Introduction to Field Station Databases

John Porter

University of Virginia

Roadmap

- Why do we need field station databases?
- Challenges for Ecological Databases
- Evolving a Database
- Software Tools and Hardware
- Partnerships – sharing the load

WHY have Scientific Databases?

- Improvement of data quality
  - multiple users provides multiple opportunities for detecting and correcting problems in data
- Cost
  - data costs less to save than to collect again
  - with environmental data, often data cannot be collected again at any cost

WHY have Scientific Databases?

- Environmental Policy and Management
  - environmental policy decisions require data that are regional or national, but most ecological data is collected at smaller scales
  - numerous Federal initiatives
    - NEON – National Ecological Observatory Network
    - NII - National Information Infrastructure
    - FGDC - Federal Geographic Data Committee

WHY have Scientific Databases?

- **New Science**
  - Long Term
    - long-term studies depend on databases to retain project history
  - Synthesis
    - use of data for a purpose other than which it was collected
  - Integrated, multidisciplinary projects
    - depend on databases to facilitate sharing of data

Why Have Field Station Databases?

- Not all field station needs are science-based – field stations need databases for:
  - Scheduling and Tracking Facility Use
  - Reports and Publications
  - Billing & Payments

Attracting Researchers Which do you choose?

- Field Station A
  - Beautiful mountain forest setting
  - Modern Laboratories
- Field Station B
  - Beautiful mountain forest setting
  - Modern Laboratories
  - Real-time Climate and Meteorological Data
  - Biodiversity Data
  - Soils Data
  - Topographic Data

The Field Station Information Challenge

- Can we make accurate information available to field station users and operators:
  - in ways they can **locate** the information they need?
  - with information in forms they can readily **use**?

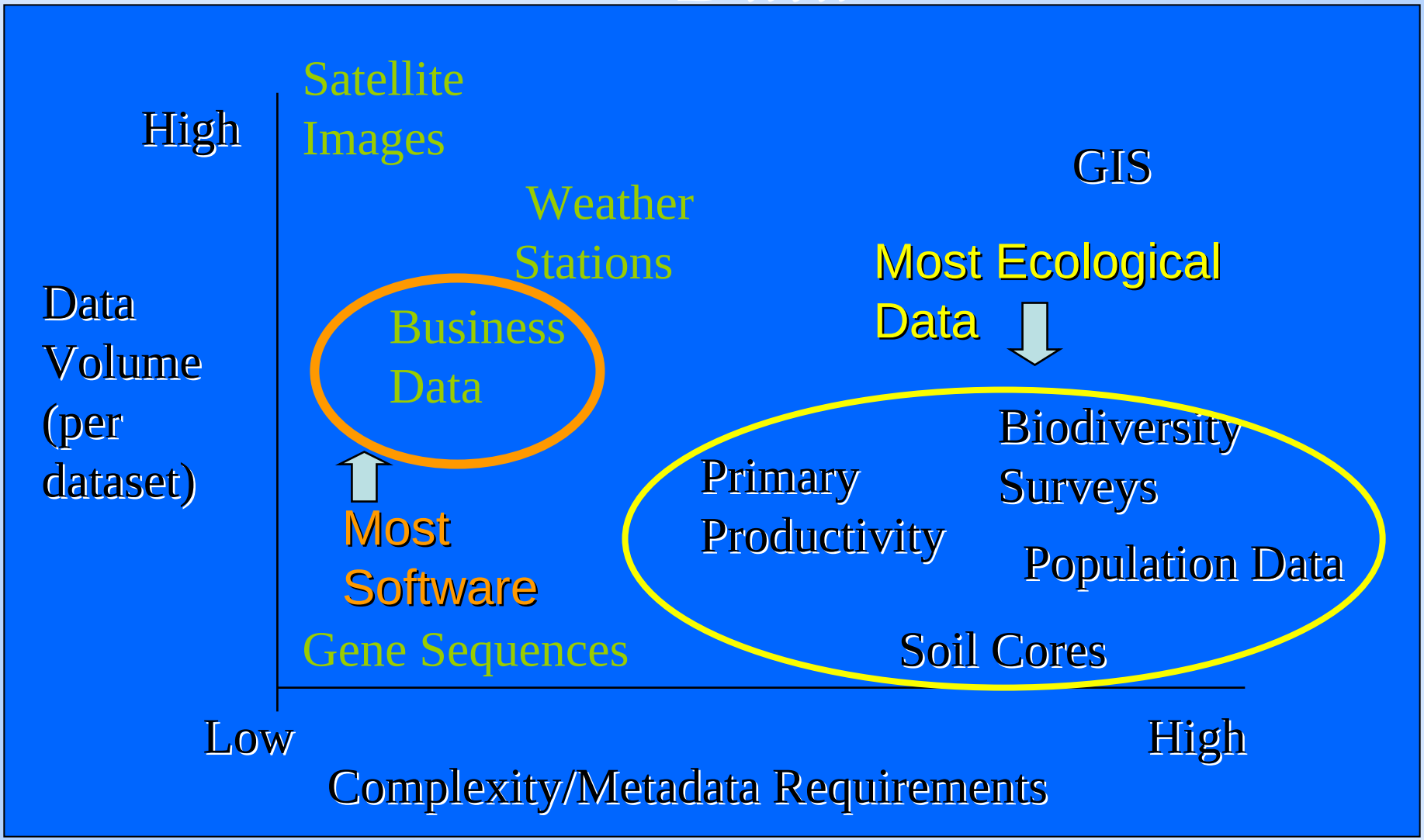
Challenges for Scientific Databases

- Long-term perspective
  - without databases, most data do not outlive project that collected them
  - goal: data that is **accessible** and **interpretable** 20-years in the future
    - technological - need persistent media that does not become technologically obsolete
    - contextual - need to capture context of data collection
    - semantic - terms need to be well-defined

Challenges for Scientific Databases

- Deal with Diversity
  - science means asking NEW questions
    - new kinds of queries
  - scientific data is heterogeneous and diverse
  - scientific users have different backgrounds and goals
  - the user community for a given database will be dynamic

Characteristics of Ecological Data



Challenges

- Resources
  - Equipment
- Resources
  - Operational expenses
- Resources
  - Personnel

Evolving a Database

- Development of a database is an evolutionary process
- Implement system based on current priorities - but think ahead!
- Seek scalable solutions
  - avoid bottlenecks
  - adding the 1000th piece of data should be as easy as adding the first (or easier)

Developing a Database - Questions to Ask

- Why is this database NEEDED?
- Who will be the USERS of the database?
- What types of QUESTIONS should the database be able to answer?
- What INCENTIVES will be available for data providers?

Organization for Tropical Studies

where science and nature converge

Español | Search | Tropiweb

La Selva Biological Station

Location	at the confluence of two major rivers in the Caribbean lowlands of northern Costa Rica, La Selva comprises 1,600 hectares (3,900 acres) of tropical wet forests and disturbed lands. It averages 4 m (over 13 feet!) of rainfall that is spread rather evenly throughout the year.
Principal Biome Main Communities	Located within the tropical and premontane wet forest, the Station has about 73% of its area under primary tropical rain forest.
History	La Selva was originally established in 1954 by Dr. Leslie Holdridge, as a farm dedicated to experimentation on mixed plantations for the improvement of natural resources management. It was purchased in 1968 by the Organization for Tropical Studies and declared a private biological reserve and station. Since then, it has become one of the most important sites in the world for research on tropical rain forest. Over 240 scientific papers are published yearly from research conducted at the site.

Learn More about La Selva Biological Station...

Sample Web Page Contents

- About– a brief description of the organization and its mission
- Directory – contact information for participants etc.
- Datasets
  - File-based
  - Query-based
  - Online GIS
- Bibliography /library
- Visitor Tools – info, application forms
- News
- Calendar of Events

Virginia Coast Reserve LTER – IM Milestones

- 1988- **DBASE III system**
- 1990- GIS Lab Established
- 1990 - Data Access Policy
- 1992 - Electronic Mail Calendar
- 1992 - Gopher Information Server
- 1993 - WWW Server
- 1994 - Online Research Summaries
- 1995 - **Web-based Personnel Directory**
- 1996 - Automated System for Research Summaries
- 1996 - **Biodiversity Database**
- 1997 - **Web form-based Information Management Tools**
- 2002 – Wireless Internet connection to island field site
- 2004 – **EML 2.1 Metadata**
- 2005 – EML-based analysis tools
- 2005 – **Web page in Content Management System**

Software

Web pages and DBMS are complimentary – what one does well the other does badly

Web Pages

- Good
  - Easy Interface
  - Flexible formats
- Bad
  - Static
  - Cumbersome to update

Database Management Systems

- Good
  - Powerful manipulation of data
  - Easy updates
- Bad
  - User-hostile interface

Web page + DBMS

- Combining web pages and DBMS leads to new power for both
  - Easy to use forms-based interfaces
  - Dynamic generation of web pages using the most current data
    - Web page updates need not come just from the webmaster
  - Multiple web pages can share the same data

Commonly Used Types of Software

- Input and Analysis tools
- Information sharing tools – WWW
- Database Management Systems (DBMS)

Input and Analysis

Spreadsheets

- Good
  - Widely used, easy to learn for simple graphical and statistical analyses
  - Commonly already installed on most computers
- Bad
  - Can encourage “bad practices” – create data that can’t easily be used
  - Poor support for sophisticated analyses
  - Lack of auditability – hard to “back track” how data were manipulated

Statistical Packages

- Examples: SAS, SPSS, Statistica, R etc.
- Good
  - Powerful analysis tools
  - Auditable: Can store programs – fully document details of analysis
- Bad
  - Harder to learn
  - Less common on computers
  - Can be expensive

Other Input

- DBMS – Database Management Systems

Database Management System (DBMS) Types

- **Filesystem-based**
  - simple
  - inefficient
  - few capabilities
- Hierarchical
  - phylogenetic structures
  - geographical images
- Network
  - very flexible
  - not widely used
- **Relational**
  - widely-used, mature
  - table-oriented
  - restricted range of structures
- Object-oriented
  - developing -few commercial implementations
  - diverse structures
  - extensible

DBMS Advantages and Disadvantages

- Advantages
  - additional capabilities
    - sorting
    - query
    - integrity checking
  - easy access to data
- Disadvantages
  - few graphical or statistical capabilities
  - proprietary formats may limit archival quality of data
  - require expertise and resources to administer

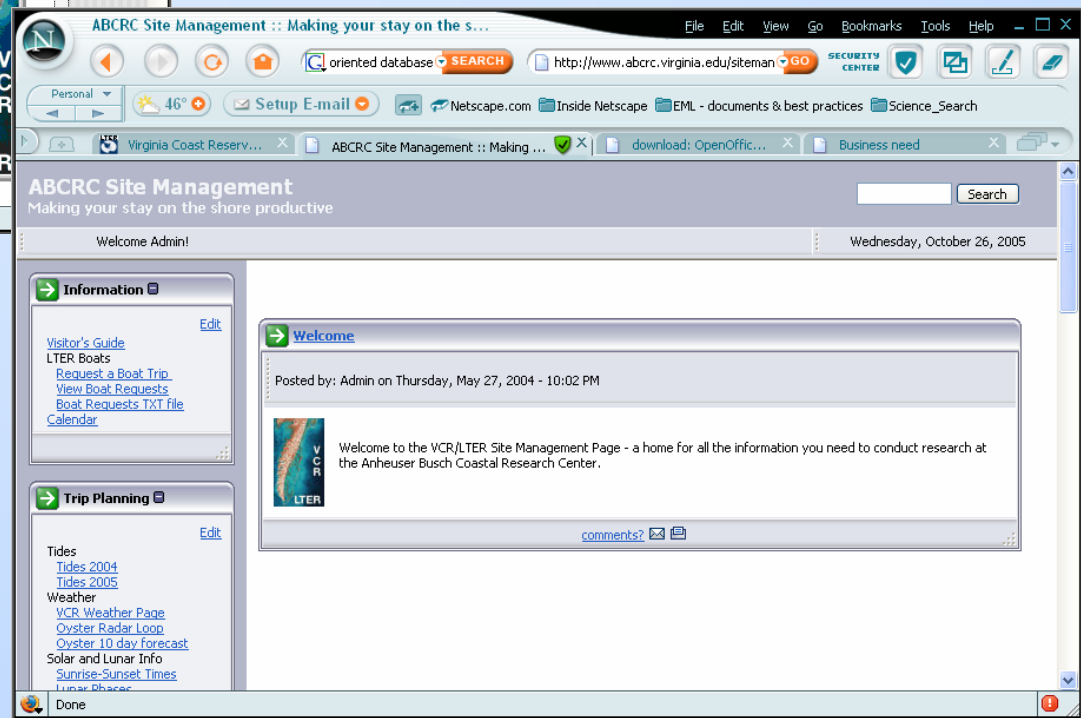
Database Management Systems

- Commercial Products
  - Microsoft ACCESS (part of Microsoft Office)
  - Microsoft SQLserver
  - Oracle
- Freeware
  - MySQL
  - PostgreSQL
  - MiniSQL

DBMS Backends

- Increasingly DBMS are being used as tools that support the “behind the scenes” activities in support of web sites
  - You may not interact with the database itself, but rather with a TOOL that interacts with the database
- Tools such as Content Management Systems (CMS) use programs that in turn use DBMS to perform their functions

Web Page Using CMS



The same content (stored in a DBMS) can appear different ways at the touch of a button.

You don't need to know very much about databases to administer a CMS site…

Information Sharing Tools

- WWW servers
  - Apache Web Server
    - Free
    - Based on open standards
    - Runs on PCs, Macintosh and Unix
  - Microsoft Web Server
    - Free, often distributed with Windows
    - Links to Microsoft tools
    - Proprietary - runs only under Windows

What are the “Best Software”?

- **SORRY! – there is no one list that is the correct answer for everyone!**
- A knowledgeable user, rather than the particular software used, controls what can be accomplished
- Costs
  - Cost of software
  - Cost of administration
  - Life-cycle costs
  - Costs of migration

Computer Systems

- UNIX/Linux
  - mature, full-functioned system
    - strong on multitasking
    - more reliable and robust
  - steep learning curve
  - lots of free software
  - software can be expensive
  - wide array of WWW tools

- PCs & Macs
  - rapid improvements in operating system design facilitate network access
  - software & hardware inexpensive
  - tools are more user-friendly
  - number of tools rapidly growing

Computer Systems

- UNIX/Linux
  - expensive
  - wide array of WWW tools
- PCs & Macs
  - number of tools rapidly growing

These days it JUST DOES NOT MATTER Which system you use. All are capable of supporting all but the most specialized needs.

ments
stem
e
ss
user-

Partnerships

Adding Features

- Sometimes the best way to add content and functionality to your web page is to use functionality provided by others
- Simple Example: Weather Underground "Stickies"

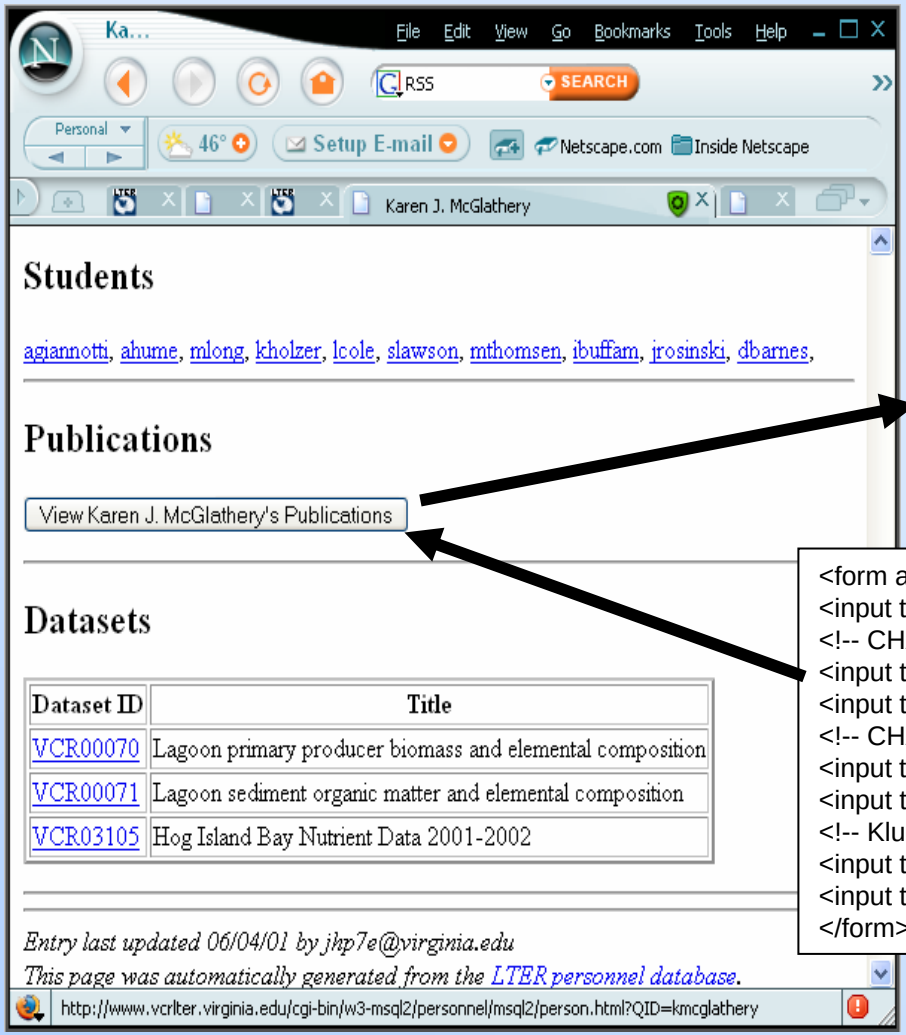
Adding Features

- Another source are Really Simple Syndication (RSS) feeds, where news items from other web sites appear on your web site
- For field stations there are more data-specific tools and partnerships that may be useful

Example: Publications

- The LTER Network operates a web-based publication database
- I would like to add a list of publications to the personnel directory listing for each researcher
- How? – use “hidden” form fields to automate the search for publications by a particular researcher

Sample Literature Search




```
<form action="http://search.lternet.edu/biblio/reftableadvancedselect.php" method=get>
<input type=hidden name=PARAMTYPE1 value=site>
<!-- CHANGE THE FOLLOWING VALUE TO YOUR SITE ID -->
<input type=hidden name=PARAMSTR1 value=VCR>
<input type=hidden name=PARAMCON2 value=AND>
<!-- CHANGE THE FOLLOWING VALUE TO THE AUTHOR ID -->
<input type=hidden  name=PARAMTYPE2 value="author">
<input type=hidden name=PARAMSTR2 value="McGlathery, Karen J.">
<!-- Kludge- collectionid should be provided automatically in later versions -->
<input type=hidden name=collectionid value=9999>
<input type=submit value="View Karen J. McGlathery's Publications"><br>
</form>
```

Features you can add:

- Data Catalog Search
  - Use the KNB Metacat system for managing field station data
  - You can customize the link used to initiate searches so that only YOUR data is returned automatically
- OBSF Personnel Database
  - Provided by the LTER Network Office
  - Again, queries can be customized so that only an individual station's data is searched

Adopting Tools & Standards

- Ecological Metadata Language
  - Morpho
  - Metacat
  - Open source, available

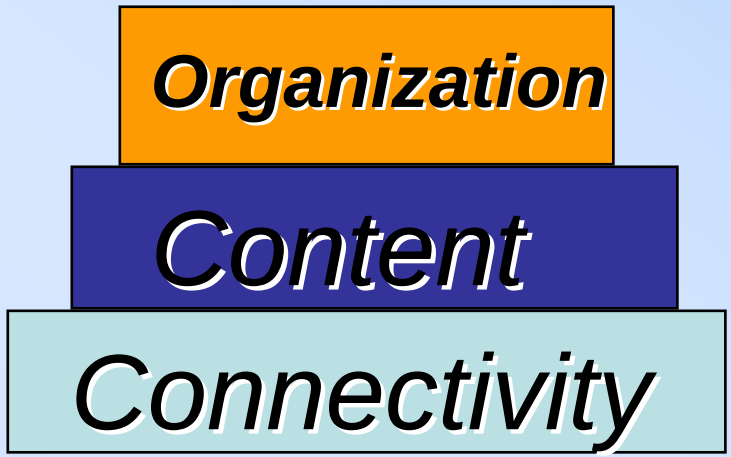
Future?

- There are potentially a large number of tools that could be transparently shared across the Organization of Biological Field Stations
  - The tools need to be identified
  - Funding needs to be obtained to pursue them
  - The community needs to be educated on how to include them

Final Thoughts

- Ecological databases are increasingly setting the boundaries for science itself
- Databases evolve, but they don't spontaneously generate

Database Building Blocks



Acknowledgements

- This material is based upon work supported by:
- The National Science Foundation under Grant Numbers: 0080381, 0129792 ,9980154, 0225676 and 0072909.
- Collaborators: University of New Mexico (Long Term Ecological Research Network Office), NCEAS (UC Santa Barbara), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research), University of Virginia, University of California Berkeley (Hastings Biological Station), University of Wisconsin
- The Andrew W. Mellon Foundation