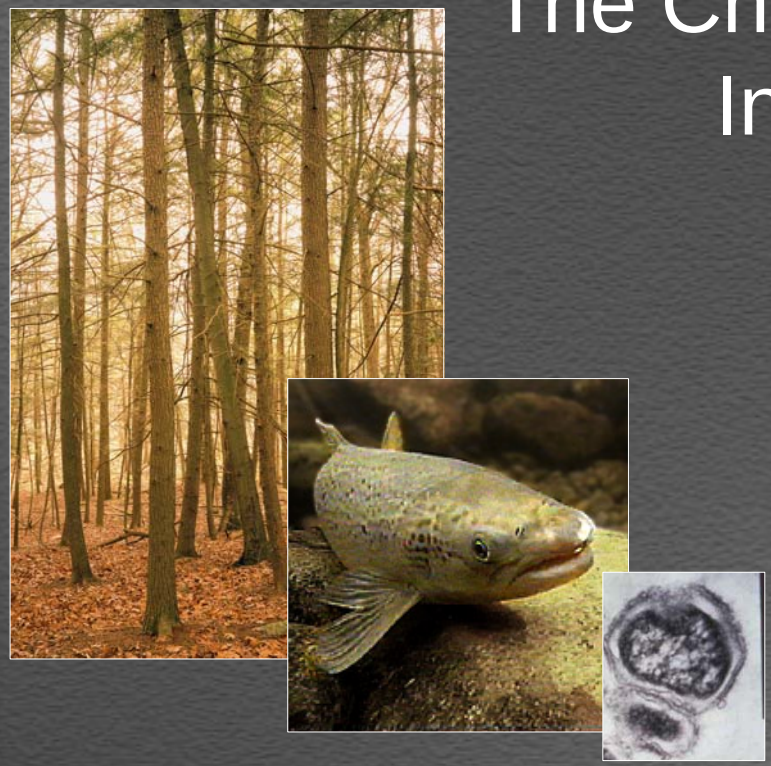
Biological Informatics:

The Challenge of Data Integration



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

Aquatic food webs

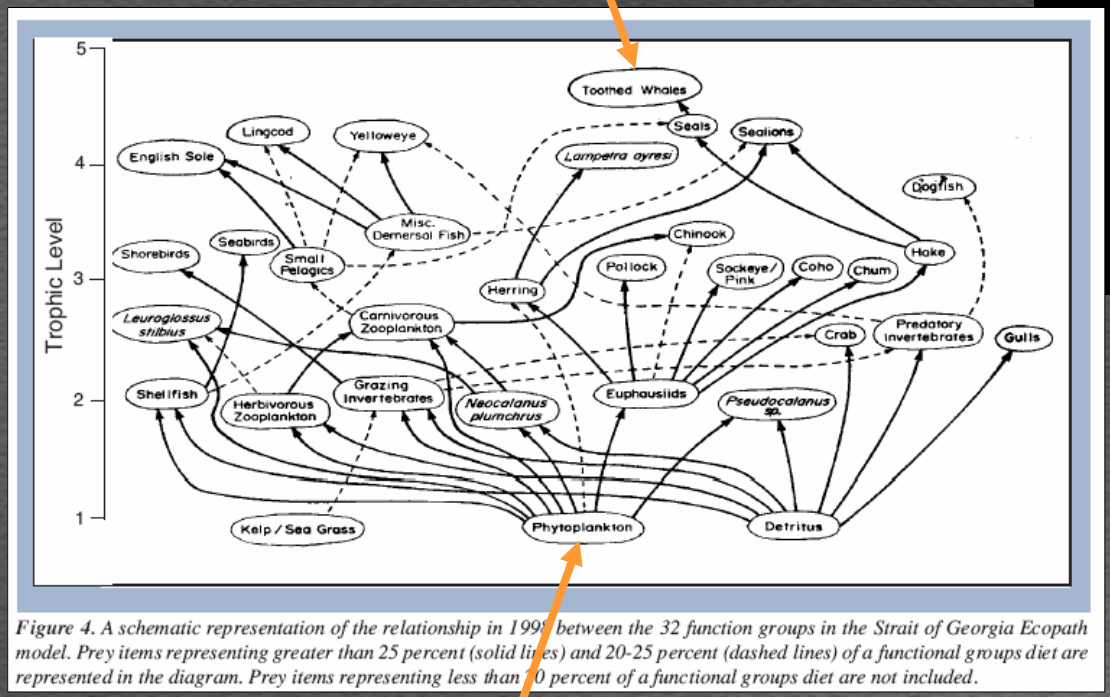
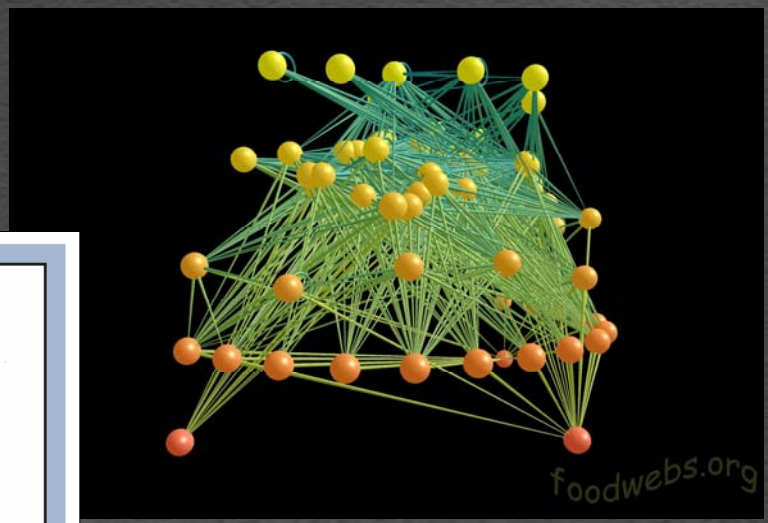


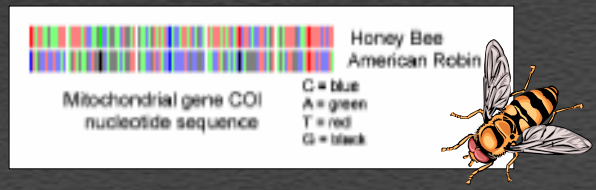
Figure 4. A schematic representation of the relationship in 199[illegible] between the 32 function groups in the Strait of Georgia Ecopath model. Prey items representing greater than 25 percent (solid lines) and 20-25 percent (dashed lines) of a functional groups diet are represented in the diagram. Prey items representing less than [illegible]0 percent of a functional groups diet are not included.

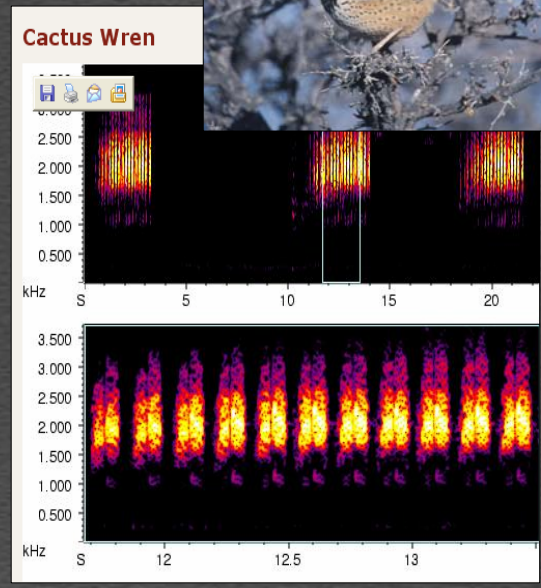
Beamish et al., 2003



Provided by the PEaCE Lab (www.foodwebs.org)

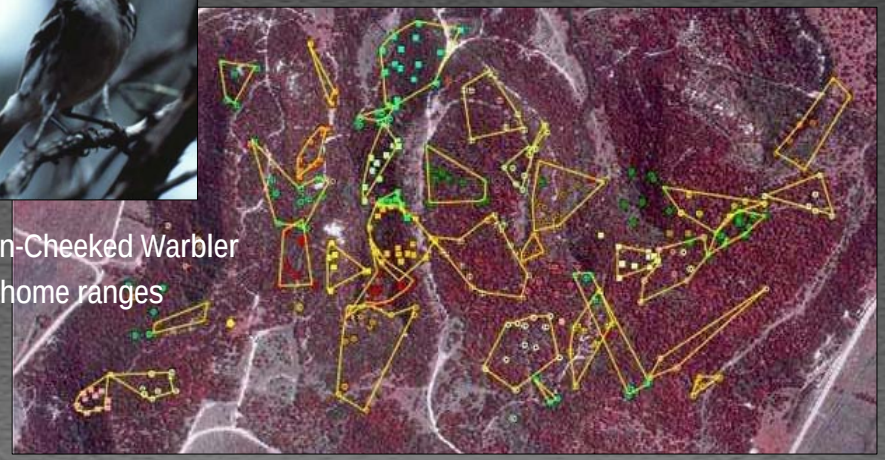
USGS





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

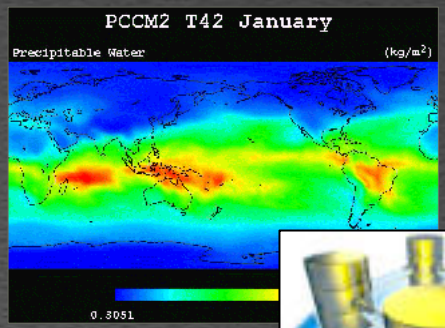
Golden-Cheeked Warbler
home ranges

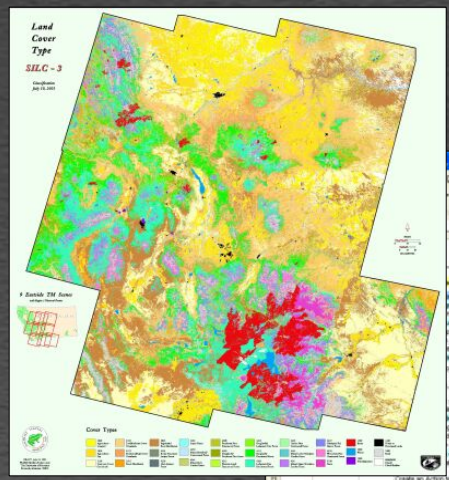


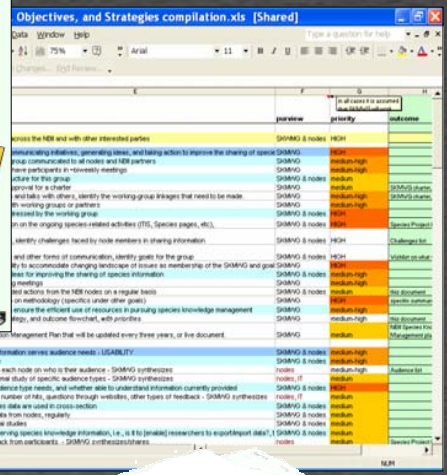
USGS

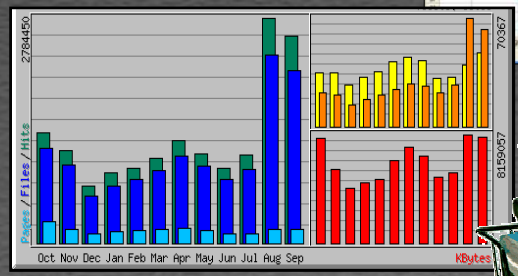
Biological information

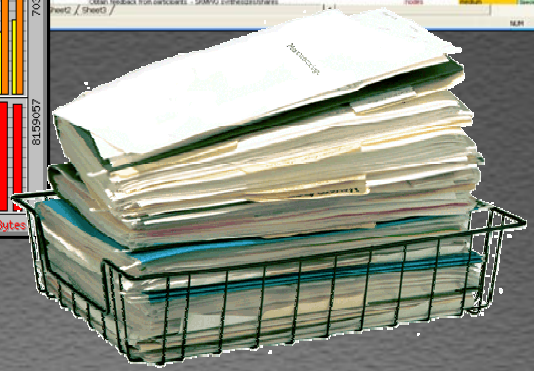
Data format diversity

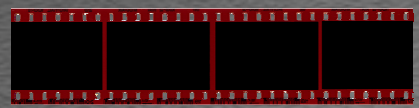












USGS

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.

USGS

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

…..

USGS

One Goal - Ecological Forecasting

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

University of Kansas

USGS

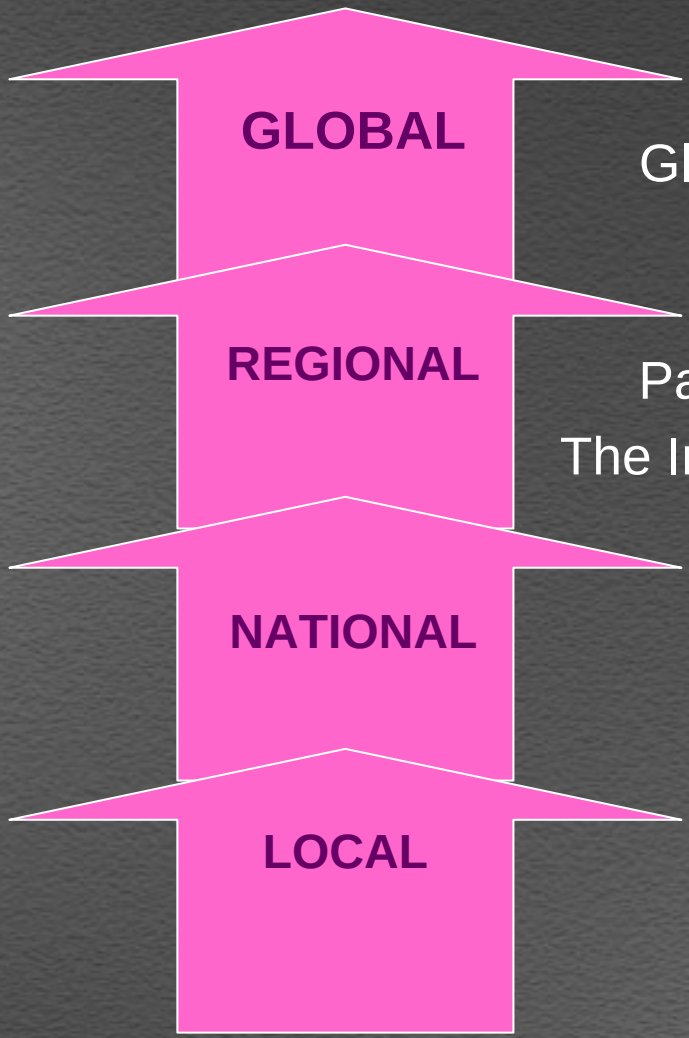
Data and Information Needs, continued

- 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.

USGS

Biological Informatics Systems



World Data Center for Biodiversity and Ecology (WDCBE), Global Biodiversity Information Facility (GBIF), Clearinghouse Mechanism (CHM)

Pacific Biodiversity Information Forum (PBIF), The Inter-American Biodiversity Information Network (IABIN)

NBII (US), CBIN (Canada), ERIN (Australia)

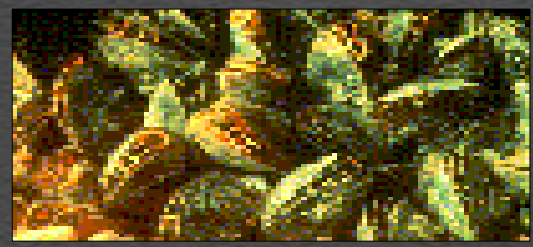
State Heritage Programs, GAP Analysis, Universities, County Park Information

USGS

NBII

The National Biological Information Infrastructure

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



www.nbii.gov

USGS

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.
- …..

USGS

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

CONABIO

USGS

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

USGS

Challenges, continued

Linking data of different scales within:

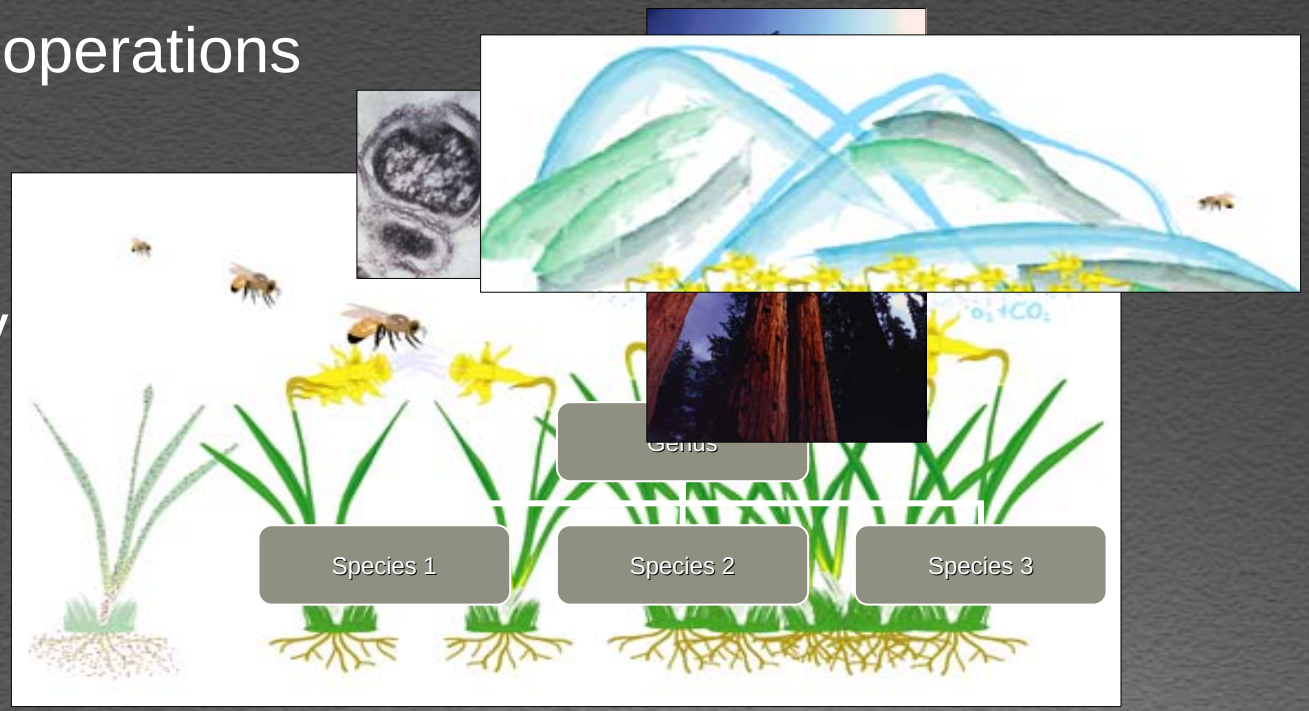
- Biology
- Space
- Time

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

USGS

Biological Scales

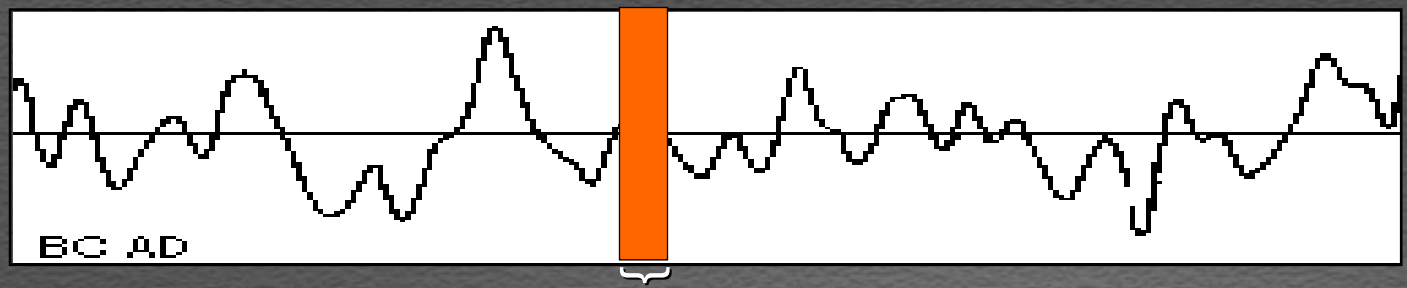
- Organism size
- Biological operations
- Taxonomy



USGS

Space and Time - Continuous dimensions

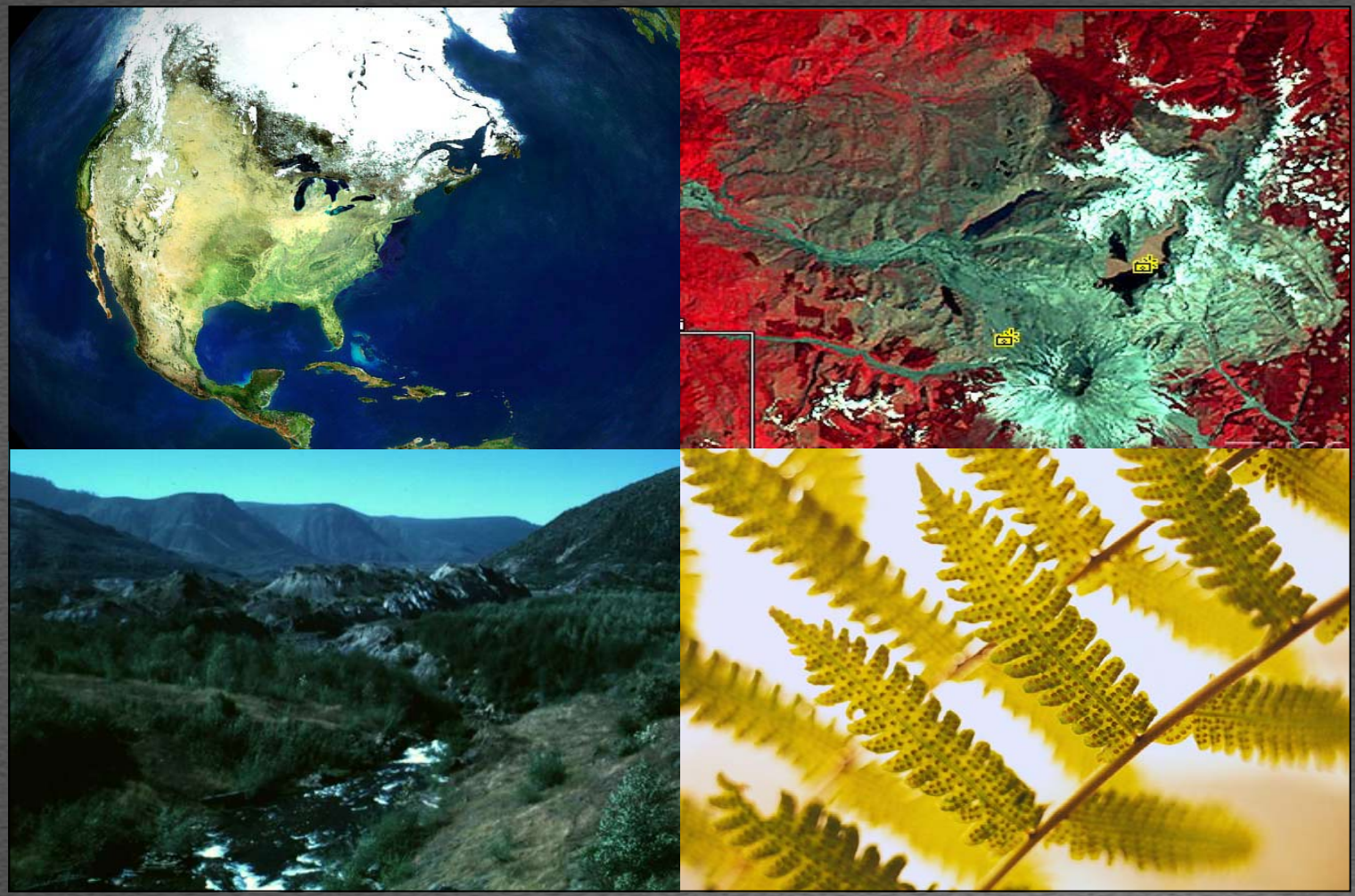
1 centimeter, or 10 millimeters



100 years

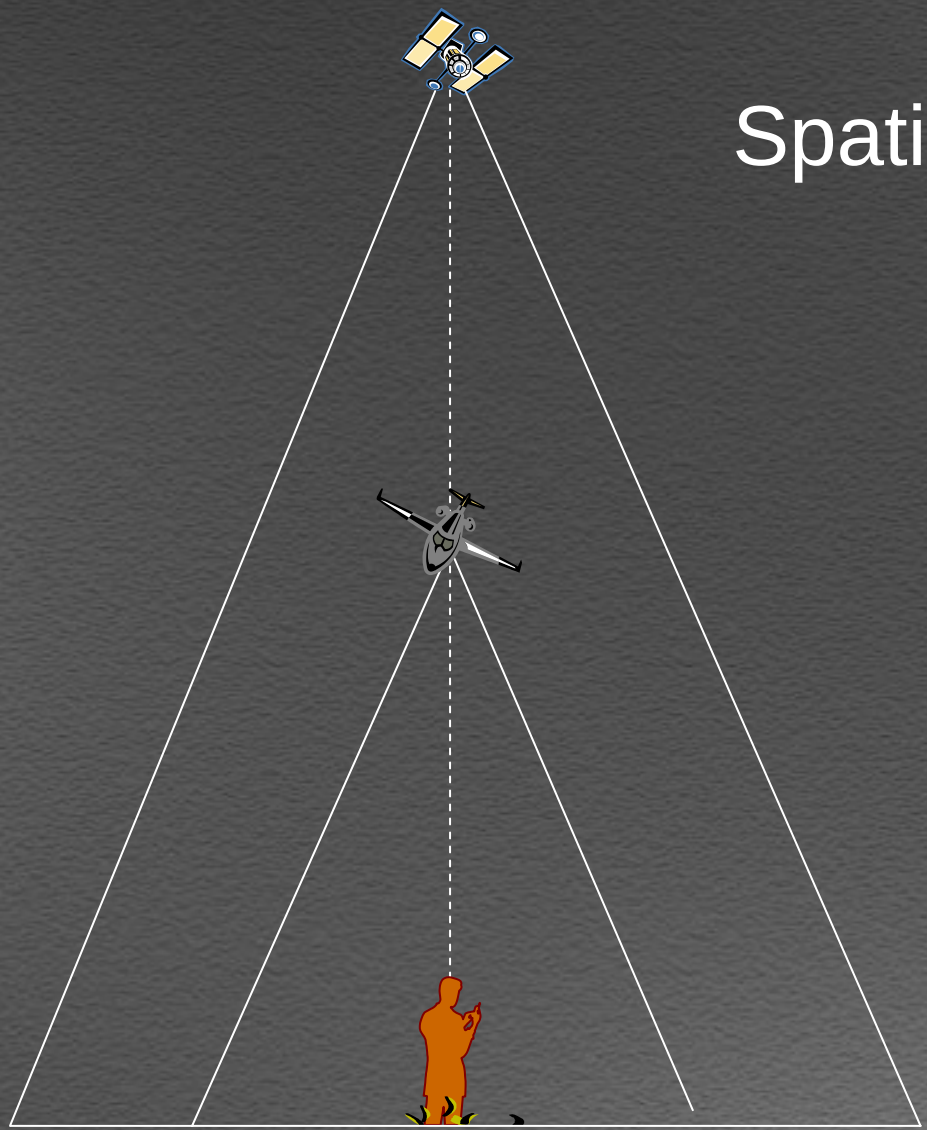
USGS

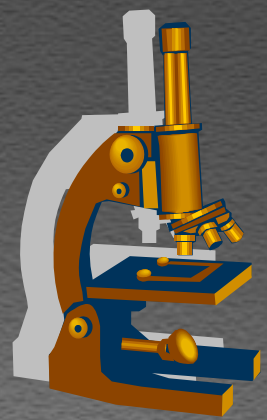
Different scales of observation



USGS

Spatial resolutions





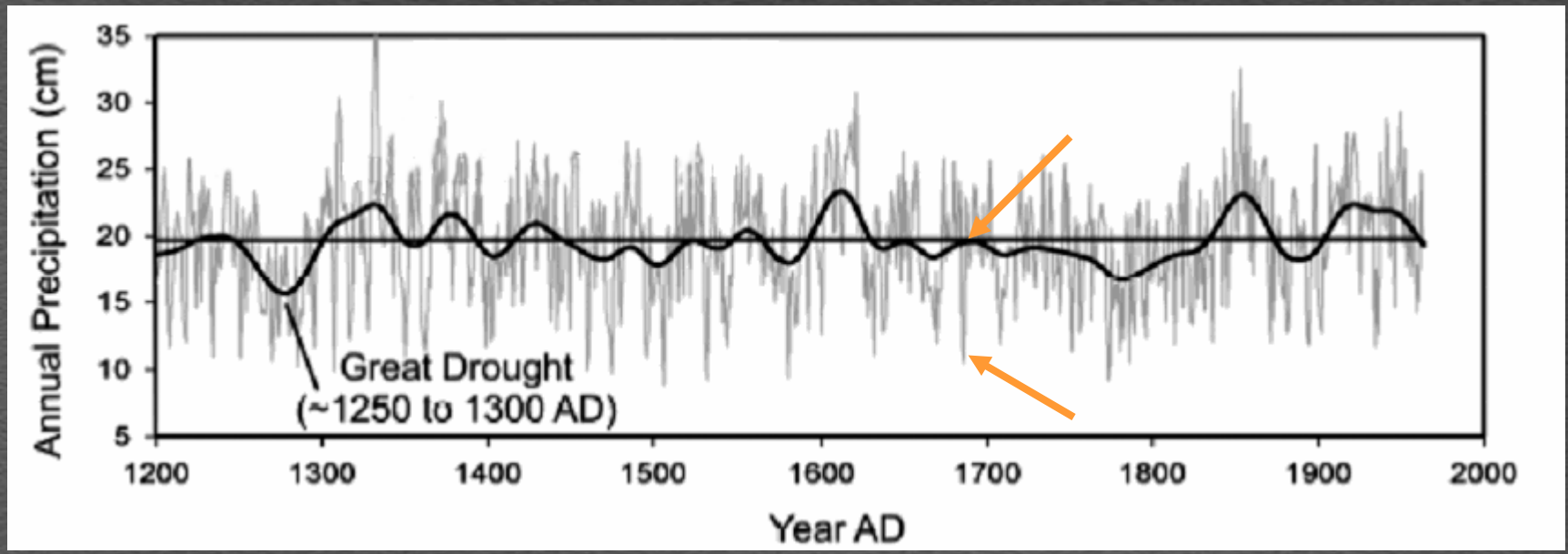
USGS

Challenges with integrating data of different scales

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

USGS

Variability at different scales



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

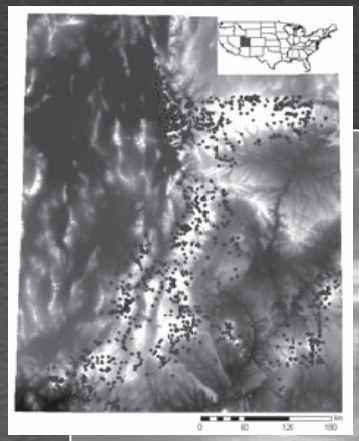
USGS

Challenges with integrating data of different scales, continued

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

USGS

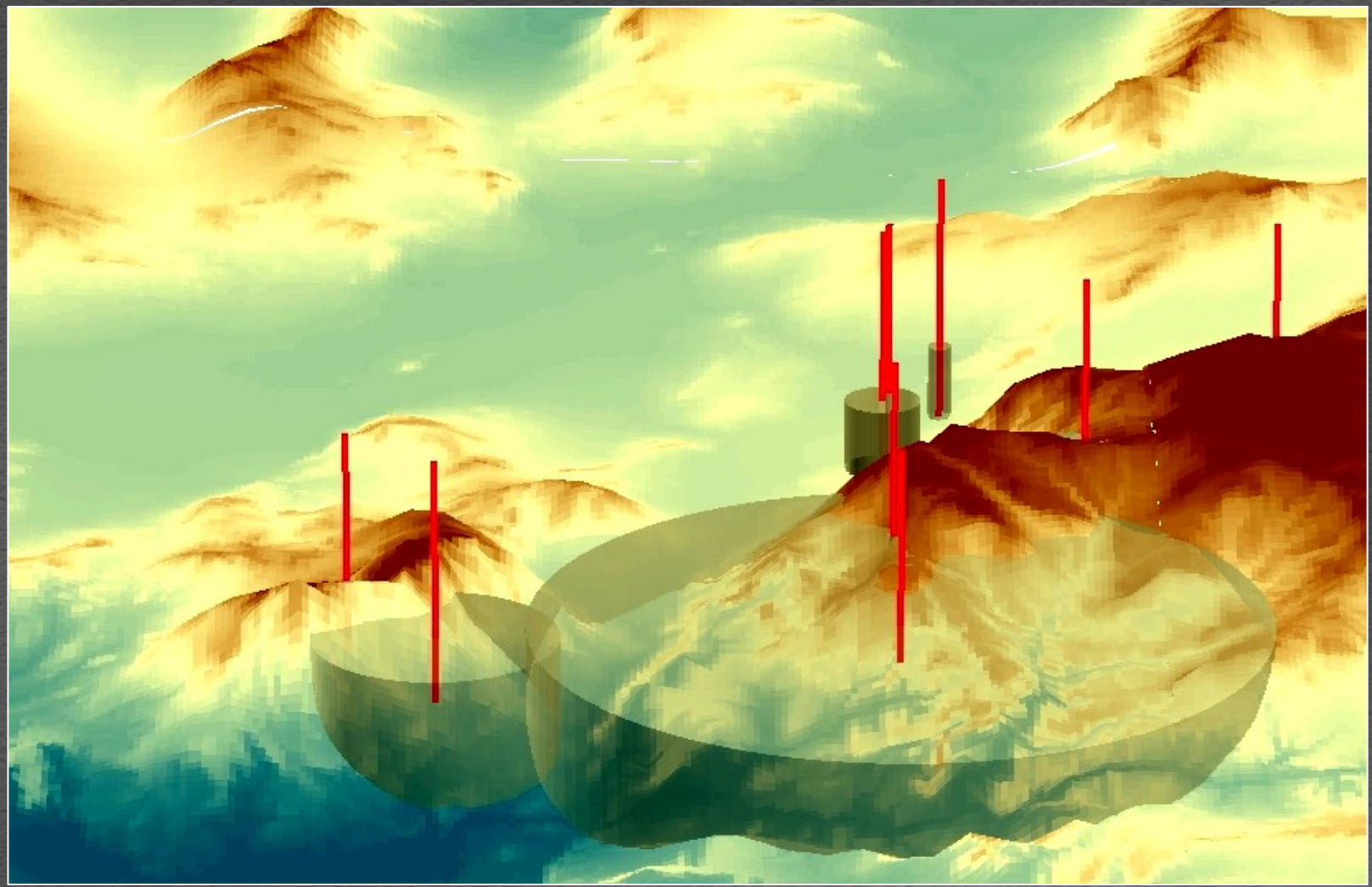
Spatial Uncertainty



USGS

Rowe RJ. 2005. Journal of Biogeography 32: 1883-1897

Uncertainty in 3 dimensions



USGS

Rowe RJ. 2005. Journal of Biogeography 32: 1883-1897

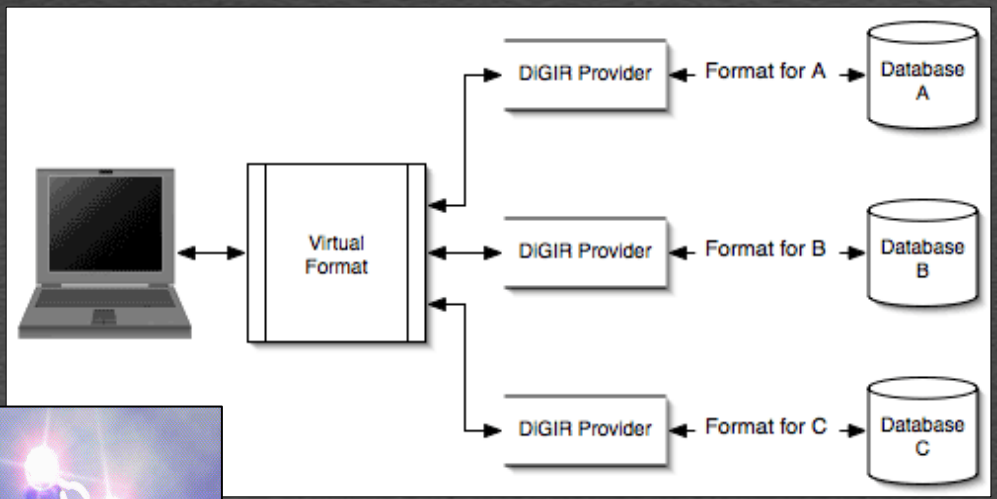
Solutions to the various challenges in biological informatics

Chip Clark, NMNH

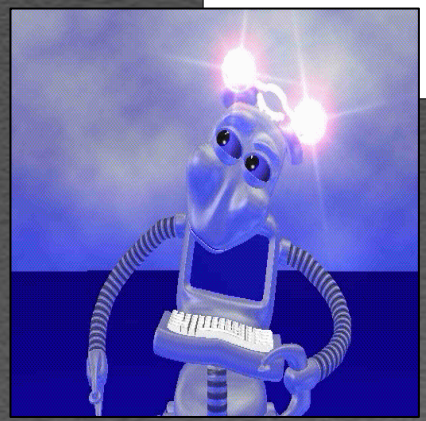
USGS

Data Linkages

- New query protocols,
  e.g., DiGIR



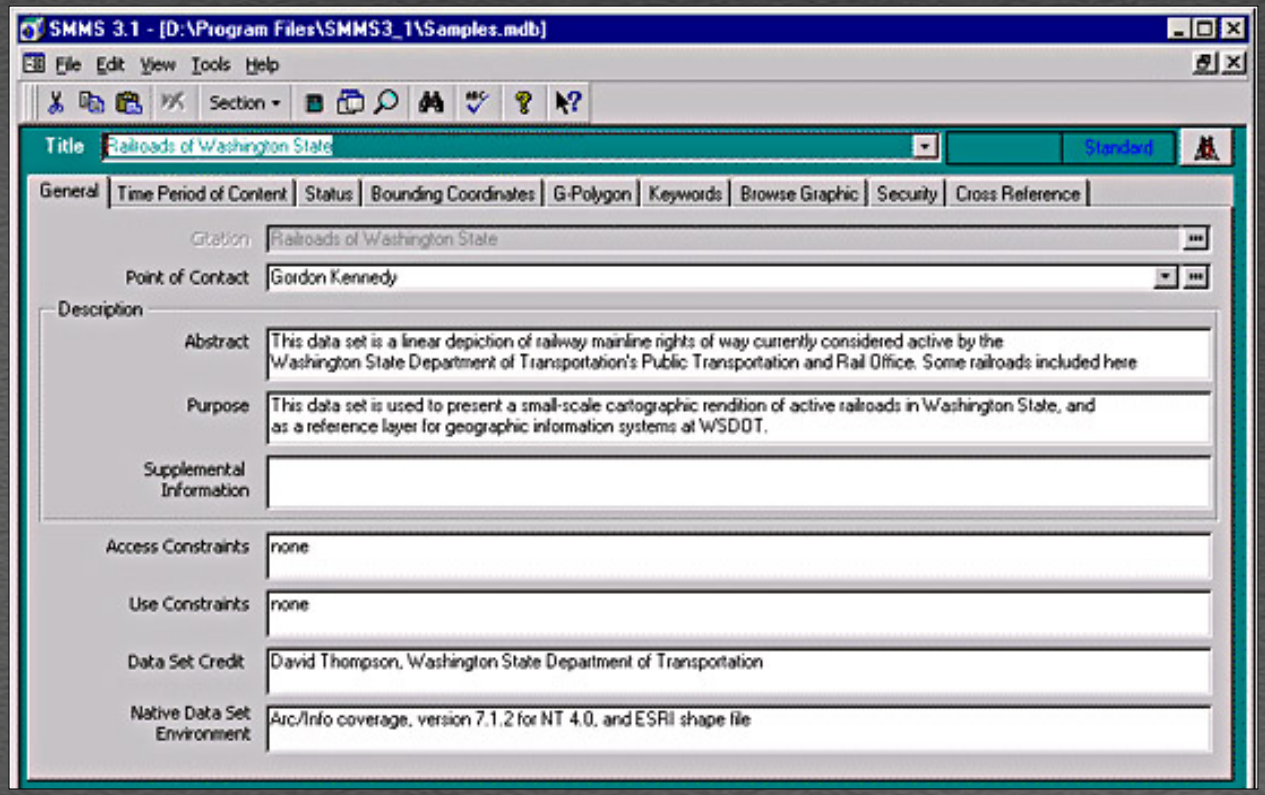
- Intelligent search agents,
  e.g., BioBot



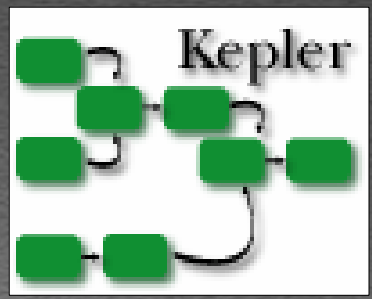
- Customization
- Multi-lingual interfaces
- Improved visualization

USGS

Information describing the data

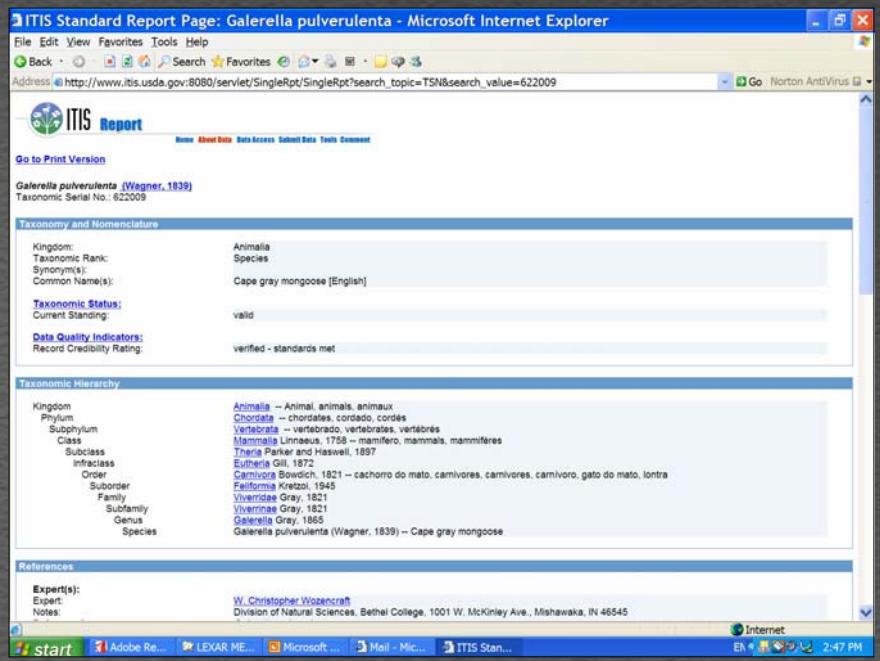


Metadata

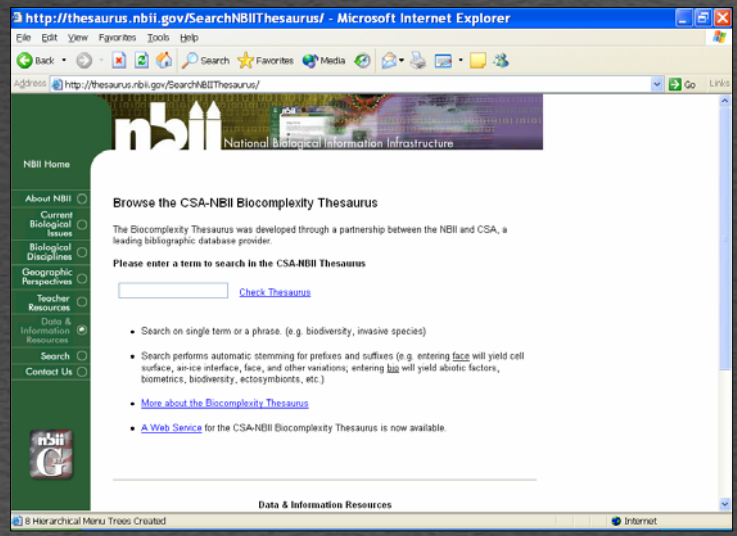


USGS

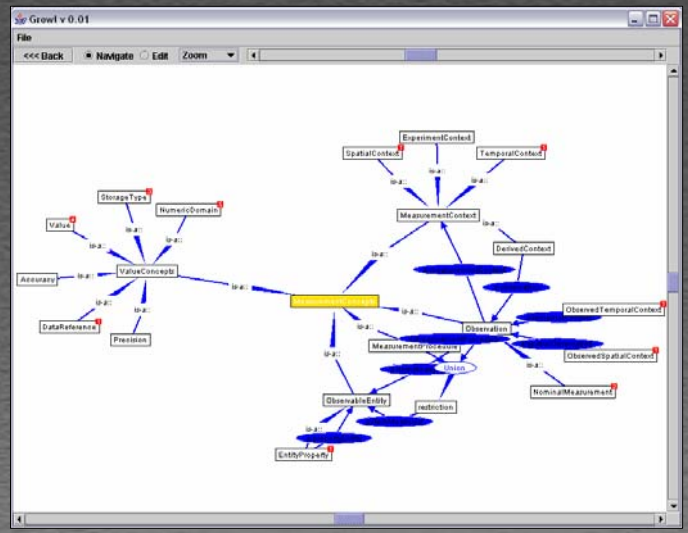
Terminologies and research protocols



Integrated Taxonomic Information System



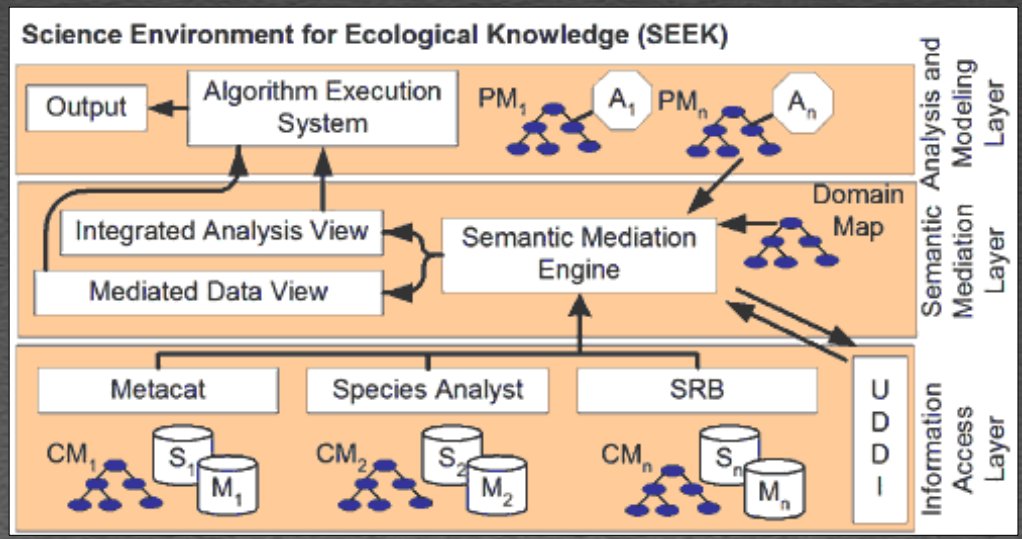
Biodiversity Complexity Thesaurus



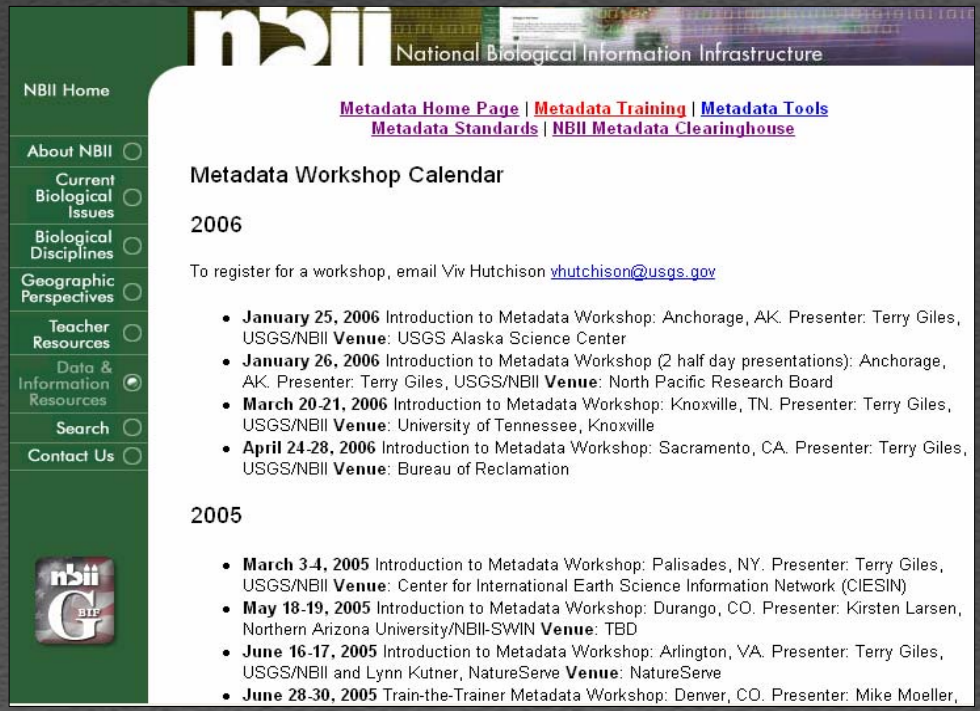
Pacific Ecoinformatics and Computational Biology Lab

Processing power and tools

- Mass computation capacity
- EcoGrid computing



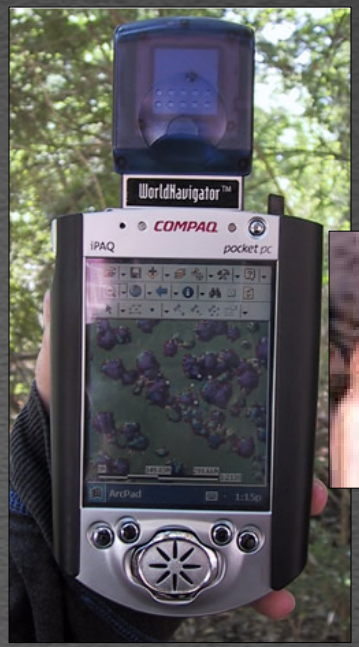
Training and education



IGERT
Integrative Graduate Education and Research Traineeship
NATIONAL RECRUITMENT PROGRAM

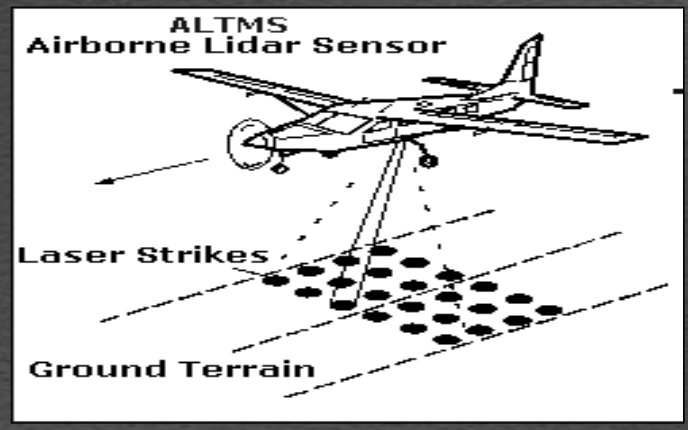
USGS

New Data Collection Techniques



Field computers

Smart Dust



LIDAR

UAV

USGS

New opportunities for Science Disciplines

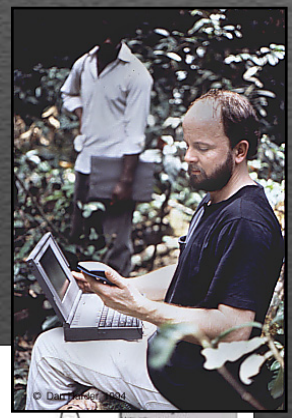
- Interdisciplinary collaborations allows investigation into new questions

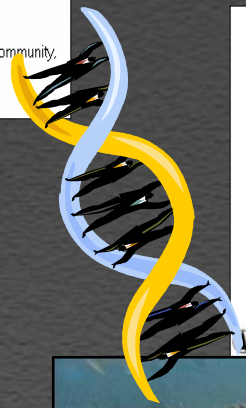
- More communication
  - Highlighting important results
  - Illustrating interlocking biological systems
  - Providing ecological forecasts

- Sharing data
  - More valuable to more people
  - Longer-lived data

USGS

The Vision for the Future





"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)

USGS

Thank you

Gladys Cotter
Associate Chief Biologist for Information
Biological Informatics Office, USGS
Gladys_Cotter@usgs.gov

USGS