Post ASM 2006, Report of Workshops I, II, and III[1]

Judy Cushing and Nicole Kaplan, organizers, Grasslands ANPP Data Integration Workshops

Description of the Grasslands Data Integration Project (GDI): Between 2003 and 2006, computer scientists and information managers (IM's) developed a prototype ANPP Grasslands Database. The work was funded by the LTER Network Office, which supported some IM travel, and indirectly by the National Science Foundation[2] as part of the Canopy *Databank* project efforts to develop informatics tools for field ecologists. We applied concepts from *Databank* to manage, synthesize, and analyze vegetation data from three LTER grassland sites (SGS, SEV, and JRN) and began to design the Grasslands DataIntegration project (GDI).

At a workshop at the LTER All Scientists' Meeting (ASM) in 2006, we presented preliminary results of the GDI to LTER Principal Investigator's (PI's) and ecologists, and consulted with them on its scientific value and direction. There was considerable enthusiasm to continue this work, both for the value that this particular database would hold to the ecological community and for the value of sustainable tools to integrate long term NPP data for cross-site analysis and synthesis in the LTER Network and beyond. Alan Knapp, Daniel Milchunas, Esteban Muldavin, and Debra Peters, LTER PI ecologists, expressed interest in the project, and suggested changes to our data transformation, documentation and analysis, as well as inclusion of other sites, e.g., Konza LTER and the Kruger National Park in South Africa.

Two post ASM workshops proposals were then funded by the LTER Network (with some matching funding from Cushing's NSF BDI grant) for the purpose of solving the problems of integrating long-term ANPP data from now five grassland sites, and documenting and publishing the work. Since 2006, we have continued the Grasslands Data Integration Project (GDI), a successful collaboration between ecologists, computer scientists, and information managers to address data integration challenges and produce analyses from the centralized and standardized database of long term NPP data that were collected very differently, but all from grassland sites. The datasets represent a core area of research within the LTER community, and the resulting integrated GDI database represents a powerful resource for ecologists to perform cross-site analyses.

Project objectives were to make the integration process more efficient, enable cross-site analysis, conserve fine levels of data granularity, and eventually accommodate ANPP data from sites outside the grassland biome, as well as other grassland sites. In addition to the post ASM06 workshops I and II that we report on below, the GDI information managers and computer scientists met in March of 2008 at the Jornada Experimental Range LTER office, with support of the original *Databank* Project, to continue work on data products and publications. As of March 20, 2008, the GDI database contained 113,500 distinct NPP observations from five sites, and was 73 MB in size. Because the database creation process brought data errors to light, the integrated database contains fewer errors than the original source data.

In addition, we developed *Specifik*, a web-application to assist a user in adding USDA PLANTS species codes to a species table to standardize data described to species level within the GDI database or similar projects. Currently, the GDI database is available upon request from the SEV information manager and *Specifik* is open source and freely available for download from the internet (http://alala.evergreen.edu/~mallettj/specifik/). Lastly, our collaborators are currently producing multiple newsletter articles, posters, and peer-reviewed journal articles on our technical solutions, lessons learned, and ecological results.

May 10-11, 2007 Post ASM Grasslands Data Integration Workshop I, Colorado State University, Fort Collins, CO

Participants: Judy Cushing (1), Nicole Kaplan (2), Judith Kruger (3), Christine Laney (4), Carrie LeRoy (1), Alan Knapp (2), Juli Mallett (1), Daniel Milchunas (2), Esteban Muldavin (5), Ken Ramsey (4), Susan Stafford (6), and Kristin Vanderbilt (4), and Lee Zeman (1)

---

[1] Support for workshop III provided by NSF Canopy Database Project (NSF Grants: DBI-0417311, DBI-0319309), JRN-LTER (NSF Grant: DEB-0080412), SEV-LTER (NSF Grant: DEB-0080529), and SGS-LTER (NSF Grant: DEB-0217631).

[2] DBI-0417311, RUI: Forest Canopy Databases and Database Tools – Branching Out to Ecological Synthesis

From: (1) The Evergreen State College, (2) SGS LTER, (3) Kruger National Park, S. Africa, (4) JRN LTER, (5) SEV LTER, (6) Univ. Minnesota.

The objective of this workshop was to present preliminary analyses of the prototype integrated database, and to work with the ecologists to articulate semantic differences in experimental design among the sites and scientific questions of interest. We also conducted community analysis and multivariate analyses and discussed results with ecologists.

1. Grasslands ANPP Data from three LTER sites:
    a. JRN 1990-1998 from the web site, SEV 1999-2004 (with USDA Plants species codes), SGS 1983-2005 (with their own species codes) was integrated into one MS Access database, with a species code table linking site species code to USDA plants code.
    b. The integrated database has both species codes originally on the dataset as sent from the LTER site and derived ANPP per m2
    c. There are approximately 10,000 rows of data. From the integrated database, we abstracted two data sets: annual ANPP observations per m2 by species code and annual ANPP observations per m2 by plant family. This was used to conduct exploratory data analyses.
2. We identified challenges related to analysis:
    a. There is NO YEARS where all three sites overlap, so we identified additional years needed across sites to create an integrated database with overlapping years. Several of these were later added to the database.
    b. Preliminary analysis – the data alone don't show why ANPP varies.
        i. Site "explains" only 3.9% variability, so (1) might not be an issue.
        ii. Dominant vegetation type also explains only 3.5% variability.
        iii. Year only explains 0.4% of the variation
        iv. LTER sites precipitation data from climDB will be used to help increase explanatory power.
        v. The large size of the database is limiting the kinds of analysis that can be done, given the tools we have. So, community analysis, e.g., species diversity, will be done on FAMILY alone. Also there were too many distinct species to manipulate, so we identified the need to for data transformation tools that can work with large datasets.
        vi. Experimental design affects level of granularity at which data should be statistically analyzed. Level of granularity and aggregation steps for data analysis were determined by ecologists
3. We identified problems related to data errors and how to handle them:
    a. Species code missing from site's species table. If an error occurred in the recording of species, mark it, fix it, or drop it. If a missing species code, add it, and rerun.
    b. Species code does not correspond to a single species (i.e. is it an x or a y). (what % of data?), determine species
    c. Species code for site x corresponds only to a genus. Throw these observations out for analysis at species level and analyze at genus level.
    d. Data outliers. Note them in the database, listed as error and ignored, and notify the site.

June 5-8, 2007 Post ASM Grasslands Data Integration Workshop II, The Evergreen State College, Olympia, WA

Participants: Judy Cushing (1), Jincheng Gao (2), Nicole Kaplan (3), Christine Laney (4), Carrie LeRoy (1), Juli Mallett (1), Ken Ramsey (4), Mark Servilla (5), Kristin Vanderbilt (6), Lee Zeman (1).

From: (1) The Evergreen State College, (2) KNZ LTER, (3) SGS LTER, (4) JRN LTER, (5) LTER Network Office, (6) SEV LTER.

The purpose of this workshop was to articulate what we have learned and would like to share about challenges to integrating extensive and diverse ecological datasets, and to plan for subsequent publicatiosn of the integrated database, how to sustain the GDI database within the LTER, and where to publish project results:

1. Why is it important to integrate these data?

2. Roles of Information Managers, input from Principal Investigators, expertise of statistician, and the skills of Computer Scientists all important.
    a. Computer scientist input involved defining semantics, parsing skills, ability to take data in different formats and configure into other formats, fit into design of a larger database, apply the concept of data decomposition.
    b. Data structures were a challenge to manipulate, how to approach; as a result of writing better scripts, some integration processes that took over night, now take an hour.
    c. Several challenges required human intervention, so we suggest for future data integration:
    d. machine-readable format (save as CSV), separate taxonomic names into genus and species,(can't parse a binomial name because the code must represent one thing), better document changes over time, have each site produce a mapping from site codes to USDA plants database codes.
3. Need to communicate and document methodologies clearly
4. Need to determine how to aggregate data to for statistically valid analysis across sites, as each site has a unique experimental design and comparable level of granularity for data needs to be determined to perform valid statistical tests
5. Need to document instructions for the workflow to derive comparable unit of analysis (g/m2)
6. Need to determine what metadata, contextual data should be collected and how it should be stored , i.e., When, Where, What and How Much, now the metadata for the site needs to be processed and integrated, just as we have integrated the data.
7. There are important reasons to maintain site-specific information, such as local species codes
    a. Suggestions on how to maintain this information (e.g., flags)
8. Need to know treatments, if any.
9. Need to determine what contextual data (aka ANPP drivers) should be collected and stored (eg drivers, such as ppt, temp, PDSI).
10. Need to communicate with other groups within the community to find out what we can learn (e.g. EcoTRENDS).
11. When adding new sites, need to determine at what level of granularity data are comparable
12. *Need to develop a list of tricks for working with large datasets.*

March 23-25, 2008 Post ASM Grasslands Data Integration Workshop III, New Mexico State University, JRN LTER Office, Olympia, WA

Participants: Judy Cushing (1), Nicole Kaplan (2), Christine Laney (3), Juli Mallett (1), Ken Ramsey (3), Kristin Vanderbilt (4), Lee Zeman (1).

From: (1) The Evergreen State College, (2) SGS LTER, (3) JRN LTER, (4) SEV LTER.

The purpose of this two day session was to plan the transfer of the GDI database from Evergreen to the LTER site (SEV), to train LTER IMs in the use of the site-to-PLANTS species code table mapper (*Specifik*), to determine a possible validation process for the integrated database, and to plan for the first publication of the work we have done.

Future plans for International Workshop to integrate NPP data from South Africa and China

Participants: Judy Cushing (1), Nicole Kaplan (2), Christine Laney (3), Juli Mallett (1), Ken Ramsey (3), Kristin Vanderbilt (4), Lee Zeman (1), and representatives from South Africa (Ukulinga Research Station) and a dataset from China (Xilingol, an ILTER site).

From: (1) The Evergreen State College, (2) SGS LTER, (3) JRN LTER, (4) SEV LTER.

Kristin Vanderbilt, IM from Sevilleta successfully applied for an International Supplement grant in order to host a workshop to analyze the GDI data with the addition of another dataset from South Africa (Ukulinga Research Station) and a dataset from China (Xilingol, an ILTER site).