Using Web-Accessible Large Data Sets for Teaching & Student Research

Tom Langen
15 April 2011

Overview

Priority research problems in the natural and social sciences require integration & synthesis of multiple sources of data, at appropriate temporal and spatial scales. Funding agencies now require that data collected by monitoring and research projects be made available to the public. Other entities have taken as part of their mission data archiving or serving as a clearing-house of diffuse data sources on some theme, and they make these data accessible to the public through the web.

The web brings to a student's desktop computer real-time data acquisition (sensors, monitoring programs), and large-scale data appropriate for answering significant questions in the natural and social sciences. Students are already accustomed to finding and downloading entertainment and educational materials from the web. Guided inquiry or problem-based learning exercises that task teams of students to test hypotheses using real data can prepare them for the one of the emerging essential research skills: data mining and data set integration using existing data from diffuse sources.

These data provide a tremendous opportunity to introduce into undergraduate ecology courses inquiry-based activities that would previously been impossible to do. There are at least four challenges to using large data sets, including (1) locating the data, (2) figuring out how to access and download it in format amenable to analysis, (3) knowing how to determine what the data actually provide: *the metadata*, and (4) understanding data authorship and ownership – how to ethically use the data and credit the providers of it.

Pedagogical Goals

- ✓ Develop science process skills via inquiry or project-based learning using multiple sources of real data
- ✓ Develop the capacity to frame multiple hypotheses and test their predictions by integrating multiple types of data
- ✓ Foster collaborative research skills
- ✓ Improve data analysis skills
- ✓ Improve ability to locate, evaluate, and use publicly-available data
- ✓ Experience research using data at appropriately large spatial and temporal scales

Skill Development

- ✓ Geographic fluency (geographic literacy)
- ✓ Fluency at using and problem-solving in standard and idiosyncratic data archive formats
- ✓ Efficient search methods for data location and retrieval
- ✓ Evaluation of metadata
- ✓ Evaluation of data quality and relevance
- ✓ Ethical methods of using the data and crediting the data authors

Several draft student-active exercises appropriate for undergraduate ecology students can be downloaded at https://groups.nceas.ucsb.edu/big-data/front-page ; contact info for authors is also available at this site. These activities are drafts produced by the NCEAS Distributed Seminar *Engaging Undergraduate Students In Ecological Investigations Using Large, Public Datasets*, sponsored by the *National Center for Ecological Assessment & Synthesis* (NCEAS), *Neon Inc.*, and the *Ecological Society of America* (ESA). These activities will eventually appear on *Teaching Issues & Experiments in Ecology* (TIEE http://tiee.esa.org/), and ESA online publication of peer-reviewed educational resources.

SOME RECOMMENDED DATA SOURCES

GOVERNMENT AGENCY DATA PORTALS

National Atlas http://www.nationalatlas.gov/
Geospatial data on the environment, economy, and people of the US.

US Department of Agriculture Census of Agricultural Data http://www.agcensus.usda.gov/
Authoritative data on all aspects of agriculture in the US.

Center for Disease Control & Prevention Data & Statistics http://www.cdc.gov/datastatistics/
Comprehensive data on all aspects of disease epidemiology.

USGS Water Data for the Nation http://waterdata.usgs.gov/nwis
Hydrological and water-quality data from across the US.

EPA Wadeable Stream Assessment http://water.epa.gov/type/rsl/monitoring/streamsurvey/web_data.cfm
Downloadable data on a national survey of stream water quality.

USGS Survey Disease Maps http://diseasemaps.usgs.gov/index.html
US County-scale maps of incidence patterns of various mosquito-vectored diseases.
The Multi-resolution Land Characteristics Consortium (MRLC) National Land Cover Database
http://www.mrlc.gov/
Land cover or land use, canopy cover, and impermeable surface area of the entire US, at a resolution of 30 m x 30 m, based on remote sensing data from satellite imagery.

US Fish & Wildlife Service National Wetlands Inventory http://www.fws.gov/wetlands/
Wetlands greater than 1 acre are mapped and classified throughout the US, Puerto Rico and US territories. Data can be examined using the Wetland Mapper http://www.fws.gov/wetlands/Data/Mapper.html and then downloaded for use by a GIS application, or can by inspected directly using Google Earth http://www.fws.gov/wetlands/Data/GoogleEarth.html .

USDA Forest Inventory and Analysis National Program http://fia.fs.fed.us/ , **Forest Inventory Data Online (FIDO)** http://fia.fs.fed.us/tools-data/default.asp
Highly-detailed periodic surveys of forest composition at sites throughout the US.

US Geological Survey http://www.usgs.gov/
Reports, data analysis, maps, and raw data on a diversity of topics related to environmental science, including biodiversity and emerging diseases.

NOAA National Climate Data Center http://www.ncdc.noaa.gov/oa/ncdc.html
Extensive data archives of climate data, including paleoclimate.

ENVIRONMENTAL DATA CLEARINGHOUSES

Ecotrends http://www.ecotrends.info/EcoTrends/
Data archive and data visualization tools for ecological data at sites distributed around the US.

NASA Global Change Master Directory http://gcmd.nasa.gov/index.html
Data on all aspects of global change, includes data on climate, land use, biodiversity and human dimensions.

Oak Ridge National Laboratory **Distributed Active Archive Center for Biogeochemical Dynamics** (ORNL DAAC) http://daac.ornl.gov/index.shtml
A NASA-sponsored source for biogeochemical and ecological data and models useful in environmental research.

Pole to Pole Ecological Research Lattice of Sites (P2ERLS) http://www.p2erls.net/
Portal to research stations and research networks, including their data archives.

Weatherspark http://weatherspark.com/
Visualized time-series data on local climate at sites around the globe.

Long Term Ecological Research (LTER) Network http://www.lternet.edu/
Network of research stations that have standardized monitoring programs as well as site-specific research. Sites are mandated to make data publicly available on the web.

BIODIVERSITY DATA CLEARINGHOUSES /ARCHIVES

International Union for the Conservation of Nature (IUCN) Redlist http://www.iucnredlist.org/
Searchable list of the world's threatened and endangered plants and animal species on the IUCN Redlist.

Conservation International Global Biodiversity Hotspots http://www.biodiversityhotspots.org/Pages/default.aspx
Detailed data on the attributes and threats to the world's global biodiversity hotspots.

National Biological Information Infrastructure http://www.nbii.gov/
Data archive and clearinghouse for biological data from the US. Also provides standards for metadata.

Biological Inventories of the World's Protected Areas
http://www.ice.ucdavis.edu/bioinventory/bioinventory.html
Searchable species occurrence records and species lists for over 1,400 protected areas around the globe.

Global Biodiversity Information Facility http://data.gbif.org/welcome.htm
An enormous clearinghouse of biodiversity data .

Global Population Dynamics Data Base http://www3.imperial.ac.uk/cpb/research/patternsandprocesses/gpdd
5000 population size time series for 1400 species, most of which have at least ten years of data. There are data on the natural history of the organism and the location & method of sampling.

USGS Breeding Bird Survey http://www.pwrc.usgs.gov/BBS/
Breeding bird survey data back to 1966.

Bird Point Count Database. http://www.pwrc.usgs.gov/point/
Depository of bird point-count data from across the US.

Bird Studies Canada **Nature Counts** http://www.bsc-eoc.org/birdmon/default/main.jsp
Bird survey data archive for Canada, includes point counts and many other types of surveys.

Avian Knowledge Network http://www.avianknowledge.net/content/datasets
Archive of aggregated bird surveys from many organizations and studies across throughout the western hemisphere, including Latin America.

NatureServe http://www.natureserve.org/getData/index.jsp
Data on species of plants and animals in the Western Hemisphere, including detailed range maps.

RESEARCH PROJECT DATA ARCHIVES

Dryad http://datadryad.org/
Data archives for bioscience data from peer-reviewed journal articles from a large consortium of journals.

Ecological Society of America (ESA) Data Registry http://esapubs.org/archive/archive_D.htm
Archive of ecological and environmental data from ESA publications.

National Center for Ecological Assessment & Synthesis (NCEAS) Data Repository
http://knb.ecoinformatics.org/knb/style/skins/nceas/index.jsp
Data archive of contributed data sets of all types of ecological data.

NCEAS Scientific Computing Database http://www.nceas.ucsb.edu/scicomp/
Clearinghouse of climatological, geospatial, and other data. Also has shareware software for analysis.