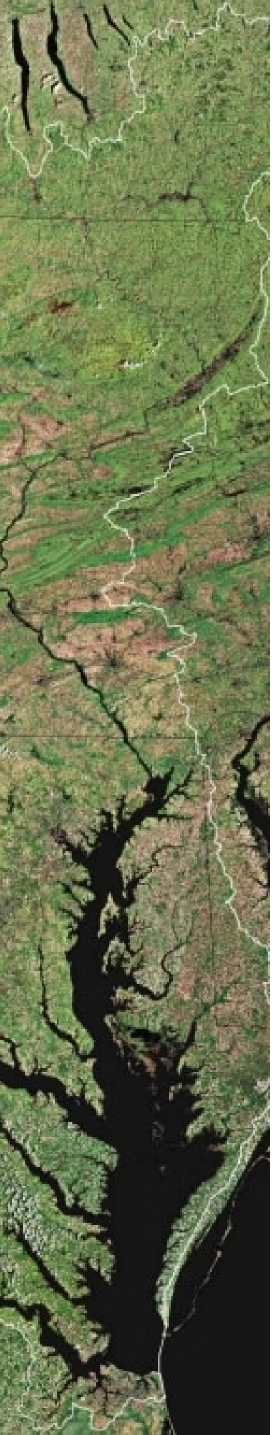
Staging repository (library service)

USAEP staging

Metadata, data

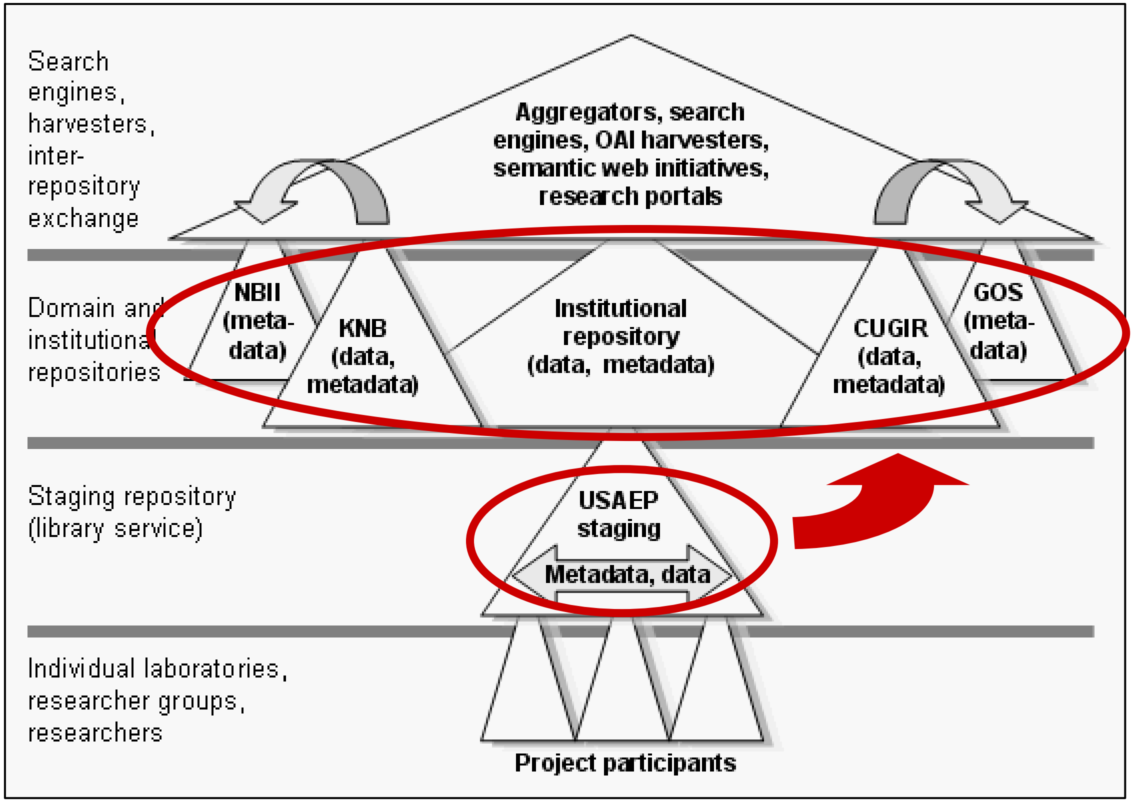
Individual laboratories, researcher groups, researchers

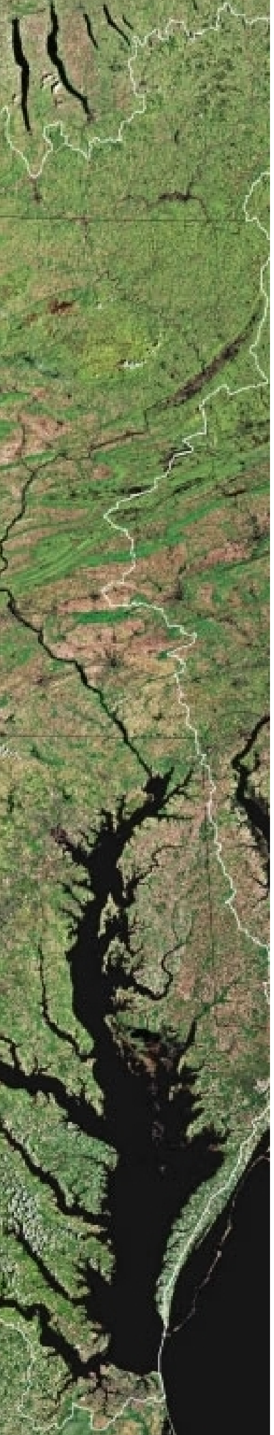
Project participants



Strategy

- Local support for data and metadata preparation
- Local and/or discipline-based “publication” of data, metadata


Search engines, harvesters, inter-repository exchange
Aggregators, search engines, OAI harvesters, semantic web initiatives, research portals
Domain and institutional repositories
NBII (meta-data)
KNB (data, metadata)
Institutional repository (data, metadata)
CUGIR (data, metadata)
GOS (meta-data)
Staging repository (library service)
USAEP staging
Metadata, data
Individual laboratories, researcher groups, researchers
Project participants



“Staging repository”

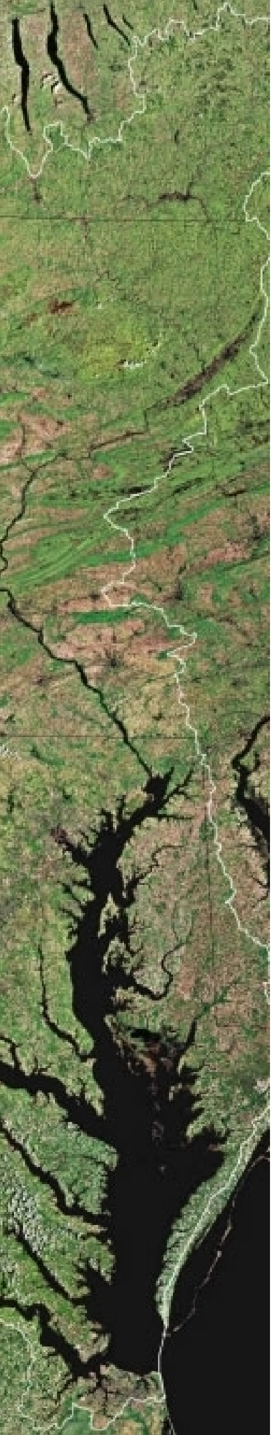
- Use discipline-specific metadata standards and tools

  >> Ecological Metadata Language (EML)
  >> Morpho

- Provide a place to share pre-publication data within the group

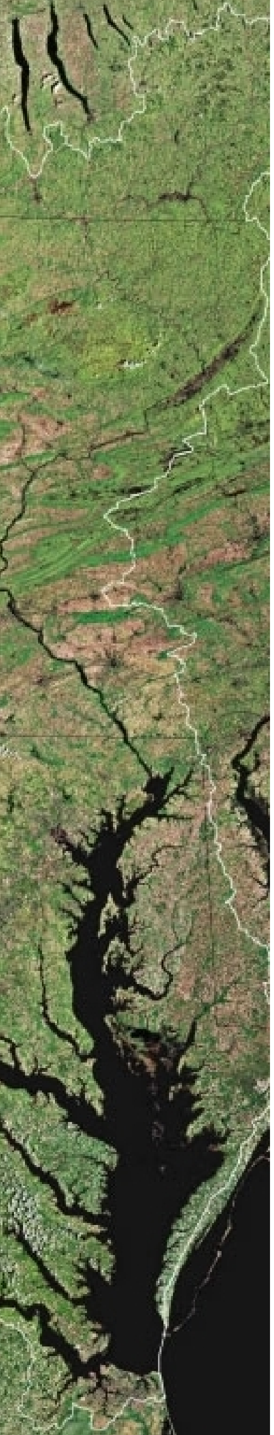
  >> Metacat

- Provide training and recommendations on data and metadata preparation



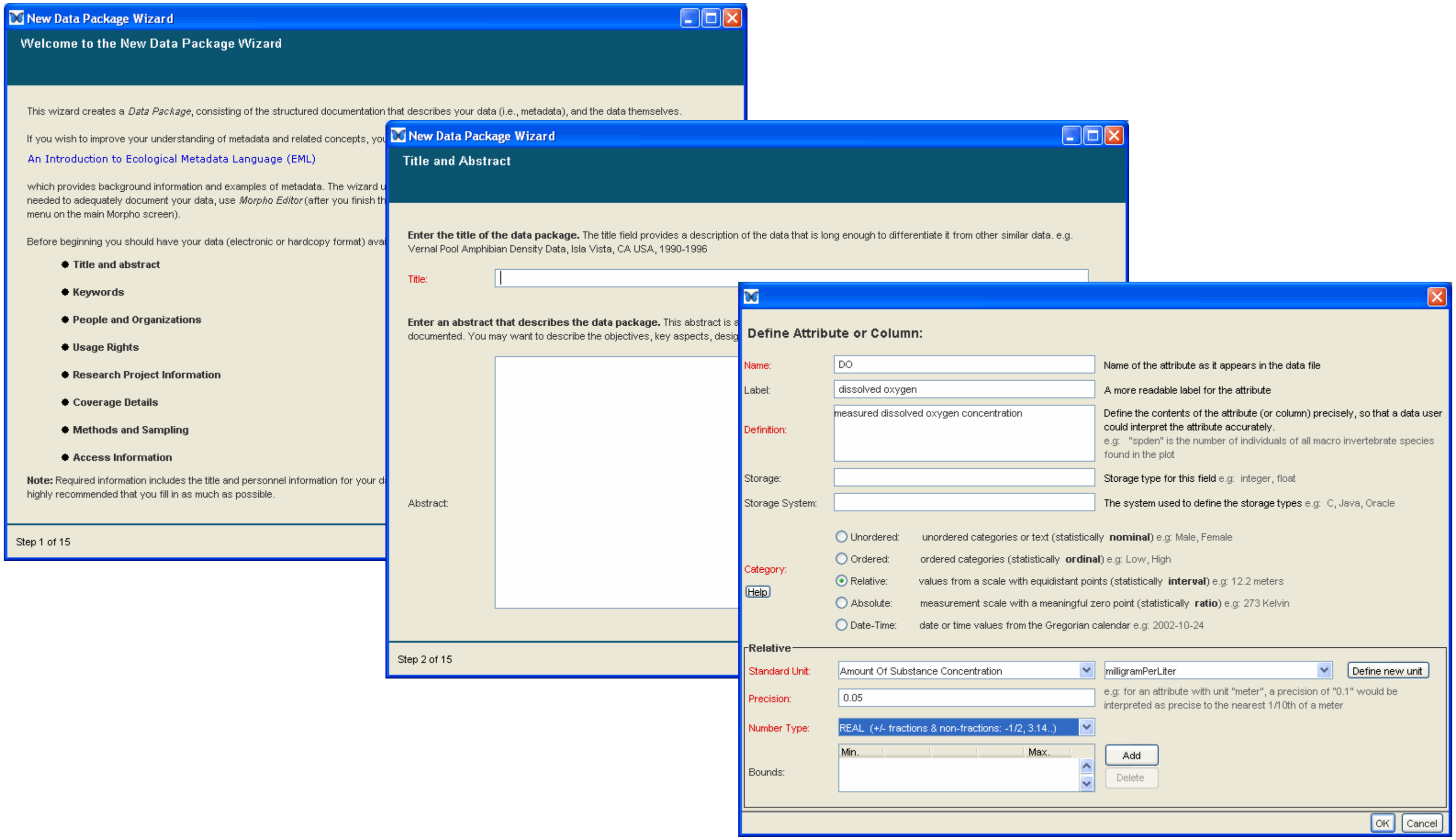
Ecological Metadata Language: EML

- Developed specifically for ecological data (NCEAS, LTER)
- Modular and extensible XML-based standard
- Accommodates information on methods, geographic coverage, temporal coverage, detailed descriptions of tabular data
- http://knb.ecoinformatics.org/software/eml/
- Comes with tools!



Morpho

- Easy to use, platform independent metadata editor.
- Interacts with Metacat: allows users to upload metadata and data; allows users to search, view, and export public data and metadata.



Welcome to the New Data Package Wizard

This wizard creates a *Data Package*, consisting of the structured documentation that describes your data (i.e., metadata), and the data themselves.

If you wish to improve your understanding of metadata and related concepts, yo

An Introduction to Ecological Metadata Language (EML)

which provides background information and examples of metadata. The wizard u
needed to adequately document your data, use *Morpho Editor* (after you finish th
menu on the main Morpho screen).

Before beginning you should have your data (electronic or hardcopy format) avai

- **Title and abstract**
- **Keywords**
- **People and Organizations**
- **Usage Rights**
- **Research Project Information**
- **Coverage Details**
- **Methods and Sampling**
- **Access Information**

Note: Required information includes the title and personnel information for your d
highly recommended that you fill in as much as possible.

Step 1 of 15

Title and Abstract

Enter the title of the data package. The title field provides a description of the data that is long enough to differentiate it from other similar data. e.g. Vernal Pool Amphibian Density Data, Isla Vista, CA USA, 1990-1996

Title:

Enter an abstract that describes the data package. This abstract is a
documented. You may want to describe the objectives, key aspects, desig

Abstract:

Step 2 of 15

Define Attribute or Column:

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

Label: dissolved oxygen — A more readable label for the attribute

Definition: measured dissolved oxygen concentration — 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

Storage: — Storage type for this field e.g: integer, float

Storage System: — The system used to define the storage types e.g: C, Java, Oracle

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

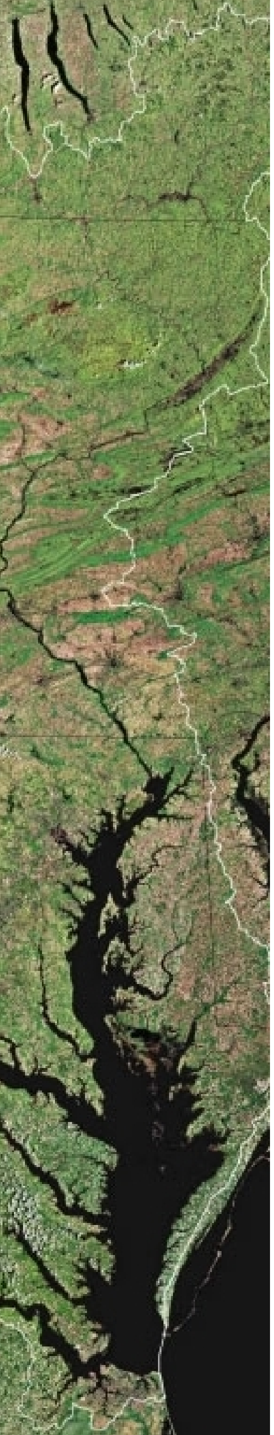
Relative

Standard Unit: Amount Of Substance Concentration — milligramPerLiter

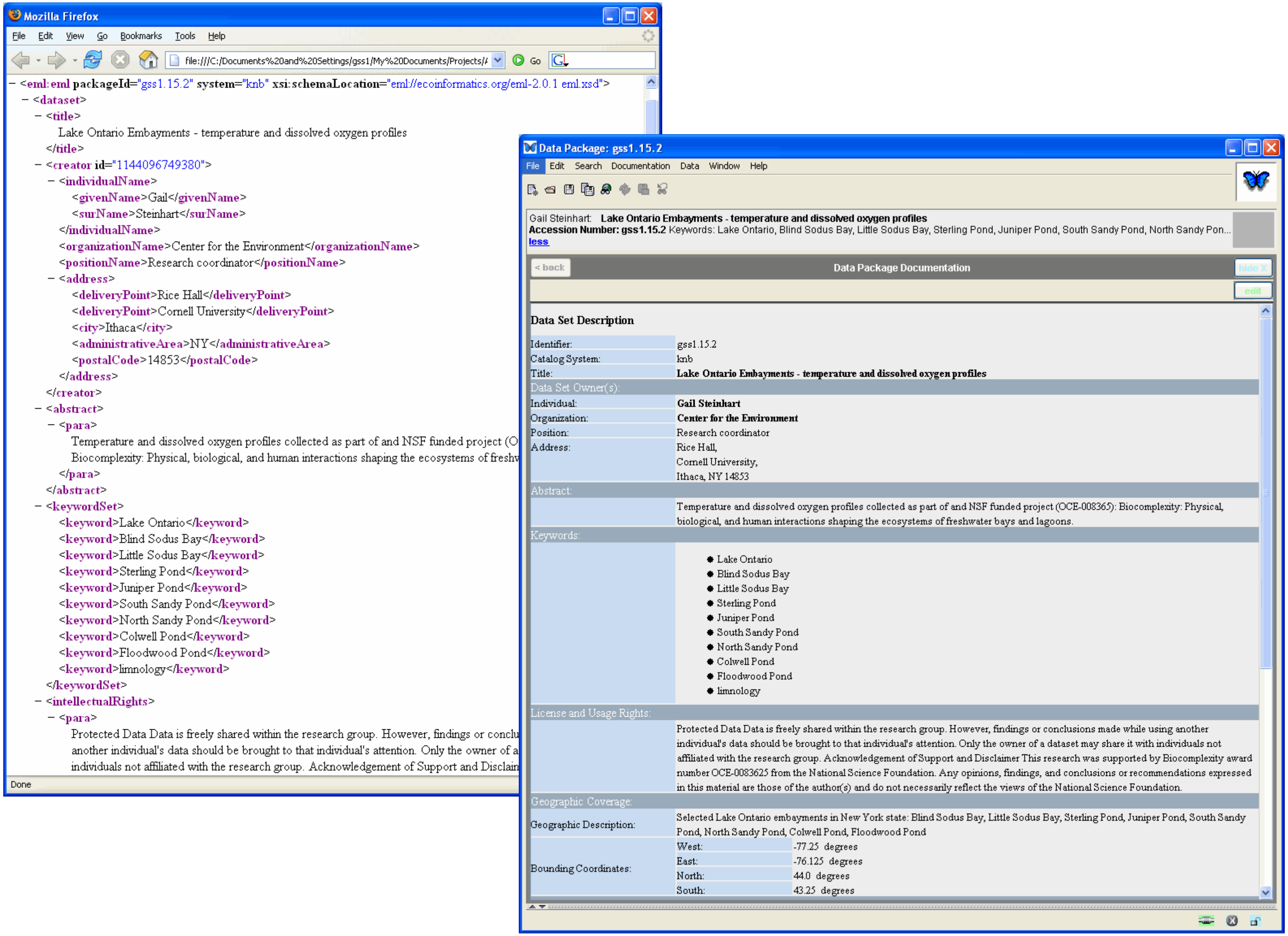
Precision: 0.05 — e.g: for an attribute with unit "meter", a precision of "0.1" would be interpreted as precise to the nearest 1/10th of a meter

Number Type: REAL (+/- fractions & non-fractions: -1/2, 3.14..)

Bounds: Min. Max.



EML record



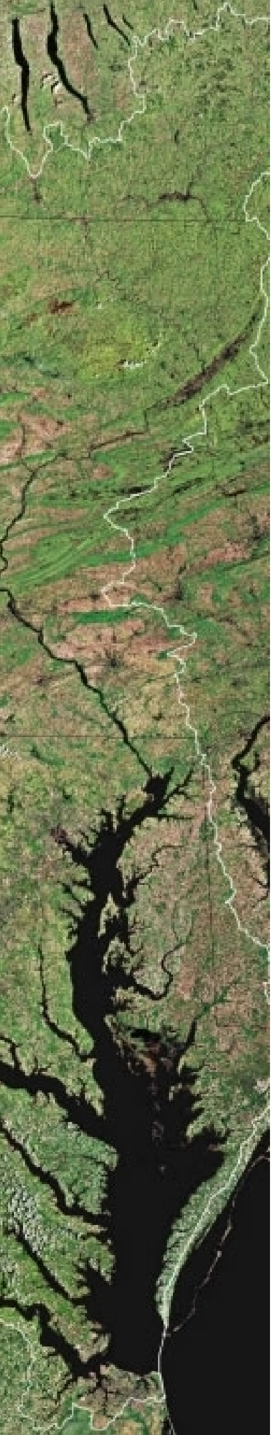
Mozilla Firefox
File Edit View Go Bookmarks Tools Help
file:///C:/Documents%20and%20Settings/gss1/My%20Documents/Projects/
Go
Done

```
– <eml:eml packageId="gss1.15.2" system="knb" xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 eml.xsd">
  – <dataset>
    – <title>
        Lake Ontario Embayments - temperature and dissolved oxygen profiles
      </title>
    – <creator id="1144096749380">
      – <individualName>
          <givenName>Gail</givenName>
          <surName>Steinhart</surName>
        </individualName>
        <organizationName>Center for the Environment</organizationName>
        <positionName>Research coordinator</positionName>
      – <address>
          <deliveryPoint>Rice Hall</deliveryPoint>
          <deliveryPoint>Cornell University</deliveryPoint>
          <city>Ithaca</city>
          <administrativeArea>NY</administrativeArea>
          <postalCode>14853</postalCode>
        </address>
      </creator>
    – <abstract>
      – <para>
          Temperature and dissolved oxygen profiles collected as part of and NSF funded project (O
          Biocomplexity: Physical, biological, and human interactions shaping the ecosystems of freshw
        </para>
      </abstract>
    – <keywordSet>
        <keyword>Lake Ontario</keyword>
        <keyword>Blind Sodus Bay</keyword>
        <keyword>Little Sodus Bay</keyword>
        <keyword>Sterling Pond</keyword>
        <keyword>Juniper Pond</keyword>
        <keyword>South Sandy Pond</keyword>
        <keyword>North Sandy Pond</keyword>
        <keyword>Colwell Pond</keyword>
        <keyword>Floodwood Pond</keyword>
        <keyword>limnology</keyword>
      </keywordSet>
    – <intellectualRights>
      – <para>
          Protected Data Data is freely shared within the research group. However, findings or conclu
          another individual's data should be brought to that individual's attention. Only the owner of a
          individuals not affiliated with the research group. Acknowledgement of Support and Disclaim
```

Gail Steinhart: **Lake Ontario Embayments - temperature and dissolved oxygen profiles**
Accession Number: gss1.15.2 Keywords: Lake Ontario, Blind Sodus Bay, Little Sodus Bay, Sterling Pond, Juniper Pond, South Sandy Pond, North Sandy Pon...
less

Data Set Description

Identifier:	gss1.15.2
Catalog System:	knb
Title:	**Lake Ontario Embayments - temperature and dissolved oxygen profiles**
Data Set Owner(s):	
Individual:	**Gail Steinhart**
Organization:	**Center for the Environment**
Position:	Research coordinator
Address:	Rice Hall, Cornell University, Ithaca, NY 14853
Abstract:	
	Temperature and dissolved oxygen profiles collected as part of and NSF funded project (OCE-008365): Biocomplexity: Physical, biological, and human interactions shaping the ecosystems of freshwater bays and lagoons.
Keywords:	
	● Lake Ontario ● Blind Sodus Bay ● Little Sodus Bay ● Sterling Pond ● Juniper Pond ● South Sandy Pond ● North Sandy Pond ● Colwell Pond ● Floodwood Pond ● limnology
License and Usage Rights:	
	Protected Data Data is freely shared within the research group. However, findings or conclusions made while using another individual's data should be brought to that individual's attention. Only the owner of a dataset may share it with individuals not affiliated with the research group. Acknowledgement of Support and Disclaimer This research was supported by Biocomplexity award number OCE-0083625 from the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Geographic Coverage:	
Geographic Description:	Selected Lake Ontario embayments in New York state: Blind Sodus Bay, Little Sodus Bay, Sterling Pond, Juniper Pond, South Sandy Pond, North Sandy Pond, Colwell Pond, Floodwood Pond
Bounding Coordinates:	West: -77.25 degrees East: -76.125 degrees North: 44.0 degrees South: 43.25 degrees



“Publication” of data

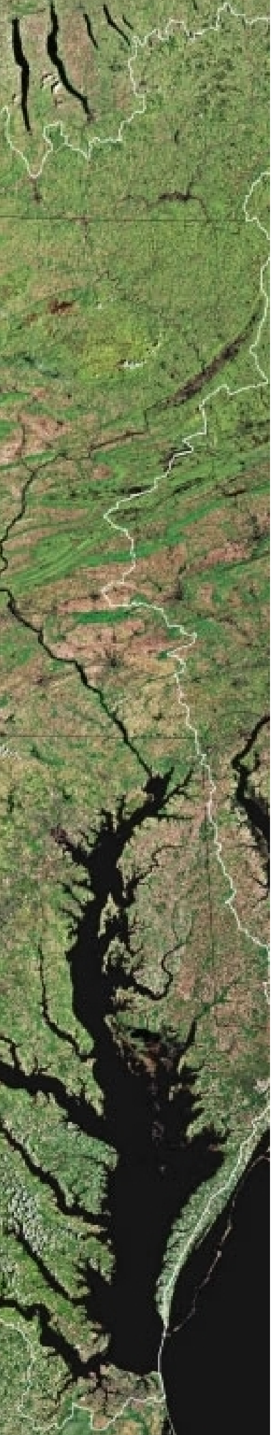
- Deposit in institutional repository

  >> DSpace

- Submit metadata (and possibly data) to discipline-specific repository

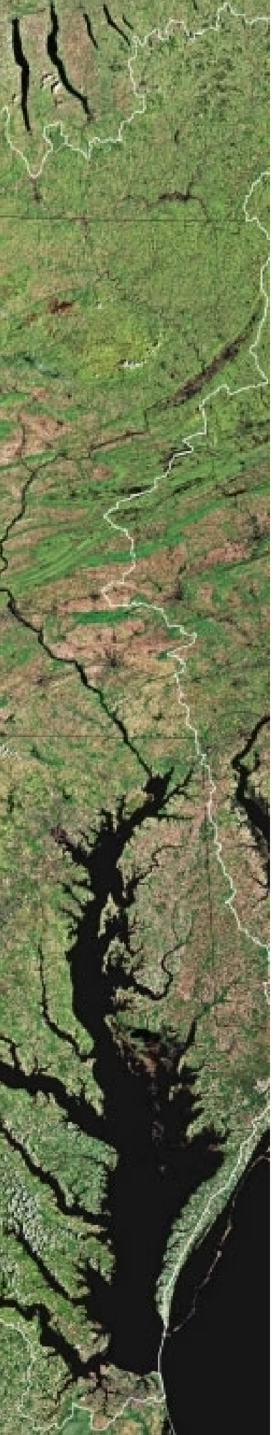
  >> KNB, other?

- Link from project web portal: http://www.usaep.mannlib.cornell.edu/



Test case: Historical data

- Observational data from last 30 years
- Original format: Quattro Pro workbooks with multiple pages
- Various errors (apparent duplicate records, misaligned columns, out of range values)
- Missing or ambiguous information (methods, units, geographic locations)
- *Extensible model?*



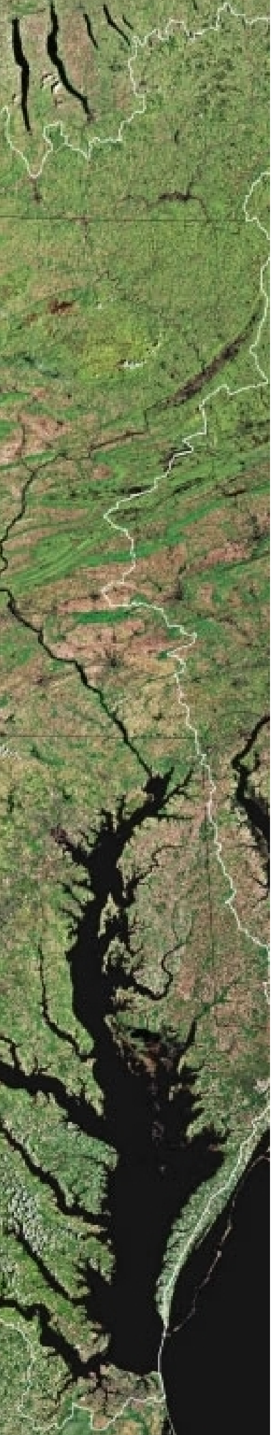
Summary – curation skills

- “Traditional” library and archiving skills (metadata, preservation, interoperability, appraisal and selection)

- Understanding of CI

- Subject area knowledge:
  - Understanding of research practices, tools, and culture (may be discipline-specific)
  - Awareness of standards and tools related to data

- Productive partnerships with researchers (or ability to develop them)



Thank you

Gail Steinhart
Research Data & Environmental Sciences Librarian
Albert R. Mann Library, Cornell University

GSS1@cornell.edu