Developing a cross-site system to improve access to vegetation synthetic databases:

Veg-DB Workshop II

Sevilleta Field Station, New Mexico
April 30-May 2, 2013

Introduction

A second workshop was held at Sevilleta Field Station to further consider the development of a system to provide access to vegetation-related synthetic databases that would help foster cross-site research in the LTER network. In the initial workshop held at Harvard Forest, participants from a range of sites representing very different biomes and measurement methods met for two days to consider how this system might be structured, what it might provide, and how it might be implemented. In the second workshop held at Sevilleta Field Station the participants, also representing a wide set of ecosystem types and expertise ranging from field scientists to information management specialists, reviewed the results of a survey sent out to the LTER sites to gauge needs and interest in the system. They also build upon the findings of the first workshop, designing the general system architecture and identifying key aspects of the system's function. The following report summarizes the findings of this second workshop, but to provide background the general findings of the first workshop are first summarized.

Findings of Workshop I

The aspects of the Veg-DB system that were covered in the first workshop report included: an overview of the problem to be solved, the basic objective of the system, possible uses for the system and questions it could help address, types of data that would be needed, the resolution of the system in terms of time, space and taxa, the range of complexity of computing variables such as biomass and NPP across the LTER sites, the possible structure of the system and how it might evolve over time, and the relationship of this effort to other vegetation-related systems and other kinds of databases that might help support interpretation of vegetation data. The highlights of each of these topics were:

1. **Objective of the Veg-DB System.** The objective of Veg-DB would be to deliver reliable and consistent vegetation-related data to users via a single web-based portal. The focus would not be on primary/raw data which can be either currently gathered either from individual sites or future from a network system such as PASTA. Rather the system would provide access to a value-added, secondary data product with standardized units as well as the ancillary information needed to interpret these data. The kinds of data would include population, community, and ecosystem parameters, but not data on seed production, germination, flowering, phenology, and spatial arrangements such as those represented on stem maps.

2. **Benefits of Veg-DB.** Several benefits that Veg-DB could provide include help to sites not usually estimating the variables to be provided by Veg-DB, inform investigators what data is being collected at which sites, allow long-term data to be shared in a meaningful and useful manner, enhance the LTER network capacity to lead ecological synthesis efforts, and help address research problems that currently viewed as data limited.

3. Uses of Veg-DB. The kinds of topics and related hypotheses that Veg-DB could enable addressing include: Individual plant growth rates versus size/age of plant; Temporal trends in mortality related to climate variability and change; Temporal trends in NPP related to climate variability and change; Successional patterns of biomass accumulation and NPP; The relationship between diversity (richness, evenness, etc) and NPP; Correlation of temperature, precipitation, and other abiotic factors with broad-scale patterns of NPP and biomass. Veg-DB could also be a helpful resource in educational activities such as laboratories and course projects. Veg-DB would potentially be an important resource for parameterizing and testing simulation models.

4. Types of Data. Veg-DB will need to do more than deliver raw data and would integrate the raw measurement data and supporting data (e.g., plot areas, species information, conversion factors, biomass equations) to generate value-added output data. It will also need to connect to other database systems to provide the ancillary data needed to interpret these output data. The value added output data will include ecosystem, community, and some key population variables that can be derived from a common set of raw data. This includes at the ecosystem level: 1) live biomass and carbon stores, 2) NPP, 3) net change in live biomass, 4) mortality and litterfall, 5) ingrowth/birth of new biomass, and 6) herbivory; At the community level: 1) presence/absence of species, 2) dominance expressed as cover, basal area, density, volume, biomass, and carbon, and 3) diversity expressed as richness and evenness; and at the population level: 1) density of individuals, 2) recruitment into minimum size class measured, and 3) mortality of individuals. It is unlikely all output variables will be available for all sites. Therefore sites will initially provide the data they have and will be encouraged to supply the missing data as resources allow.

5. Data Resolution. Veg-DB will provide data at specific levels of spatial, temporal, and taxonomic resolution. The minimum time step of the data would be one year, either as a cumulative value (NPP and other fluxes), an annual average (biomass), or peak value (cover). The spatial resolution of the data might range from individuals to subdivisions of plots, to plots, to a level of "logical" plot aggregation such as a watershed, marsh, stand or tract. Data will be available at the species, life-form (e.g., herbs, shrubs, trees), and fully aggregated levels (i.e., all plants). Additional levels such as genera should be easily derived from Veg-DB resolutions. The minimum size of plants in terms of height or diameter will be determined by the sites supplying the data.

6. Range of Complexity of the Problem. A key issue to be resolved is how to deal with the range of complexity of vegetation collected at the sites. At some sites biomass is directly harvested whereas at other biomass and NPP cannot be measured directly and requires many kinds of indirect data on dimensions and numbers to derive these variables. This indicates that each site or groups of similar sites may need to develop their own calculation methods.

7. Limitations to Developing Veg-DB. There are several factors that might limit the development of Veg-DB: willingness of investigators to share data; lack of rewards; and excessive costs relative to other activities that are expected.

8. **Possible Structures and Evolution of Veg-DB.** Veg-DB could have two possible roles as part of a new system. In the simplest configuration, Veg-DB would primarily be an aggregation and filtering tool that draws upon data in PASTA. A more complicated configuration would have Veg-DB serve several roles as a data filter and aggregator for users, but it would also create the value-added databases removing that responsibility from individual sites. The workshop participants recommended a two stage

development process. The first phase would have the sites create the value-added databases and upload them into PASTA. The second phase of development would transfer the responsibility of creating the value-added data from sites and make that a network level activity. Sites would provide all the raw and supplemental data to create the value added database to PASTA. The Veg-DB system would periodically rerun the calculations to produce the value-added database and enter it into PASTA. Veg-DB would also be used to retrieve, aggregate, and report the value-added data. Another aspect of evolution to consider is the number of sites involved. While including all sites with vegetation data is a desirable final goal, it would be better to start with a smaller set of prototype sites. While initially designed for the LTER network, Veg-DB could be used by other entities with similar data and missions such as the US Forest Services experimental forests.

9. **Relationship to Other Efforts.** There are a number of other efforts that Veg-DB should take advantage of including: Existing site level scripts and programs to create value-added data; The Veg-X exchange format for vegetation data; Veg-DB designers could learn from other previous efforts such as SiteDB, ClimDB, HydroDB, and CTFS-SIGEO; Use other database systems such as Site-DB in a complementary manner.

Survey Results

One of the activities identified in the first workshop was to conduct a LTER community survey to determine the kinds of vegetation data and processing infrastructure at sites, the degree of interest in participating in Veg-DB, and information on how scientists and educators might use Veg-DB.

Of the 26 sites sent the survey, 18 replied. Of the sites not replying, the majority likely did not feel they had vegetation. While vegetation is a terrestrial term, almost all the sites have primary producers, and the structure of Veg-DB would accommodate most forms of plants including aquatic ones. Four terrestrial sites (CDR, CWT, JRN, and NWT) did not reply to the survey, which might indicate a lack of interest and support for the proposed system. Of the 18 sites replying, 15 were interested in participating, 2 were potentially interested, and one was not interested. The later was largely because the form of plants studied did not appear to fit into the framework. We interpret these results to indicate there is widespread support to develop Veg-DB, but that aquatic life forms need to be accommodated if it is to encompass the entire network.

The survey participants were asked which level of biology was a research focus at their sites. All the sites were working at the ecosystem level, indicating that variable such as biomass and NPP would be valuable to include. Community level and landscape levels of biology were being examined at 16 of the sites, and the population of biology was being examined at 13 sites. Organismal biology was being examined at 9 sites. This indicates that by targeting population, community, and ecosystem levels of biology, Veg-DB could serve a large fraction of the needs of potential users, but not all.

When asked whether Veg-DB could help answer questions currently being asked by site scientists, 13 of the sites replied yes, 3 no, and one site did not answer the question. This indicates that either that Veg-DB is not focused on the right variables or that sites are addressing site level questions and the proposed system would not help on that front. Regardless of interpretation of the “no” reponses, it would seem that Veg-DB would be useful for a large share of the sites replying.

To gain a better sense of the kinds of data that sites collected on vegetation, sites were asked how they collected data and processed it to produce estimates of biomass for a range of plant life-forms. The

general conclusions are summarized below, but further details can be examined in Figures 1-6 at the end of the report. As might be expected, the type of measurement and method of conversion to biomass was quite varied with sites counting individuals, tracking individuals, or making aggregated measurements. Direct harvest of plants was most common for forbes and graminoids, whereas tracking individuals was most common for trees. The kind of response was largely a function of the ecosystem present at the site. For example, there were 8 forest sites and 8 sites tracking individual trees to estimate biomass. The conclusion is that Veg-DB will need to accommodate a wide range of data types and methods. However, a large share of the sites do track individuals, which suggests that starting the individual level of resolution would be appropriate as long as some level of higher aggregation was also included to allow the other sites to participate.

The number of sites able to provide an estimate of aboveground NPP was very high, but the same cannot be said of belowground NPP. There were 31 combinations of all the life forms in which some form of aboveground NPP could be estimated. In contrast, there were 9 for which belowground NPP could be estimated. The conclusion is that Veg-DB should initially focus on aboveground biomass and NPP.

Mortality of individuals and in aggregate is an important variable that contributes to NPP estimates, but has great value on its own considering it is likely one process involved in response to changing climate. Over half the sites collect mortality at the individual level, and a similar proportion collect data on litterfall of plant parts such as leaves.

Site were also asked if they had any concerns about the Veg-DB system. Only one site replied to this question, but raised a fundamental issue of acknowledging the providers of the data. The concern was that individuals would not be acknowledged even if sites were acknowledged. The participants recognized the need to acknowledge individuals, but also recognized that with the many datasets potentially included in Veg-DB that this may be hard to achieve. The solution would be to make sure that information on the individual contributors is included in the metadata documenting the secondary data products and that acknowledging the use of these products implies that an individual's contribution is acknowledged.

Renaming the Project

Initially the system as called VEG-DB to match the conventions used in Clim-DB, Hydro-DB, and Site-DB. However, the participants felt that this was not providing the correct image of what the system was really about. The use of DB for database implied that Veg-DB was a database, but the data used by the system would actually be stored in PASTA. When the other DB systems were envisioned, PASTA was still under development and creation of a separate database was logical. However, with PASTA coming online, the creation of separate databases, at least for long-term storage does not make sense. Given that the role of the proposed system is to provide a tool or engine for synthesis and cross-site research, the participants thought that VEG-E (Vegetation Engine) would be a more appropriate name and is used throughout the rest of the report. They also felt that rather than create parallel database systems, it would make more sense to create a suite of engines that draw upon PASTA for the data and metadata, but allow the user to manipulate these data. Veg-E would be the first of these engines targeted at specific topics.

System Architecture

As alternative system architectures were considered, it was clear that several issues needed to be addressed. First, the sites needed to maintain control of which data were being accessed by Veg-E. While it might be possible to assume that the latest version added to PASTA was the one to use, this might not necessarily be the case. Sites would be the most knowledgeable about this decision. Second, the diverse and complicated nature of the process to convert raw data such as individual records to biomass and NPP means that this task is probably best done by individual sites. While it would be possible to maintain the calculation scripts centrally, this would be difficult to maintain, for example if sites develop better ways to estimate variables. The possibility that two current versions of scripts existing is increased when calculation scripts are maintained at two locations. Third, while storing data in PASTA over the long-term is desirable, if this is the only form of data storage it means that Veg-E will need to extract the data each time it is used, which would prove inefficient and slow performance for the user.

Bearing these factors in mind the following system was envisioned (Figure 7). There are three entities involved in the system: the sites, PASTA, and Veg-E. Sites would have several responsibilities: using the raw data and associated supporting data to make estimates of the variables to be used by Veg-E, putting them in a standard format, providing the metadata for these data including the calculation methods used, uploading the data and metadata to PASTA. The sites would also maintain a list of data packages that have been uploaded to PASTA that are to be considered by Veg-E. This has several advantages. It allows sites to upload preliminary data packages to PASTA, but not make them generally available, it allows sites to allow some, but not all vegetation-related data to be part of Veg-E, and it eliminates any guessing as to which data packages are appropriate. It also puts the most complicated and diverse task at the site level where it is most likely understood the best. As long as the sites fill in the required data in the form needed, the system will work. PASTA would serve as the long-term data storage system for the individual data packages used by Veg-E, but also for periodic versions of the overall secondary data product used by Veg-E. Veg-E would store data, but primarily to enhance performance and not as a primary data storage location. Periodically Veg-E would query sites as to the data packages on their list; if new data packages have been added to PASTA, then Veg-E would extract those data and add them to the aggregated database that users can work from. Periodically Veg-E would also write the aggregated database to PASTA. This would provide snapshots of how the aggregated database changed and also would allow users to revisit earlier versions of the aggregated database to either check calculations or do comparisons. While there are complicated issues involved in creating this system, overall it is quite simple and generic allowing it to be a model for other database engines.

Creation of Site Scripts and Best Practices

Given that sites will be responsible for creating a standardized set of data variables to be used by Veg-E, there is a need to create site level calculation scripts. To assure that these have some level of standardization a set of best practices for these scripts should be developed. This would include the following steps:

- Designing and identifying the parts of the calculation system and their connections
- Developing equations and specific calculations
- Testing scripts to verify they are working as intended
- Documenting scripts with metadata including version control information

- Publishing the actual scripts and supporting data associated with the calculations in PASTA as a part of the metadata

Data Structures and Attributes

There will be three data structures that will be used by Veg-E and be uploaded by sites into PASTA: individual data, aggregated data, and species/taxa related data. To work efficiently these secondary data product will use standardized units, standardize variables, and standardize terms to the degree possible.

For the individual data the following kinds of variables would be requested. Some of the variables, such as location and time of collection would be required of all sites, others would be provided by sites depending upon availability. The kinds of variables would be as follows:

Place and time identifiers such as site, LAU (logical aggregation unit), sampling unit, subsampling unit, and time of collection. LAU would be used to aggregate sampling units into similar classes. Sampling unit could be either quadrats, plots, or transects or other systems that are used to sample vegetation.

Taxa and individual identifiers including genus, species, and a unique individual identifier. It will be important to separate the binomial into its parts so that sorting can be done by genus.

Status of the individual which could be either a survivor, an individual that died, or that appeared from the last measurement.

Dominance variables including biomass, C store, cover, basal area, and volume of the individual. In addition to the current values of these dominance variables, it would be desirable to know the change since the last measurement (e.g. delta biomass). The units of all variables

Number of individuals would include two variables: the number or density of individuals in a standardized area and the number of individuals represented by the record. The later would allow one to use the individual data format to include count data where each individual actually represents a number of individuals (e.g., number of plants in a given size class).

Other possible variables at the individual level would include litterfall and part mortality, but these would involve the use of a model that estimated the turnover of these parts based on the biomass of parts and average life-spans.

It was decided that the stores, cover, volume, etc of subparts of the plants (e.g., bark versus wood or leaves versus wood) would not be accommodated at this time in part because of the diversity in which this has been done at the sites.

The structure of the aggregated data would be quite similar to the individual data in many respects. However, all individuals would be aggregated to the level of the species binomial. The other significant difference is that the dominance variables (i.e., biomass, C, cover, basal area, volume, and number per area) would include the current value, the change since the last measurement, a amount of mortality since the last measurement, and the amount of new ingrowth since the last measurement. In addition for biomass and C there would be an estimate of NPP which would be the sum of the delta, mortality,

and ingrowth terms. Gross volume growth would be used in place of NPP for volume as that is a more appropriate term.

Taxa/species data would be provided to help interpret the individual and aggregated data. Sites would provide a species list and for each entry there would be a species code, the genus and species it represented, the taxonomic authority, the taxonomic key used so that a sense of how the name is being used is understood, information about the life form of the species (e.g., tree, shrub, forb, graminoid, planktonic, benthic, non-vacular terrestrial), and the life stages that are possible (e.g., seedling, sapling, tree). While some of this information could be derived from the other data such as the species list, others cannot. Having sites provide this information would be extremely helpful. While it requires sites to do some extra work, it should not have to be updated very frequently.

User Interface Design

There was not sufficient time at the workshop to create a mock-up of the Veg-E user interface. However, the kinds of pages that the user might encounter were discussed and our described below.

The home page would be the first one users encounter and would serve several functions. It would provide general information about Veg-E and frequently asked questions (FAQ's), the sources of funding used to develop it, the conditions of data use (such as standard acknowledgments), and would allow the user to sign in. The latter would form and agreement to abide by the conditions of use and to allow a one-time sign-in as compared to multiple sign-ins as different data are requested and used.

The next set of pages encountered would provide search and browse capability. At this point the user is gathering information about the data and creating a wish list (much like filling a shopping cart while online shopping). The information gathered might be about the availability certain kinds of data or data from certain places, information about how the data was collected or processed etc. Searching would allow the user to locate information by putting in their own set of word and terms. Browsing would be constrained by the locations, times, variables, etc that are in the database.

Once the user has found the data they want they will go to another page which allows them to select and extract the data they want to use. While this could be included in the search and browse pages, this is a separate activity and allows the user to either modify their wish list or go directly to this page to list the data they want.

After selecting and extracting the data the user wants, the next page would allow the user to aggregate the data into the time, space, and taxa resolution required for the analysis.

Analysis using graphs, simple statistical calculations, and text searching would be on the analysis page. The emphasis of this analysis system would be on rapid analysis and data exploration and not on creating publication quality graphs or sophisticated statistical analysis. Graphs would include bar charts of categories, X-Y plots, and time series. Statistical calculations would include means, standard deviations and errors, minimum, and maximum values. Text searches would allow the user to find particular values or locate the minimum and maximum values in the data set.

The final page would allow the user to output the data, metadata, acknowledgments, and log of selections the user made during their session. The latter would allow the user to repeat a session or to pick-up where they left off from a previous session.

While it would be desirable to have a page or set of pages where the user an import ancillary data (e.g., climate) to interpret the data in Veg-E, it was decided that this functionality would have to wait until the other engines are developed.

Modules

To operate Veg-E would need the following functional modules:

PASTA harvester/extractor to gather the required data

Aggregator (time, space, taxa)

Graphing tool to create time series, bar, X-Y plots, phase diagrams

A by function so that tasks can be repeated for multiple classes of objects

Date converter to get all the date conventions similar

Output tool to export data, metadata, the session log, graph

Species/taxa reconciler to assure that similar taxonomic objects are compared

Statistics tools to calculate the mean, minimum, maximum, standard deviation

A diversity calculator because these variable lie above the aggregated data

Ancillary Data

While it would be highly desirable to be able to combine the vegetation data with other kinds of data on site conditions, this may not be initially possible. The participants therefore suggested a phased development:

Phase 1-no capability to bring in ancillary data other than that used to browse, search, select vegetation data

Phase 2- an output file of sites and locations that could be uploaded to other Interface Engines to extract the relevant data

Phase 3- an ability to harvest required data from other Interface Engines

Next Steps

The next steps will be to identify the prototype sites and to agree upon the format and variables in the data sets to be created by the sites. Then the prototype sites will populate examples of these data sets and they will uploaded to PASTA so that testing of the systems functionality can begin.

Workshop Participants

Attending:

Emery Boose (HFR)
Mark Harmon (AND)
Jim Morris (PIE)
Fox Peterson (AND)
Dan Reed (SBC)
Suzanne Remillard (AND)
Roger Ruess (BNZ)
Mark Servilla (LNO)
Bob Waide (LNO)

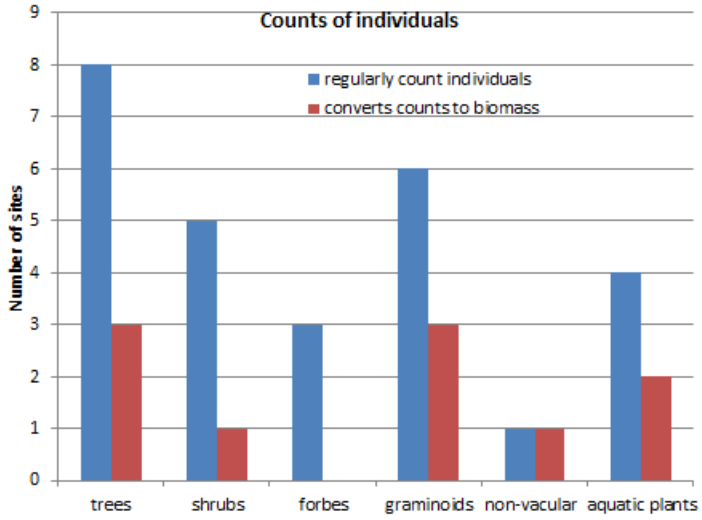

Figure 1. Number of responses for whether individuals are counted and if those data are converted to biomass by plant life-form.

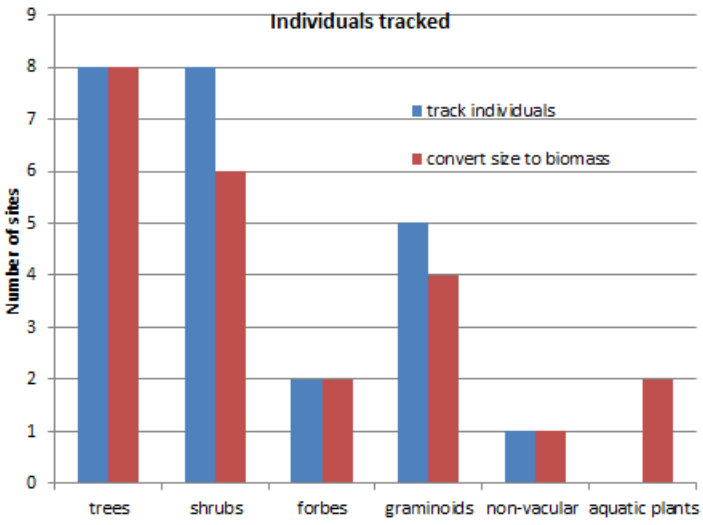

Figure 2. Number of responses for whether individuals are tracked and if those data are converted to biomass by plant life-form.

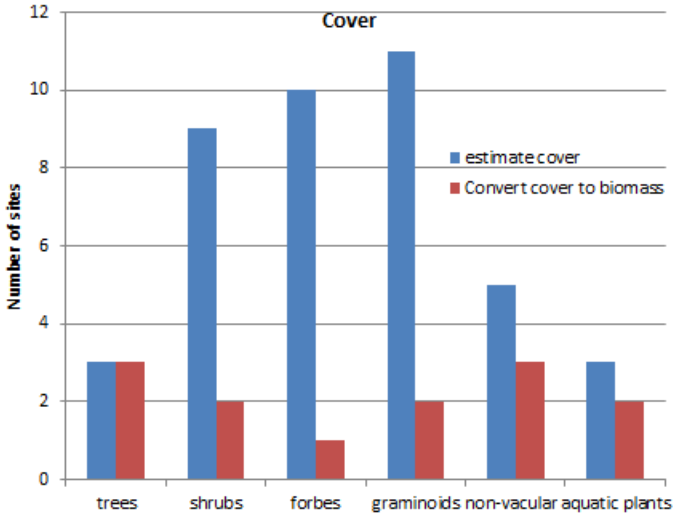

Figure 3. Number of responses for whether cover is estimated and if those data are converted to biomass by plant life-form.

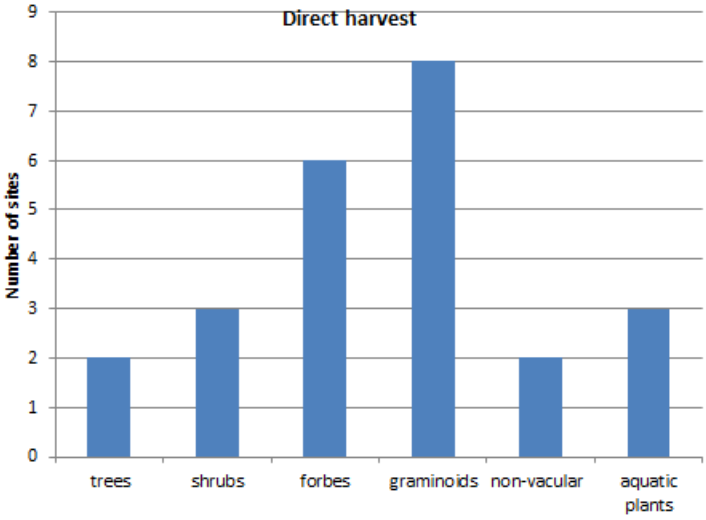

Figure 4. Number of responses for whether biomass is directly harvested by plant life-form.

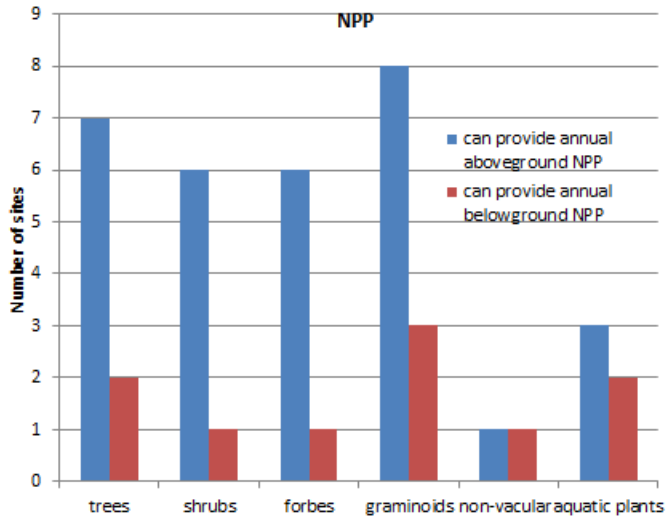

Figure 5. Number of responses for whether NPP is estimated by plant life-form.

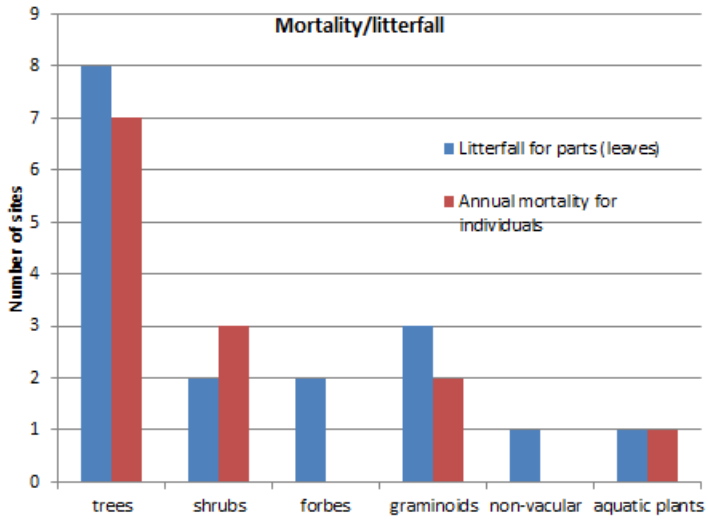

Figure 6. Number of responses for whether litterfall and individual mortality is measured by plant life-form.

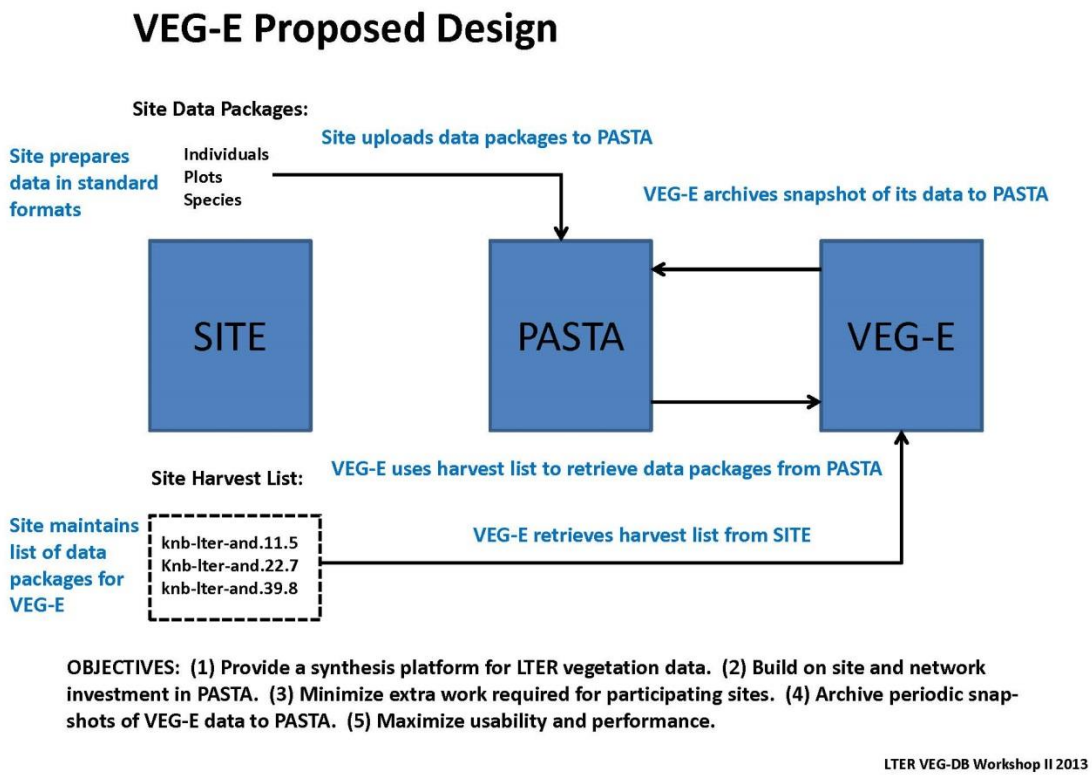

Figure 7. Proposed architecture of the Veg-E in relation to sites and PASTA.