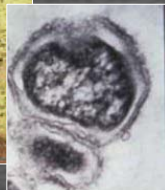


# Biological Informatics: The Challenge of Data Integration



Gladys Cotter  
Associate Chief Biologist for Information  
U.S. Geological Survey

# Aquatic food webs

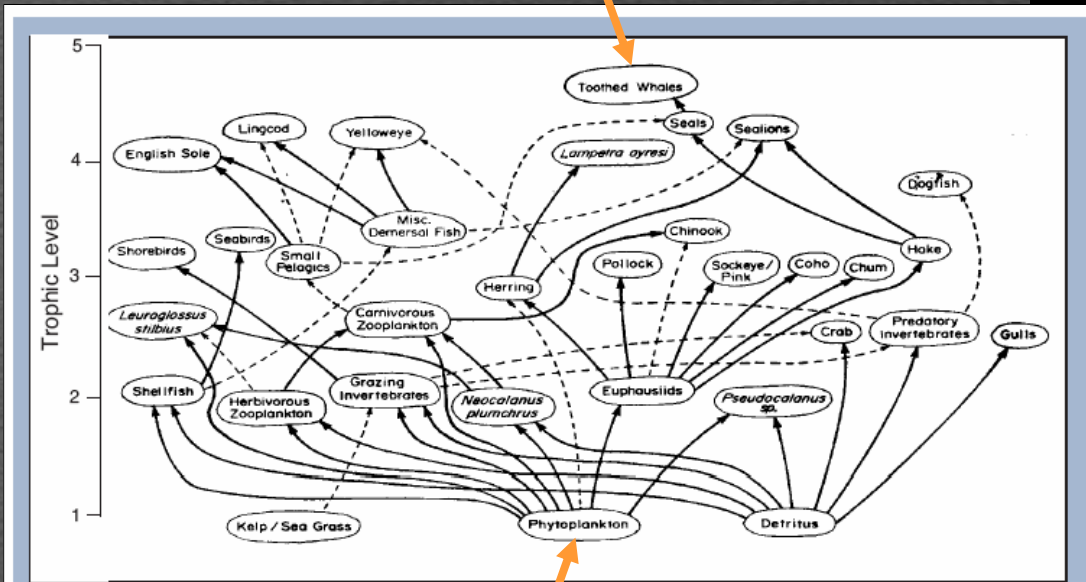
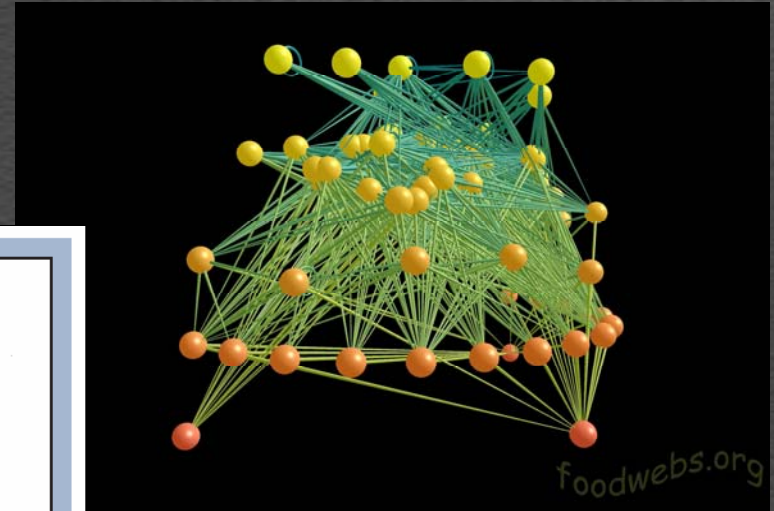
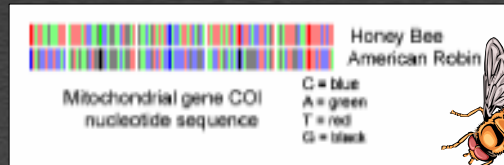


Figure 4. A schematic representation of the relationship in 1998 between the 32 function groups in the Strait of Georgia Ecopath model. Prey items representing greater than 25 percent (solid lines) and 20-25 percent (dashed lines) of a functional groups diet are represented in the diagram. Prey items representing less than 10 percent of a functional groups diet are not included.

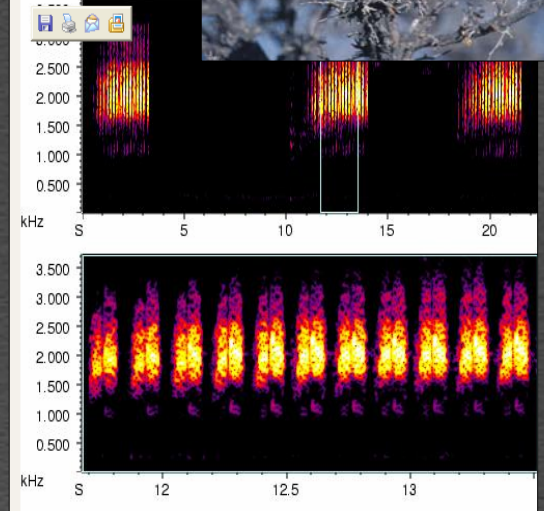
Beamish et al., 2003



Provided by the PEaCE Lab ([www.foodwebs.org](http://www.foodwebs.org))



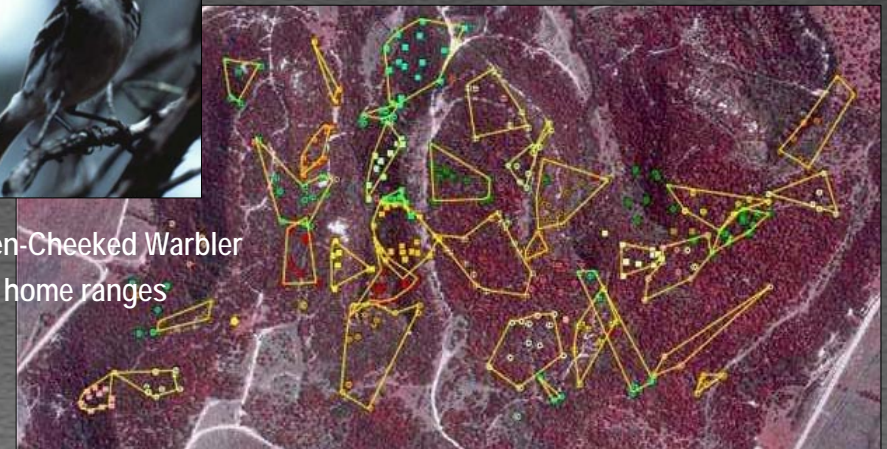
Cactus Wren



Cactus wren, *Campylorhynchus brunneicapillus*  
©2004 Cornell Lab of Ornithology

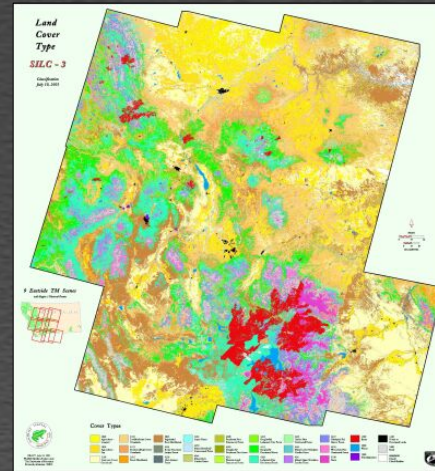
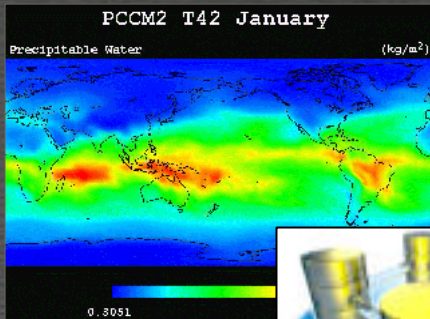


Golden-Cheeked Warbler  
home ranges



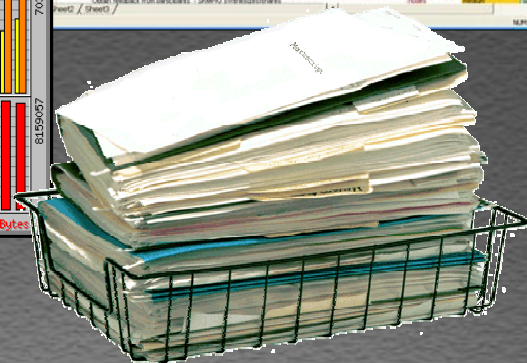
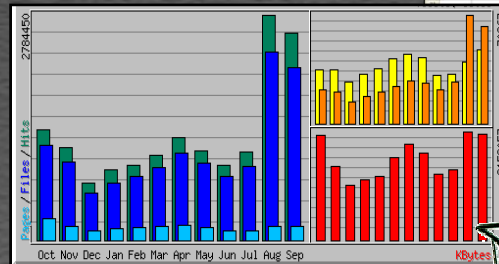
# Biological information

## Data format diversity



Objectives, and Strategies compilation.xls [Shared]

purpose	priority	indicator
access the NEB site with other interested parties	SHAWO & nodes	NEST
developing relations, generating ideas, and taking action to improve the sharing of species information	SHAWO	WATER
group communication to all nodes and NEB partners	SHAWO	medium-high
these participants in "revisited meetings"	SHAWO & nodes	medium-high
action for a charter	SHAWO	medium
workable with others, identify the working group strategies that need to be made	SHAWO	medium-high
its working groups or partners	SHAWO	medium-high
presented by the working group	SHAWO & nodes	NEST
on the ongoing species-related activities (ITS, Species pages, etc.)	SHAWO & nodes	Species Project
identify challenges faced by node members in sharing information.	SHAWO & nodes	NEST
and other forms of communication, identify goals for the group.	SHAWO & nodes	NEST
to accommodate changing landscape or issues in membership of the SHAWO and give	SHAWO	WATER
ways for improving the sharing of species information.	SHAWO & nodes	medium-high
and actions from the NEB nodes on a regular basis	SHAWO & nodes	medium
on methodology (specifics under other goals)	SHAWO	NEST
ensure the efficient use of resources in managing species knowledge management	SHAWO	medium-high
steps, and outcome flowchart, with priorities	SHAWO	medium-high
	SHAWO	medium
USGS' NEB Management Plan that will be updated every three years, or five document.	SHAWO	medium
LTG: Ensure that our species data and information serves audience needs: USABILITY	SHAWO & nodes	medium-high
1. Determine our audience	SHAWO & nodes	medium-high
Define role from each node on who is their audience - SHAWO synthesizes	nodes	medium-high
2. Conduct a formal study of specific audience types - SHAWO synthesizes	nodes	medium-high
3. Find out what each audience type needs, and whether able to understand information currently provided	SHAWO & nodes	medium
4. Find out how our species data are used in cross-section	nodes, IT	medium
5. Synthesize data from nodes, regularly	SHAWO & nodes	medium
6. Conduct formal studies	SHAWO & nodes	medium
7. Define our purpose in serving species knowledge information, i.e., is it to [enable] researchers to export/import data? - SHAWO & nodes	nodes	medium
8. Obtain feedback from participants - SHAWO synthesizes/strategies	nodes	medium



# Biological Informatics

- A new field at the intersection of biological sciences, information science, and computer technology.
- Helps save, assemble, organize, integrate, and disseminate information on the natural world.
- Covers all biological, spatial, and temporal scales.
- Data comes in a wide variety of formats, and from diverse disciplines.
- Challenge is to meet new and rapidly evolving data and information needs.

# Data and Information Needs

- Research needs:
  - make own data longer-lived and more valuable;
  - Seek additional data for use in validating methods and models; and
  - Seek additional data to answer new, complex questions that involve multiple factors

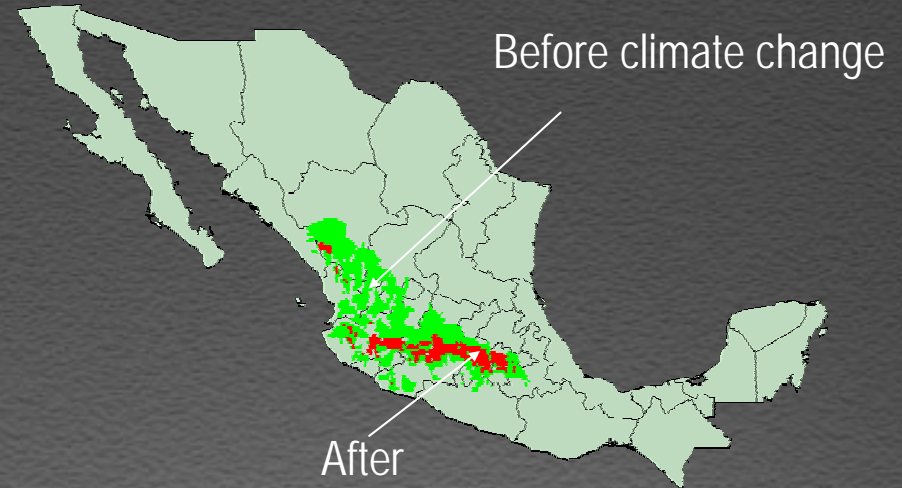
.....

# One Goal - Ecological Forecasting



The potential effect of climate change on the distribution of the Green-striped brush finch, *Atlapetes virenticeps*

University of Kansas

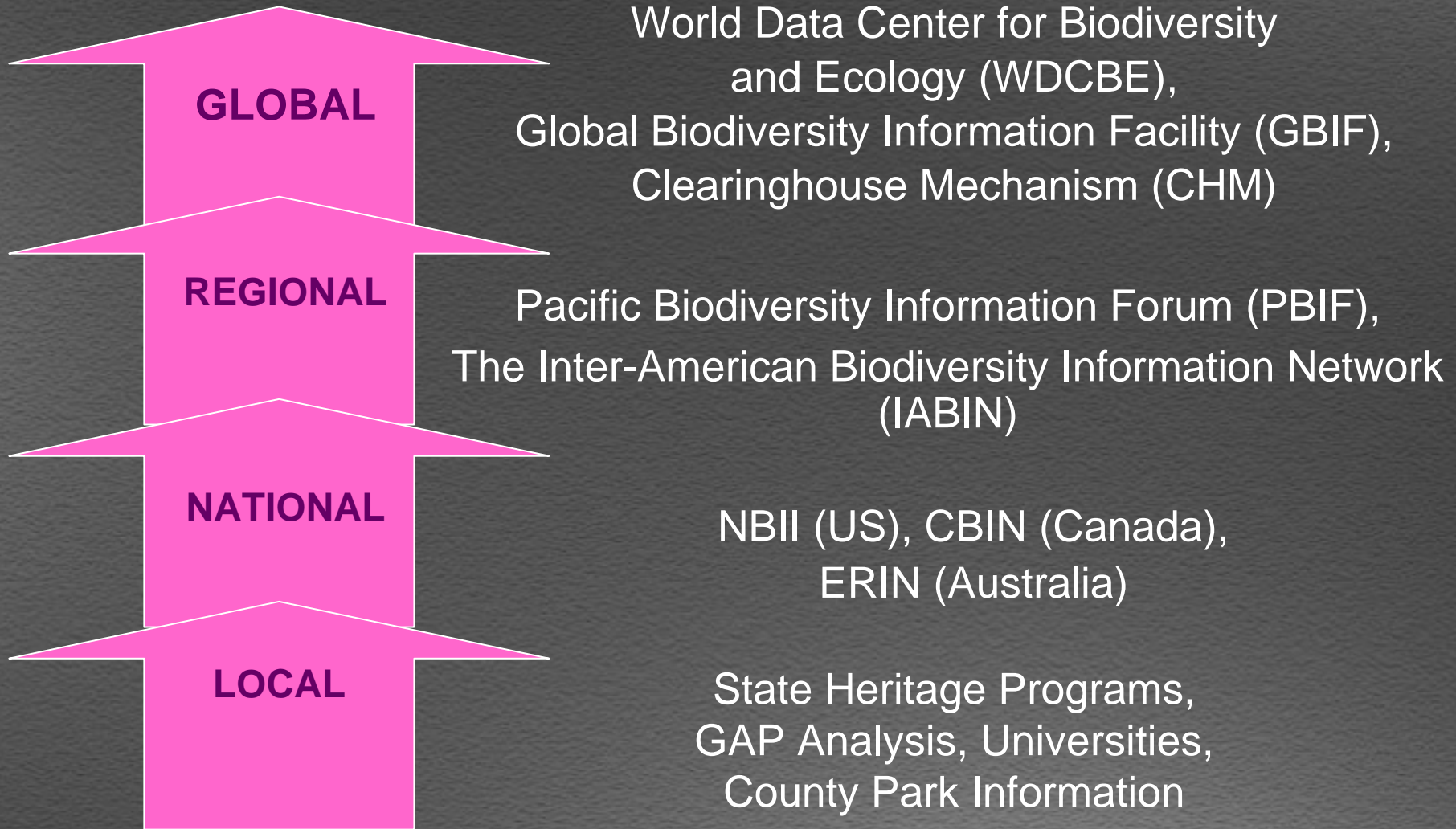


# Data and Information Needs, continued

- Research needs:
  - make data longer-lived and more valuable;
  - seek data for use in validating data and models; and
  - seek data in order to answer new, complex questions that involve multiple factors.
- Societal need to have accessible, accurate, and integrated scientific information for:
  - increased understanding; and
  - Informed decision-making about the environment.



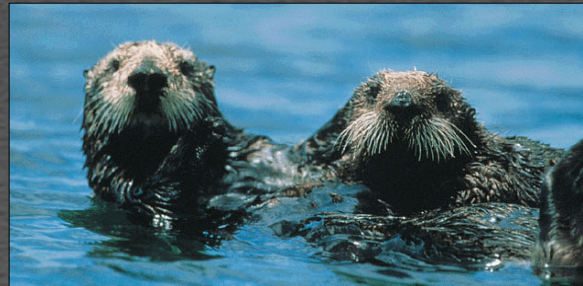
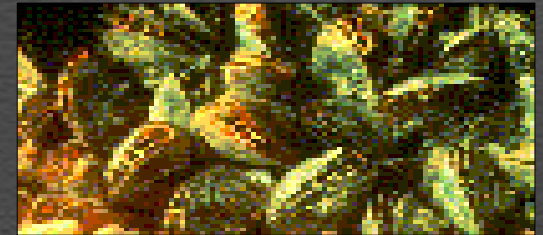
# Biological Informatics Systems



# NBII

## The National Biological Information Infrastructure

A broad, collaborative program to provide access to data and information related to our environment

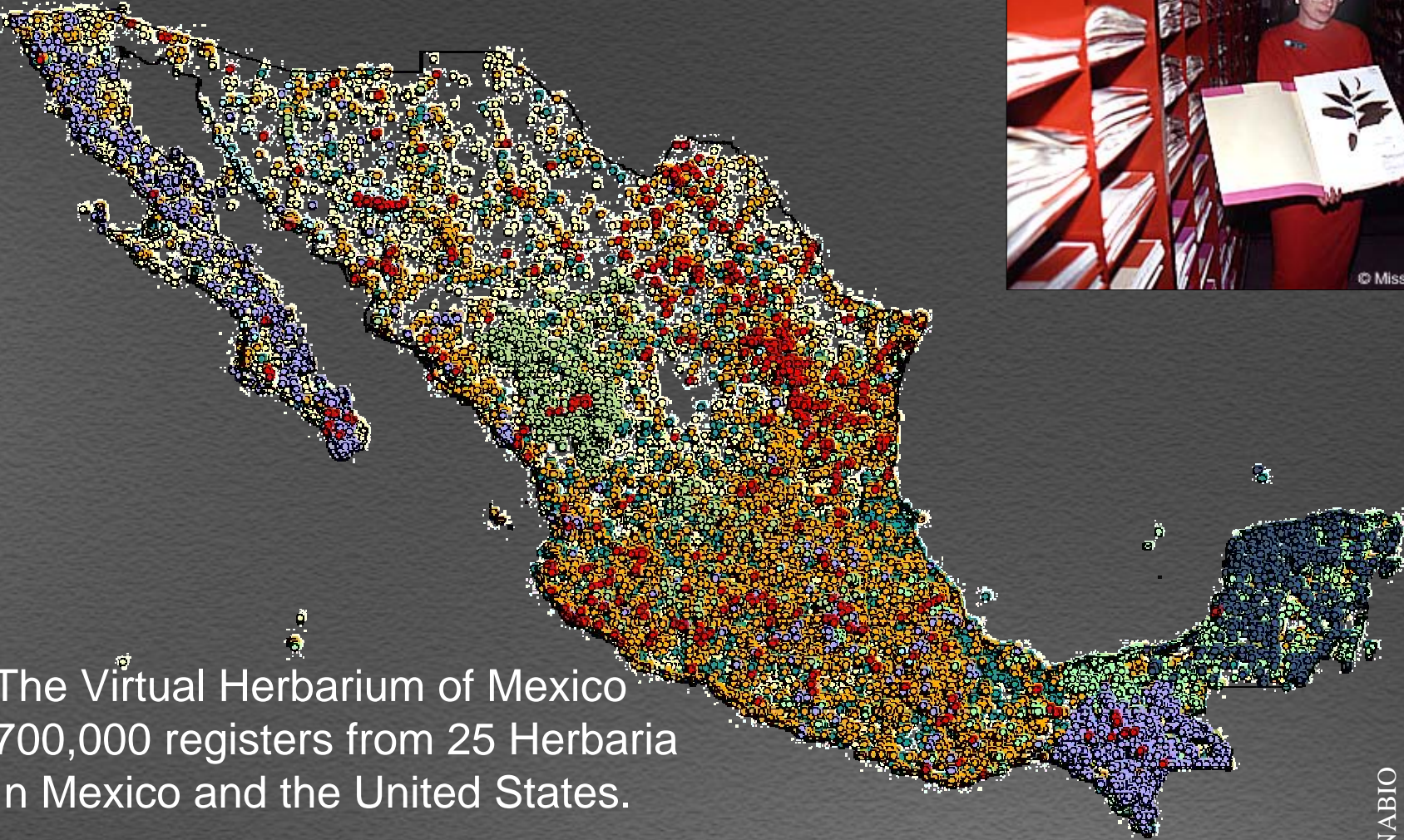


[www.nbii.gov](http://www.nbii.gov)

# Biological Informatics - Challenges in Integrating Data

- Links among complex data from within, and from outside biology (geology, meteorology)
- Different methodologies:
  - recording of a location (“Tyler, Texas” versus latitude/longitude);
  - Different equipment used; and
  - Different data formats - museum specimens versus observations versus satellite images.

• .....



The Virtual Herbarium of Mexico  
700,000 registers from 25 Herbaria  
In Mexico and the United States.



# Biological Informatics - Challenges in Integrating Data, continued

- Links among complex data from within, and then outside biology (geology, meteorology)
- Different methodologies, e.g.:
  - recording of a location (“Tyler, Texas” vs. lat/long);
  - different equipment used; and
  - different data formats - specimens, observations, satellite images.
- Inaccessible, missing, and deteriorating data
- Quality control, intellectual properties, and secure storage
- Training needed that combines biology, computers, and information science
- Linking people of different disciplines

# Challenges, continued

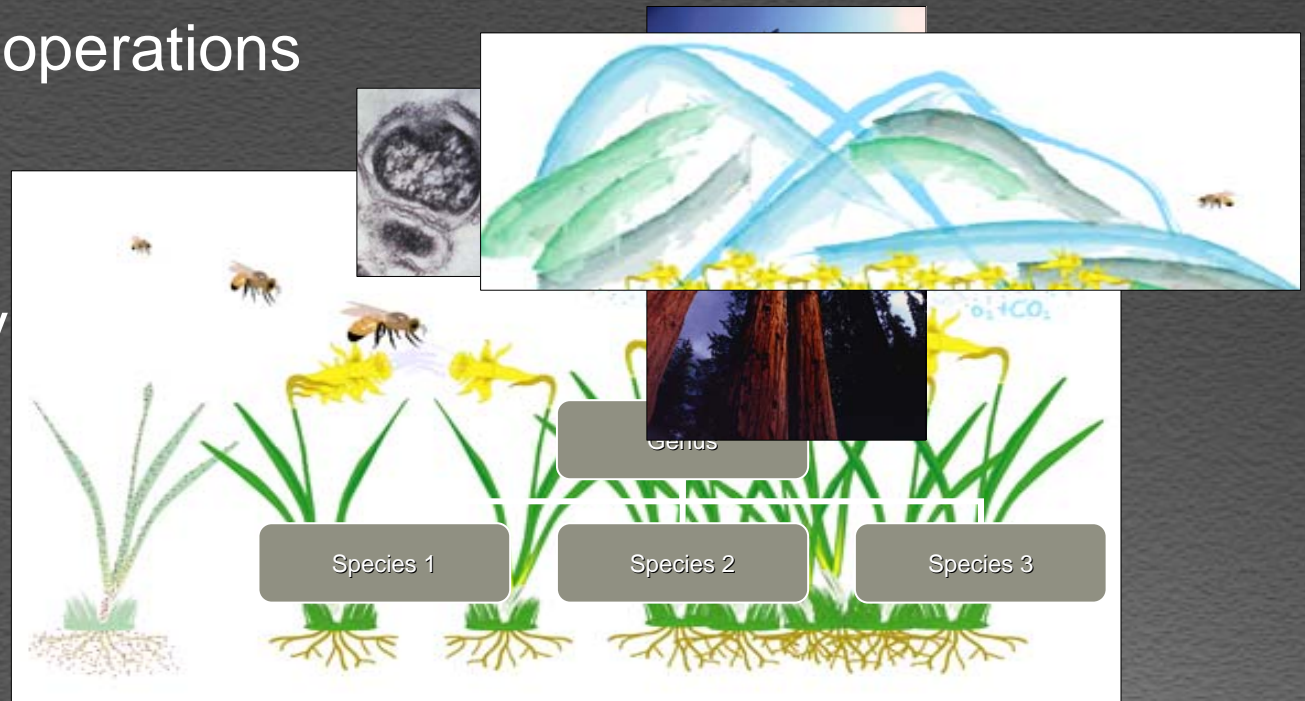
Linking data of different scales within:

- Biology
- Space
- Time

**Scale is the main challenge in integrating data for analysis and understanding.**

# Biological Scales

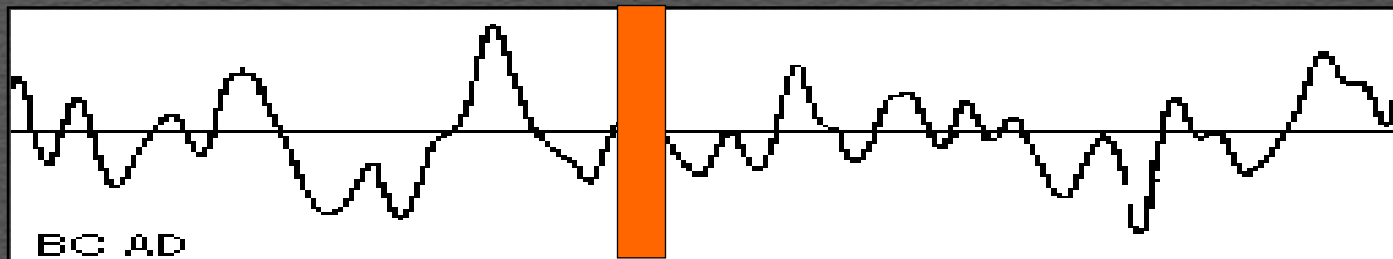
- Organism size
- Biological operations
- Taxonomy



# Space and Time - Continuous dimensions



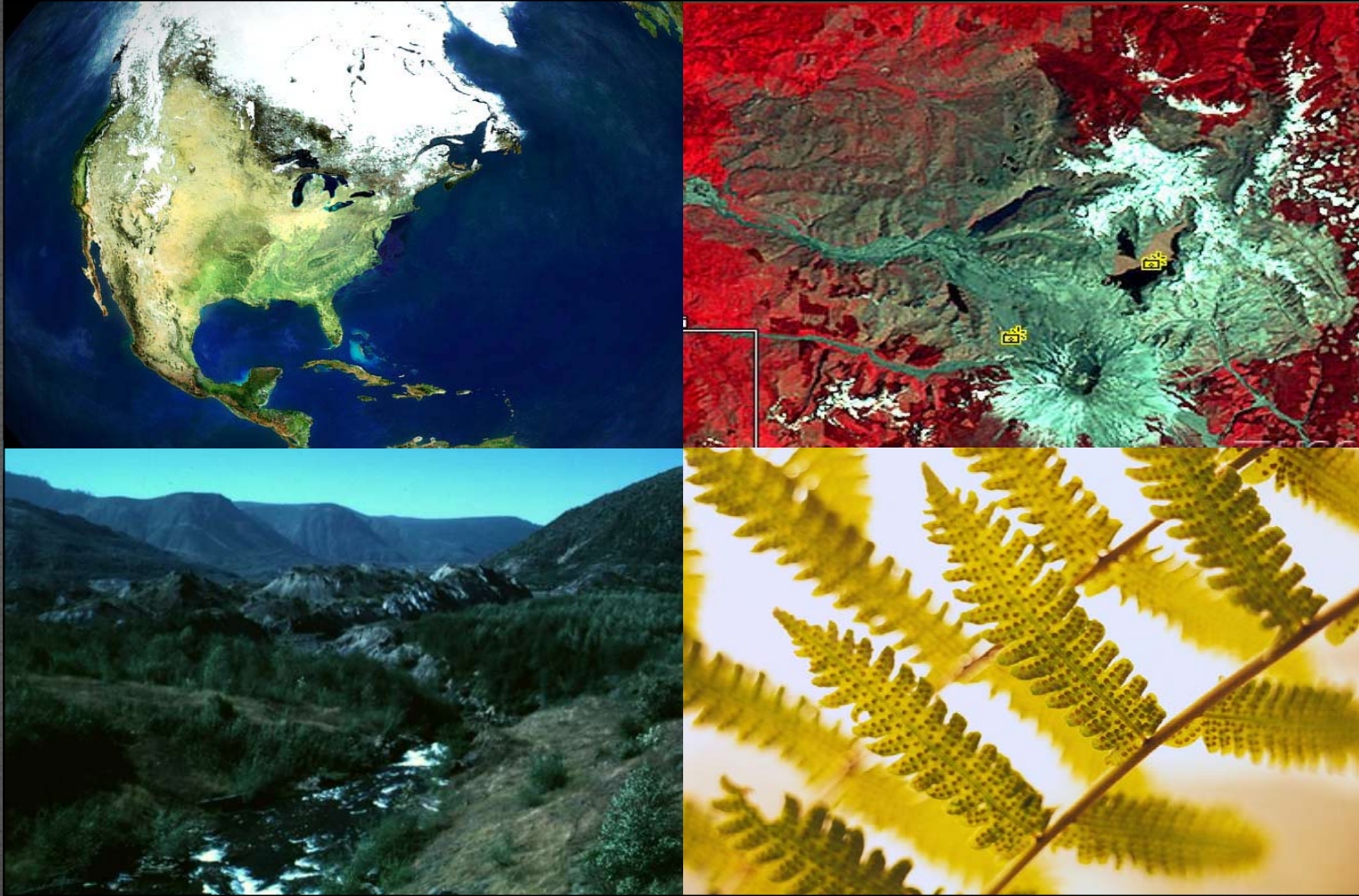
1 centimeter, or 10 millimeters



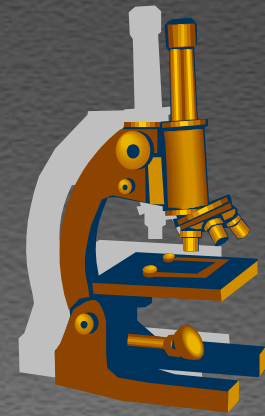
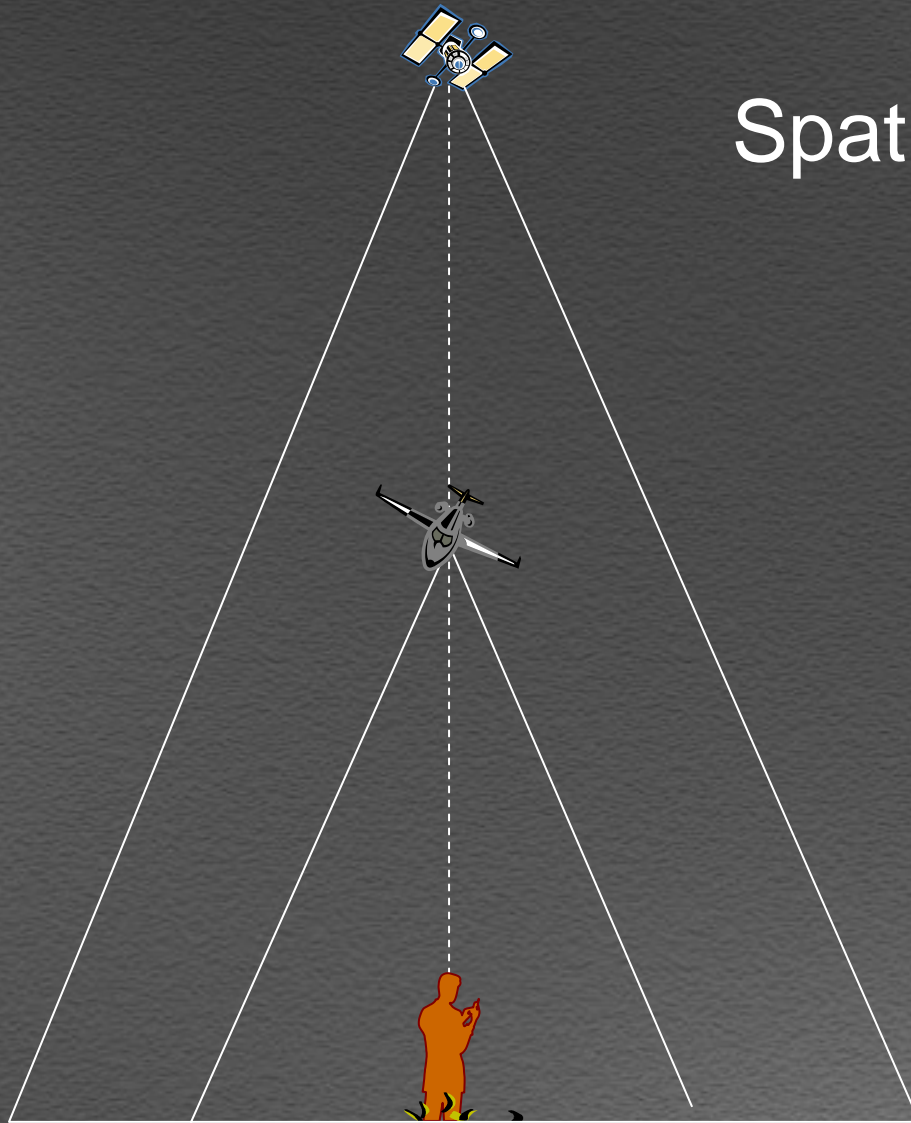
100 years



# Different scales of observation



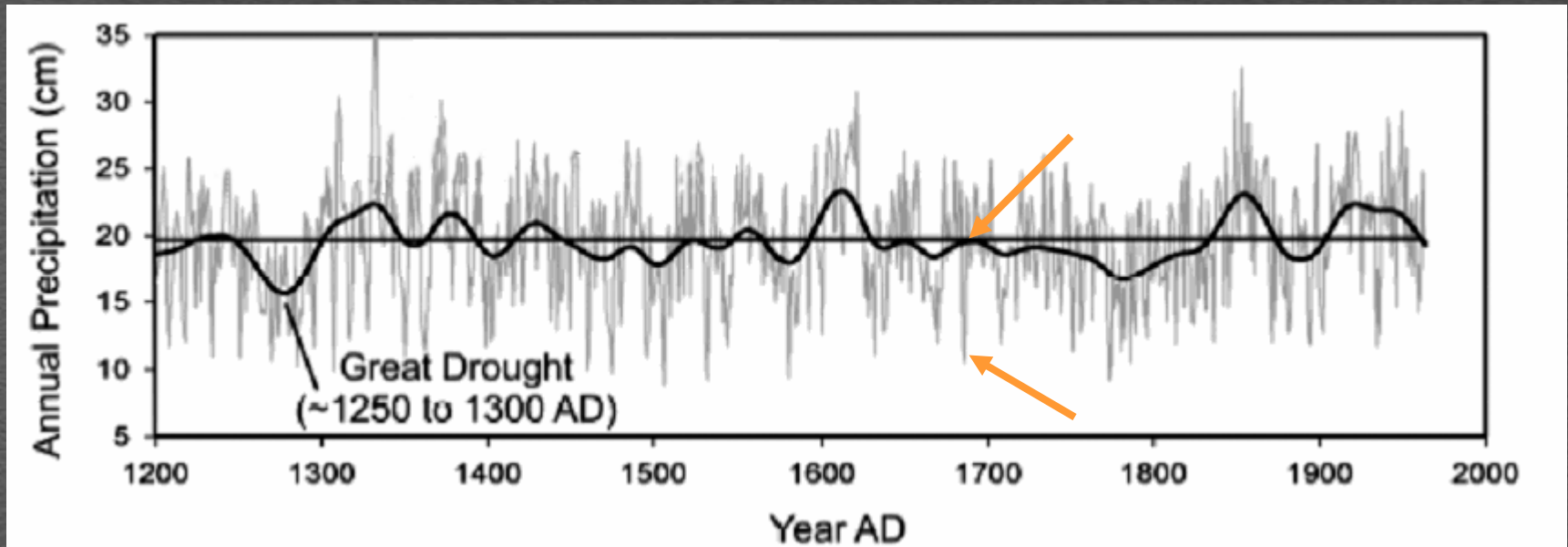
# Spatial resolutions



# Challenges with integrating data of different scales

- Different units chosen
  - Matching
  - converting
- Taking into account variability
- .....

# Variability at different scales

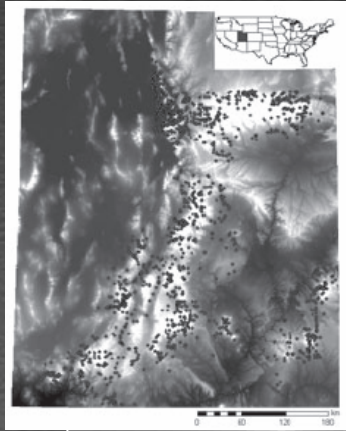


Gray et al., Ecology, in press.,  
modified with permission.

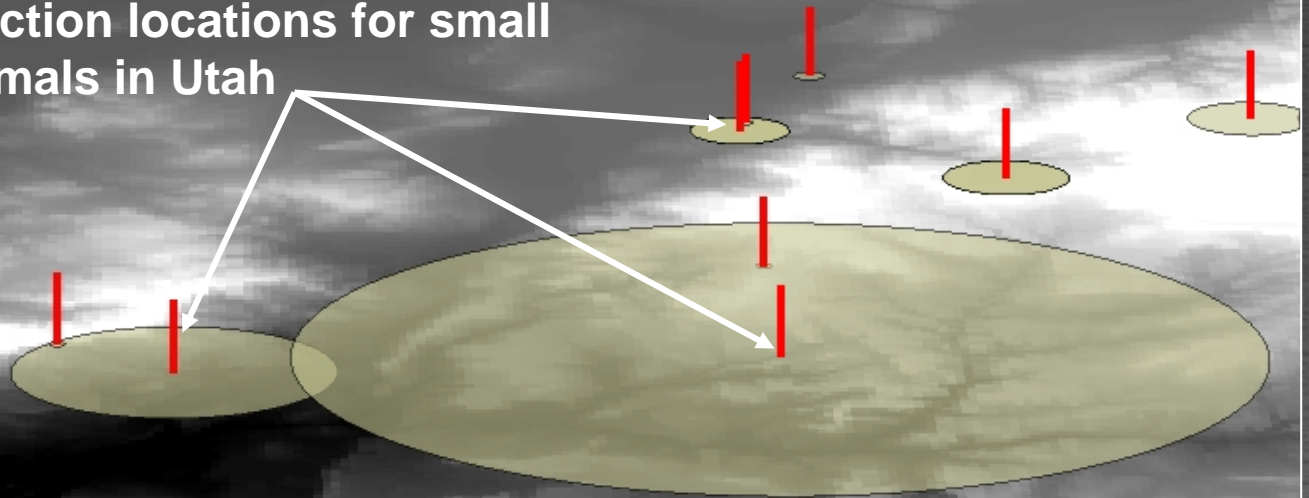
# Challenges with integrating data of different scales, continued

- Different units chosen
  - Matching
  - Converting
- Taking into account variability
- Taking into account uncertainty

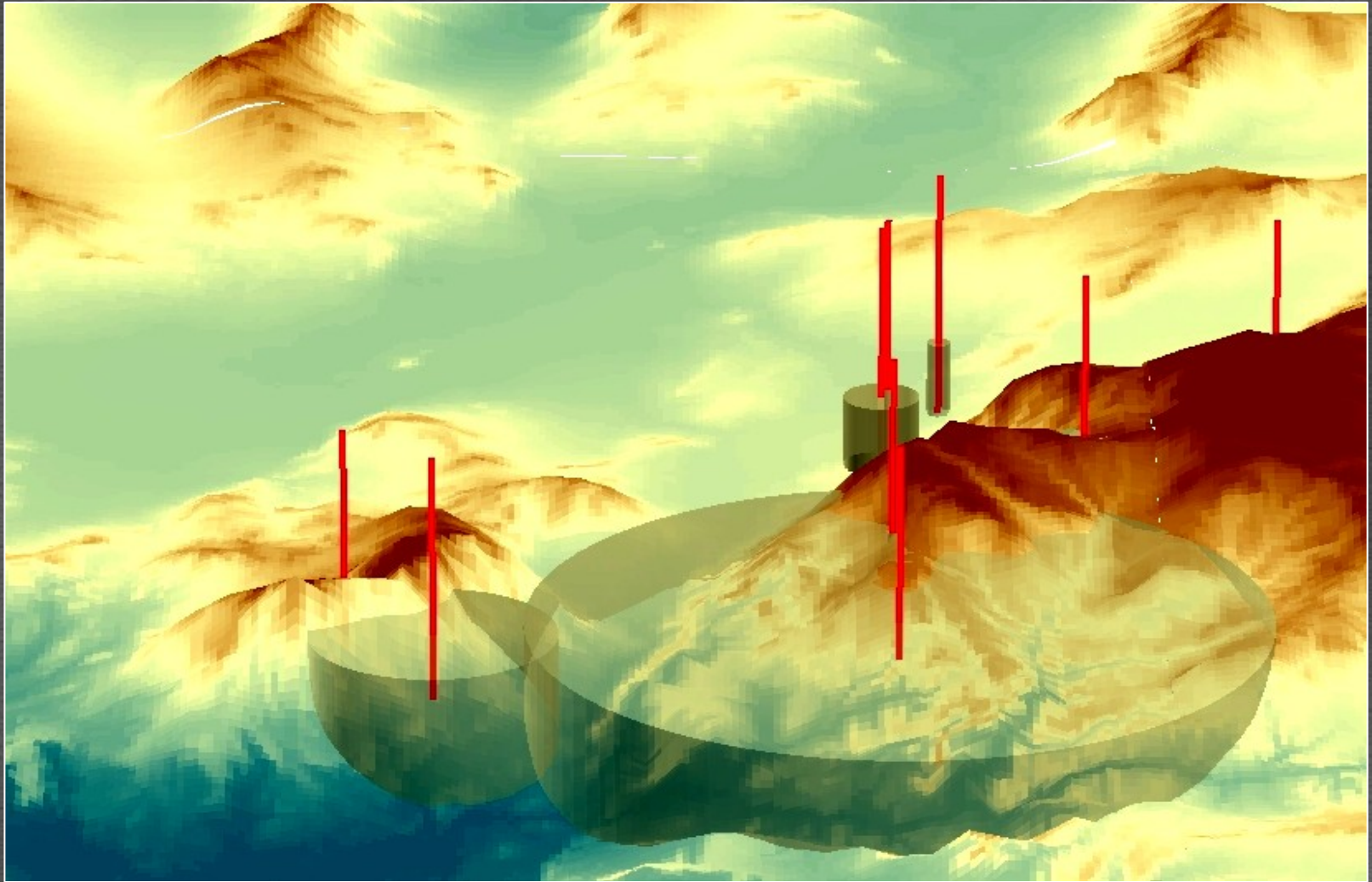
# Spatial Uncertainty



Collection locations for small Mammals in Utah



# Uncertainty in 3 dimensions



# Solutions to the various challenges in biological informatics

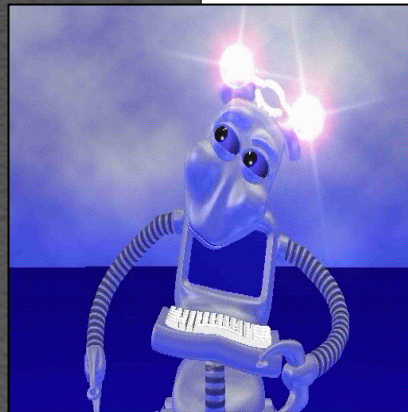
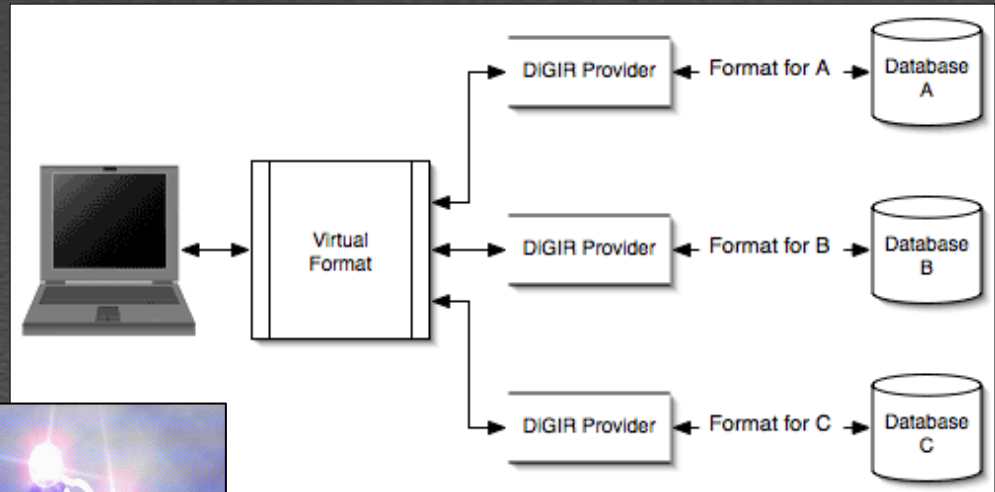


Chip Clark, NMNH



# Data Linkages

- New query protocols, e.g., DiGIR
- Intelligent search agents, e.g., BioBot



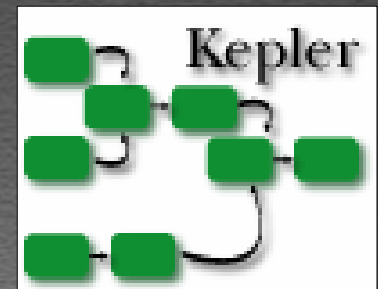
- Customization
- Multi-lingual interfaces
- Improved visualization

# Information describing the data

The screenshot shows the SMMS 3.1 metadata editor window. The title bar reads "SMMS 3.1 - [D:\Program Files\SMMS3\_1\Samples.mdb]". The menu bar includes "File", "Edit", "View", "Tools", and "Help". The toolbar contains various icons for editing and navigation. The main window is titled "Railroads of Washington State" and has a "Standard" style. The "General" tab is selected, showing the following fields:

Station	Railroads of Washington State
Point of Contact	Gordon Kennedy
Description	
Abstract	This data set is a linear depiction of railway mainline rights of way currently considered active by the Washington State Department of Transportation's Public Transportation and Rail Office. Some railroads included here
Purpose	This data set is used to present a small-scale cartographic rendition of active railroads in Washington State, and as a reference layer for geographic information systems at WSDOT.
Supplemental Information	
Access Constraints	none
Use Constraints	none
Data Set Credit	David Thompson, Washington State Department of Transportation
Native Data Set Environment	Arc/Info coverage, version 7.1.2 for NT 4.0, and ESRI shape file

Metadata



# Terminologies and research protocols

**ITIS Standard Report Page: Galerella pulverulenta - Microsoft Internet Explorer**

Address: [http://www.itis.usda.gov:8080/servlet/SingleRpt/SingleRpt?search\\_topic=TSN&search\\_value=622009](http://www.itis.usda.gov:8080/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=622009)

**ITIS Report**

Go to Print Version

**Galerella pulverulenta (Wagner, 1839)**  
Taxonomic Serial No.: 622009

**Taxonomy and Nomenclature**

Kingdom:	Animalia
Taxonomic Rank:	Species
Synonym(s):	
Common Name(s):	Cape gray mongoose [English]

**Taxonomic Status:**  
Current Standing: valid

**Data Quality Indicators:**  
Record Credibility Rating: verified - standards met

**Taxonomic Hierarchy**

Kingdom	Animalia -- Animal animals, animaux
Phylum	Chordata -- chordates, cordado, cordes
Subphylum	Vertebrata -- vertebrado, vertebrates, vertebres
Class	Mammalia Linnaeus, 1758 -- mamifero, mammals, mammiferes
Subclass	Theria Parker and Haswell, 1897
Infraclass	Eutheria Gill, 1872
Order	Carnivora Bowdich, 1821 -- cachorro do mato, carnivores, carnivores, carnivor, gato do mato, lontra
Suborder	Feliforma Kretzoi, 1945
Family	Viverridae Gray, 1821
Subfamily	Viverrinae Gray, 1821
Genus	Galerella Gray, 1855
Species	Galerella pulverulenta (Wagner, 1839) -- Cape gray mongoose

**References**

Expert(s):  
Expert: W. Christopher Woencraft  
Notes: Division of Natural Sciences, Bethel College, 1001 W. McKinley Ave., Mishawaka, IN 46545

Integrated Taxonomic Information System

**http://thesaurus.nbi.gov/SearchNBIIthesaurus/ - Microsoft Internet Explorer**

Address: <http://thesaurus.nbi.gov/SearchNBIIthesaurus/>

**NBII Home**

**Browse the CSA-NBII Biocomplexity Thesaurus**

The Biocomplexity Thesaurus was developed through a partnership between the NBII and CSA, a leading bibliographic database provider.

**Please enter a term to search in the CSA-NBII Thesaurus**

[Check Thesaurus](#)

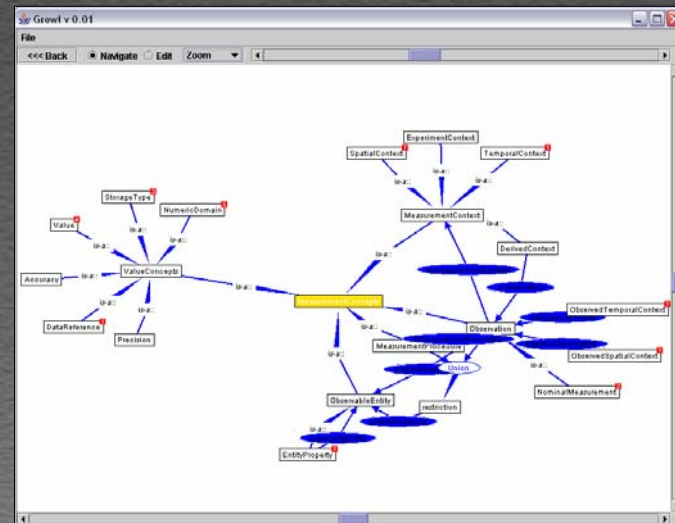
- Search on single term or a phrase. (e.g. biodiversity, invasive species)
- Search performs automatic stemming for prefixes and suffixes (e.g. entering *face* will yield cell surface, air-ice interface, face, and other variations; entering *log* will yield abiotic factors, biometrics, biodiversity, ectosymbiosis, etc.)
- [More about the Biocomplexity Thesaurus](#)
- [A Web Service](#) for the CSA-NBII Biocomplexity Thesaurus is now available.

**Data & Information Resources**

8 Hierarchical Menu Trees Created

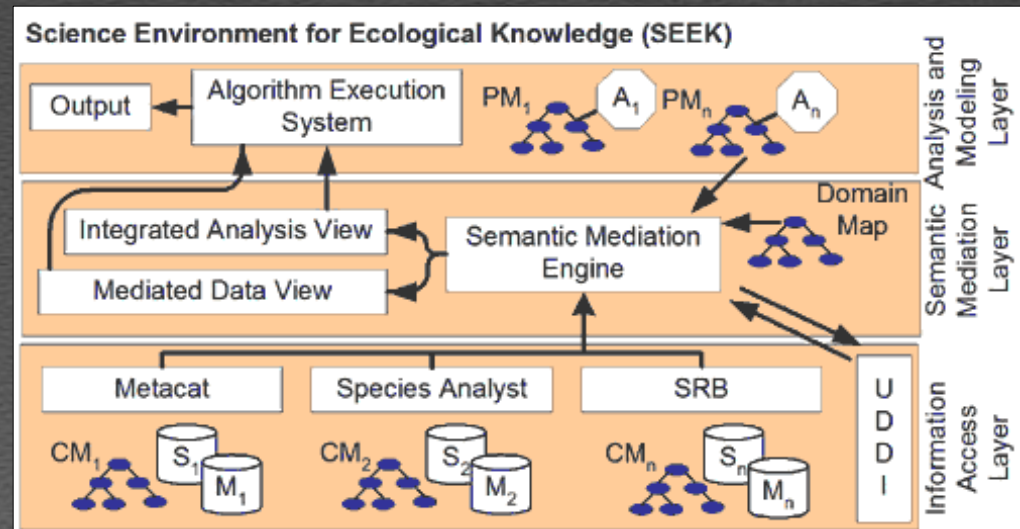
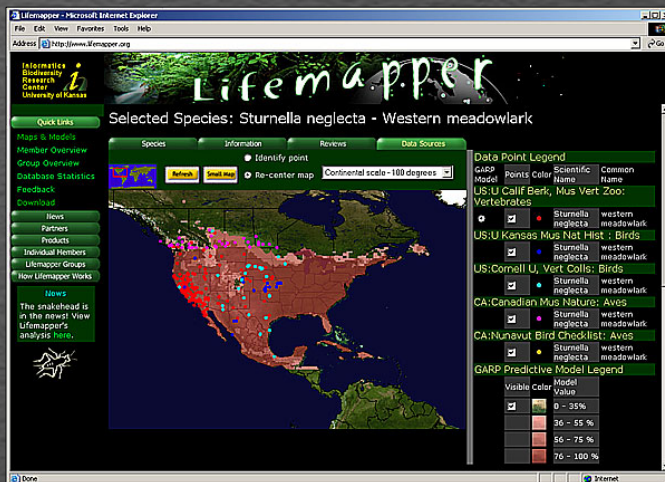
Biodiversity Complexity Thesaurus

Pacific  
Ecoinformatics  
and  
Computational  
Biology Lab



# Processing power and tools

- Mass computation capacity
- EcoGrid computing



# Training and education



**nbii** National Biological Information Infrastructure

[Metadata Home Page](#) | [Metadata Training](#) | [Metadata Tools](#)  
[Metadata Standards](#) | [NBII Metadata Clearinghouse](#)

### Metadata Workshop Calendar

2006

To register for a workshop, email Viv Hutchison [vhutchison@usgs.gov](mailto:vhutchison@usgs.gov)

- **January 25, 2006** Introduction to Metadata Workshop: Anchorage, AK. Presenter: Terry Giles, USGS/NBII **Venue:** USGS Alaska Science Center
- **January 26, 2006** Introduction to Metadata Workshop (2 half day presentations): Anchorage, AK. Presenter: Terry Giles, USGS/NBII **Venue:** North Pacific Research Board
- **March 20-21, 2006** Introduction to Metadata Workshop: Knoxville, TN. Presenter: Terry Giles, USGS/NBII **Venue:** University of Tennessee, Knoxville
- **April 24-28, 2006** Introduction to Metadata Workshop: Sacramento, CA. Presenter: Terry Giles, USGS/NBII **Venue:** Bureau of Reclamation

2005

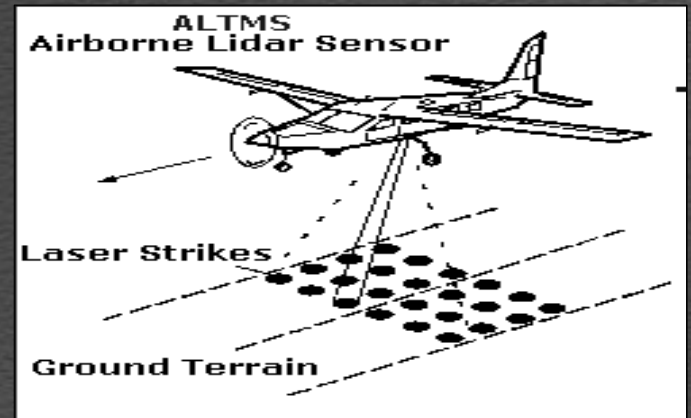
- **March 3-4, 2005** Introduction to Metadata Workshop: Palisades, NY. Presenter: Terry Giles, USGS/NBII **Venue:** Center for International Earth Science Information Network (CIESIN)
- **May 18-19, 2005** Introduction to Metadata Workshop: Durango, CO. Presenter: Kirsten Larsen, Northern Arizona University/NBII-SWIN **Venue:** TBD
- **June 16-17, 2005** Introduction to Metadata Workshop: Arlington, VA. Presenter: Terry Giles, USGS/NBII and Lynn Kutner, NatureServe **Venue:** NatureServe
- **June 28-30, 2005** Train-the-Trainer Metadata Workshop: Denver, CO. Presenter: Mike Moeller,

**IGERT**  
Integrative Graduate Education and Research Traineeship  
NATIONAL RECRUITMENT PROGRAM

# New Data Collection Techniques



Field computers



LIDAR



Smart Dust



UAV

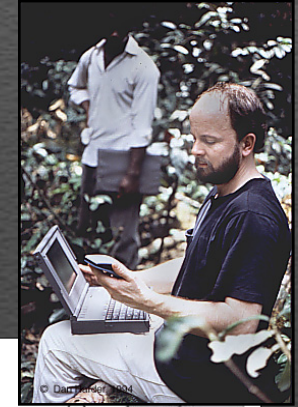
# New opportunities for Science Disciplines

- Interdisciplinary collaborations allows investigation into new questions
- More communication
  - Highlighting important results
  - Illustrating interlocking biological systems
  - Providing ecological forecasts
- Sharing data
  - More valuable to more people
  - Longer-lived data





# The Vision for the Future

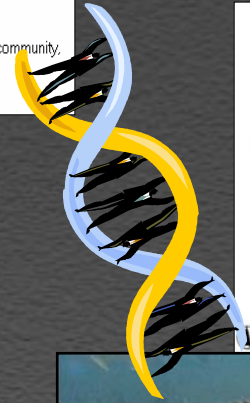


**GLOBAL INVASIVE SPECIES INFORMATION NETWORK**

**Global Invasive Species Information Network**

Welcome to the Discussion Forum on Implementation of a Global Invasive Species Information Network (GISIN)!

**Problems?** If you are experiencing difficulty in accessing/viewing the components of the GISIN community, please email Liz at [esellers@usgs.gov](mailto:esellers@usgs.gov).



*“Research is conducted in virtual laboratories in which scientists and engineers can routinely perform their work without regard to physical location, interacting with colleagues, accessing instrumentation, sharing data and computation resources, and accessing information in digital libraries”*



*(President’s Committee of Advisers on Science and Technology, 1998)*





Thank you

Gladys Cotter  
Associate Chief Biologist for  
Information  
Biological Informatics Office, USGS  
[Gladys\\_Cotter@usgs.gov](mailto:Gladys_Cotter@usgs.gov)