Research paper

An Ecoinformatics Application for Forest Dynamics Plot Data Management and Sharing

Chau-Chin Lin,[1] Abd Rahman Kassim,[2] Kristin Vanderbilt,[3]
Donald Henshaw,[4] Eda C Melendez-Colom,[5] John H Porter,[6]
Kaoru Niiyama,[7] Tsutomu Yagihashi,[7] Sek Aun Tan,[2]
Sheng-Shan Lu,[1,9] Chi-Wen Hsiao,[8]
Li-Wan Chang,[8] Meei-Ru Jeng[8]

[Summary]

Several forest dynamics plot research projects in the East-Asia Pacific region of the International Long-Term Ecological Research network actively collect long-term data, and some of these large plots are members of the Center for Tropical Forest Science network. The wealth of forest plot data presents challenges in information management to researchers. In order to facilitate the management of these data, a Forest Dynamics Plot Database and Application Workshop was held in Taiwan 2009. This paper describes the results of the workshop that produced and tested an integrated information management framework. The goal for the framework was to demonstrate how fully documented data archives can be effectively used for data discovery, access, retrieval, analysis, and integration. Results from our work included setting up a database based on the Center for Tropical Forest Science structure on a local relational database (MySQL) server, an authentication interface, a metadata query web page, and 3 workflows to test the framework.

[1] Forest Protection Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Taipei 10066, Taiwan. 林業試驗所森林保護組，10066台北市南海路53號。

[2] Forestry and Conservation Division, Forest Research Institute Malaysia, 52109 Kepong, Selangor Darul Ehsan, Malaysia. 馬來西亞林業試驗所林業與保育組。

[3] Sevilleta LTER, Department of Biology, Univ. of New Mexico, Albuquerque, NM 87131-1091, USA. 美國新墨西哥大學。

[4] US Forest Service Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, OR 97331, USA. 美國林務署太平洋西北試驗站。

[5] Institute for Tropical Ecosystem Studies, P.O. Box 70377, San Juan, PR 00936-8377. 波多黎各熱帶生態系研究所。

[6] Department of Environmental Sciences, Univ. of Virginia, 291 McCormick Rd., Charlottesville, VA 22904-4123, USA. 美國維吉尼亞大學。

[7] Tohoku Research Center, Forestry and Forest Products Research Institute, 92-25 Nabeyashiki, Shimokuriyagawa, Morioka, Iwate 020-0123, Japan. 日本森林總合研究所。

[8] Forest Technology Service Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Taipei 10066, Taiwan. 林業試驗所技術服務組，10066台北市南海路53號。

[9] Corresponding author, e-mail:sslu@tfri.gov.tw 通訊作者。

Received February 2011, Accepted September 2011. 2011年2月送審 2011年9月通過。

Key words: LTER, metadata, forest ecology, scientific workflow.

Lin CC, Kassim AR, Vanderbilt K, Henshaw D, Melendez-Colom EC, Porter JH, Niiyama K, Yagihashi T, Tan SA, Lu SS, Hsiao CW, Chang LW, Jeng MR. 2011. An ecoinformatics application for forest dynamics plot data management and sharing. Taiwan J For Sci 26(4):357-69.

研究報告

生態資訊學應用在森林動態樣區研究之資料管理與分享

林朝欽[1] Abd Rahman Kassim[2] Kristin Vanderbilt[3] Donald Henshaw[4] Eda C Melendez-Colom[5] John H Porter[6] Kaoru Niiyama[7] Tsutomu Yagihashi[7] Sek Aun Tan[2] 陸聲山[1,9] 蕭其文[8] 張勵婉[8] 鄭美如[8]

摘 要

森林動態樣區設置蒐集了長期性與大量的資料，此長期性研究是東亞太平洋國際長期生態研究網的主要項目之一，這些長期生態研究網所設立的森林動態樣區部分也是美國史密斯研究所熱帶森林中心的網絡成員。累積大量的森林動態樣區資料後，其管理與使用對研究人員來說成為一項挑戰。為了促進森林動態樣區資料管理與應用，2009年在台灣舉辦了森林動態樣區資料管理與應用的國際研討會。本文針對該次研討會中產出加以描述，探討資料如何完整建檔、倉儲、存取、與分析使用。此研討會利用台灣、馬來西亞、日本、波多黎各等四個動態樣區的資料，產出包括森林動態樣區資料管理的觀念性架構外，並建立了資料庫、認證界面、元數據查詢網頁與三個科學工作的分析流程。

關鍵詞：長期生態、元數據、森林生態、科學工作流程。

林朝欽、**Abd Rahman Kassim**、**Kristin Vanderbilt**、**Donald Henshaw**、**Eda C Melendez-Colom**、**John H Porter**、**Kaoru Niiyama**、**Tsutomu Yagihashi**、**Sek Aun Tan**、陸聲山、蕭其文、張勵婉、鄭美如。2011。生態資訊學應用在森林動態樣區研究之資料管理與分享。台灣林業科學26(4):357-69。

INTRODUCTION

Several forest dynamics plot research projects in the East-Asia Pacific (EAP) region of the International Long-Term Ecological Research (ILTER) network actively collect long-term data, and some of these large plots are members of the Center for Tropical Forest Science (CTFS) network. The CTFS is a network of forest plots that monitors trees from all around the world in more than thirty 25~50-ha plots (Ashton et al. 1999). CTFS plots involve hundreds of scientists from more than 40 institutions worldwide and share a common methodology as to measurements taken, periodicity of surveys, and identification of tree species (Condit 1995). There are also many other large forest research plots in the EAP-ILTER region that have comparable data. This wealth of forest plot data presents challenges for information management (IM) to researchers and offers an opportunity for information managers (IMgrs) to collaborate with scientists to facilitate the management, analysis, and integration of these data.

Cross-site analysis of CTFS plot data in the EAP region has been limited despite the standardization of methodologies. The inability of researchers to perform cross-site analyses seems to result from the lack of user-friendly analytical tools with which to analyze and visualize the data, and restrictive data-sharing policies. While the CTFS website provides analytical tools and training for scientists, these tools are often difficult to use without some background in R statistical programming language (Ihaka and Gentleman 1996). Additionally, CTFS data are only accessible by data owner permission and not publicly. Similar situations may also exist in many other countries with regards to the management, access, and analysis of long-term dynamic plot data. EAP IMgrs recognized that ILTER scientists working on large forest plots could benefit from application of ecoinformatics, and conceived a workshop to bring together IMgrs and scientists.

Ecoinformatics is a broad interdisciplinary science that incorporates conceptual and practical tools for the understanding, generation, processing, and dissemination of ecological data and information (Michener et al. 2002). Establishing an IM system (IMS) using ecoinformatics techniques has been ongoing and a high priority for EAP-ILTER member networks (Lin et al. 2006). The community of IMgrs has participated in a series of training workshops for scientists and data managers in this region to promote its development. A common and compatible IMS is a prerequisite for data to be efficiently and effectively shared, exchanged, and synthesized among individual scientists within member networks. The overall goals of these workshops were to improve ecological research through using information management, identifying IM requirements, and building member capacity for sharing and synthesizing data.

IMgrs from Taiwan have worked with their US counterparts since 2004 to develop a conceptual model for a regional IMS that is designed to be suitable for many Asian languages (Lin et al. 2006). The IMS assembles open-source programs and analytical tools from the products of ecoinformatics research. Currently, the IMS is being implemented in several EAP-ILTER member networks, and IMgrs are ready to engage with scientists to produce science-driven IM products (Lin 2008).

The Forest Dynamics Plot Database and Application Workshop was held on 14~18 June 2009 at the Lienhuachih Research Center, Nantou County, Taiwan. Plot scientists from Taiwan, Malaysia, and Japan attended the workshop, as well as IMgrs from Malaysia, Taiwan, and the US. The primary goals of the meeting were to have 1) scientists describe their needs and 2) IMgrs propose designs and a prototype as possible solutions. A secondary goal was to evaluate the use of time with the scientists in this type of workshop in order to obtain specifications for needed solutions. This paper describes the results of the workshop that produced an integrated information management framework. This framework demonstrates how fully documented data archives can effectively be used for data discovery, accessing, retrieval, analysis, and integration.

MATERIAL AND METHODS

Data source

The dynamic plot data management mostly follows the CTFS Data Management

System (DMS). Each member site contributes large plot census data, and the system creates a database for each plot. The CTFS provides each site with a user name and password, and the site takes responsibility for data uploading and controls access to site data. Distribution is controlled by the individual site, and access is by a single plot and can be shared at the owner site's discretion. The DMS consists of normalized databases, a hierarchical series of web forms for data entry and uploading, and PHP (Hypertext Preprocessor) scripts that allow the user to make quality-control checks and extract data from the database tables.

Data entry forms are accessed by the user in a hierarchical fashion from more- general information about the census (e.g., plot name, census number, and dates) to more-specific data (e.g., species, quadrats, and codes) and finally the census data. General plot information tables (species, quadrats, and codes) are uploaded once and are used by the system to check and filter the data entered into the census data table. The system creates a series of log and data tables that hold user session information and uploaded data. Tree data can easily be downloaded by an authorized user (PHPAdmin).

PHP scripts ensure that the data in the census tables are consistent with the general information tables. Records of census tables, the codes, species, and quadrats of which are not predefined in their corresponding files, are rejected by the system when uploaded. Other rules are also applied to issue warnings or errors such as tree diameter (dbh) values shrinking or dead trees appearing as alive. In this way, data uploads become part of a general data quality control (QC) process. In this paper, we used databases from Taiwan, Malaysia, and Puerto Rico which were created by CTFS's DMS and a database from Japan which was created using a general relational database structure in the workshop.

Conceptual development

Based on the current CTFS's DMS, the workshop had a brainstorming discussion session on the first day by asking both scientists and IMgrs to describe their vision of a "dream dynamics plot" information system. This exercise helped identify both scientific and technological challenges which need to be resolved, and describe what the system would consist of if they could have all the capabilities and functionality that they wanted. Table 1 summarizes the lists of features both communities' representatives expressed as desirable in their dream dynamics plot IMS.

This brainstorming exercise underscored that the scientists were eager to have access to user-friendly analytical and visualization tools. Further discussion resulted in a realistic vision of the types of tools needed, and particularly the suggestion that it would be valuable to utilize the software library contained in the R Package, currently available to CTFS researchers via a web page, that would hide the R complexity from the user. Incorporation of tools provided in the CTFS R package is central to the solution developed by the IMgrs.

After the brainstorming session, IMgrs met independently to propose solutions that would meet the scientists' requests for user-friendly analytical tools. The goal of the IMgr group was to identify tools and suggest realistic models for implementing R scripts available through the CTFS package. Emerging open-source tools (Table 2) impose good practices for preservation and reuse of data and rely on the development of structured metadata. These tools have been adopted by EAP-ILTER participants and have strong applications for use with CTFS and similar plot data. The IM group identified a conceptual

Table 1. "Dream dynamics plot" information management system

Information managers' wishes	Scientists' wishes
o Generic access control system o Improved data collectors' communication o Research group code sharing o CTFS community blog o Open-source data analysis and discovery tools o Facilitation of data discovery and use o Standardized database structure for CTFS data o Phenology, soils, and meteorological data o QC tools for field data	o Easy-to-use analytical tools o For a single site o For cross-site analyses o Well-documented sets of codes toolbox o For spatial analysis with support for GRASS and ArcGIS o Species history tool o Nomenclature information o Tracking changes o Interactive training and teaching interface using the CTFS data o CTFS data and related data 3-D visualization tools o For phenology or mammals trapped in the forest, for example o Field-collected data visualization tool o For more rapid QC o Single-web interface access to all desired analyses o Diversity, biomass, spatial characteristics, phylogenetic tree for plot species, growth rate by species, mortality o Interface to query and integrate data across sites, including non-CTFS sites

scientific workflow model that employs these open-source tools to provide needed functionality to meet scientists' needs.

Statistical methods

The CTFS R package is provided and consists of a set of R scripts for common calculations, analysis, and visualization of data. The package is downloadable online (http://cran.r-project.org/web/packages/CTFS/) and contains manuals with lessons for R program, CTFS datasets, and CTFS functions with examples.

RESULTS

Result of the workshop produced a metadata-driven framework which encompassed activities that are central to ecoinformatics. Using the framework, we tested general activities of ecoinformatics which included data design, data acquisition, quality assurance (QA) and QC, metadata implementation, and data archiving, access, and analysis using data from 4 sites. Details of the framework and testing of the data are described as follows.

Metadata-driven framework

Figure 1 displays the conceptual framework which applies ecoinformatics tools for forest dynamics plot databases. Three components are included in this framework.

The first component is data storage, which includes 2 database types. The first type is a census database which is a relational database using MySQL. This database is to validate and well-document data archives of plots and is password-protected using the Lightweight Directory Access Protocol (LDAP) which serves as the authentication

Table 2. Open-source tools that were considered to be chosen

Software tool	Brief tool description
EML	Ecological Metadata Language is both a content and syntactic metadata standard that allows description of ecological and other scientific data sets in eXtensible Markup Language (XML) (Michener 2000, Jones et al. 2001, Fegraus et al. 2005).
Morpho	Metadata entry software is used to create and edit metadata in the EML standard (Higgins et al. 2002).
Metacat	This is a metadata database that uses XML as a common syntax to represent metadata content standards and allows storage, query, and retrieval of XML documents (e.g., EML) (Berkley et al. 2001, Jones et al. 2001).
Kepler	This workflow software supports multiple models for computation allowing composition of complex operations from simpler components, supports data described by EML, and integrates the R statistical language (Altintas et al. 2004, Ludäscher et al. 2006).
R	This is a software environment for statistical computing and graphics (Lin et al. 2008).
PHP	This is a scripting language designed to produce dynamic web pages (Lin et al. 2008).
EcoGrid	This published web service provides a uniform and simple programming interface (Zhang et al. 2006).

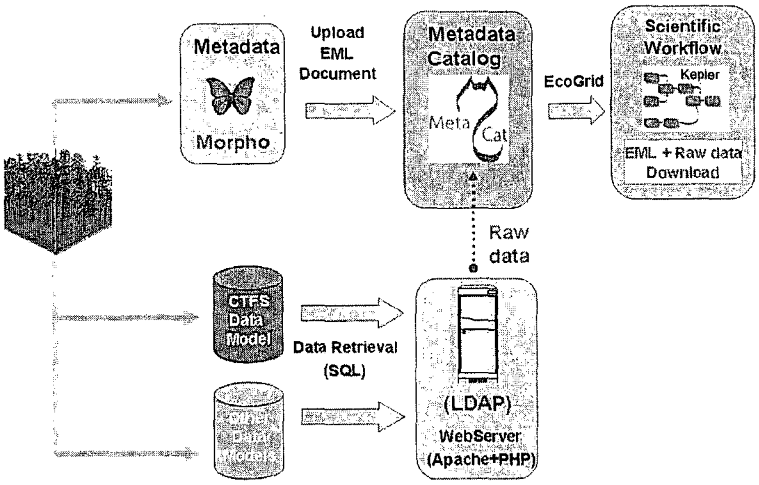


Fig. 1. Conceptual model for forest dynamics plot. Metadata (Ecological Metadata Language) including file structure information for CTFS and other plot data prepared with the Morpho metadata entry tool and uploaded into Metacat. Kepler workflow components use the Metacat to access and ingest the CTFS and other data from web servers on the EcoGrid. Other Kepler component actors provide analysis and tabular and graphical output.

control. The second type is a metadata database which is a schema-independent eXtensible Markup Language (XML) database implemented using PostgreSQL. This database is based on Ecological Metadata Language (EML) schemas. EML is a metadata standard developed by the ecology discipline but was proven to be extensible to other disciplines. The XML database structure is a storage subsystem of the Metacat framework, a metadata catalog designed by the National Center for Ecological Analysis and Synthesis (NCEAS) at the Univ. of California, Santa Barbara, CA. It is a hybrid relational database for storing XML data with arbitrary schemas. This approach permits structured, path-based queries of metadata.

The second component focuses on data discovery and handles metadata editing and querying. We chose Morpho as the metadata editor. This tool is one of the tools developed by the Knowledge Network for Biocomplexity (KNB) Project (http://knb.ecoinformatics.org) of the US. Morpho is a cross-platform, desktop application. It allows researchers to describe their data using comprehensive and flexible metadata specifications, and share their data publicly. Researchers can store their datasets and metadata either locally or on a shared metadata database described in the first component. In addition to Morpho, a data catalog web interface is another subsystem of the Metacat framework that provides a simple but powerful querying interface to assist in locating useful datasets registered within the Metacat storage subsystem.

The third component is for data analysis. EML documents provide standardized discovery and access to a collection of distributed ecological, biodiversity and environmental data and analytic resources. Published web services (EcoGrid) provide a uniform and simple programming interface for access to EML-described data and metadata. The Kepler scientific workflow system is the interface we chose as the uniform tool to allow scientists to design, execute, and monitor analytic procedures with minimal effort. Scientific applications developed at the workshop allow data acquisition, integration, transformation, synthesis, and archiving using Kepler workflows. These Kepler workflows take advantage of built-in workflow components to consume EML resources and perform a sequence of analytic and visualization operations that result in the output of tabular files and graphics. Workflows include 1) dynamically reading raw plot data based on structured metadata (EML) stored in Metacat and 2) executing of R scripts necessary to make calculations required for analysis of the data.

Raw input data are based on standard CTFS download table structures and R scripts that are extracted from the CTFS R package. Additional graphical support is provided through R script actors provided with the Kepler component library.

Scientific workflows and data analysis test

A test of the framework was conducted using restored databases from Lienhuachih (a Taiwanese site), and exported data from Pasoh (a Malaysian site), Ogawa (a Japanese site), and Luquillo (a Puerto Rican site). We retrieved 2-ha data from the Lienhuachih site, 10-ha data from the Pasoh site, 6-ha data from the Ogawa site, and 16-ha data from the Luquillo site for the test. Three scientific workflows were created using Kepler. The first workflow generated simple statistics to compare the basal areas of trees in each plot (Fig. 2). Using EcoGrid, data of the 4 plots were directly retrieved either from the MySQL databases on the TFRI server or local test server set up for the workshop which

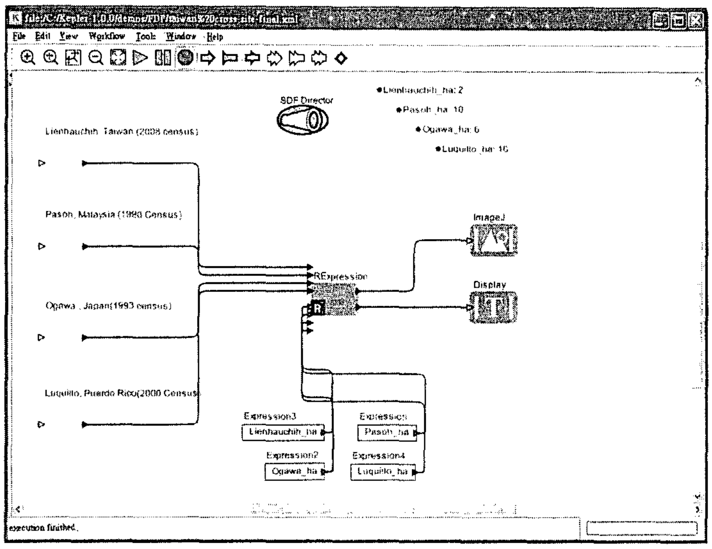
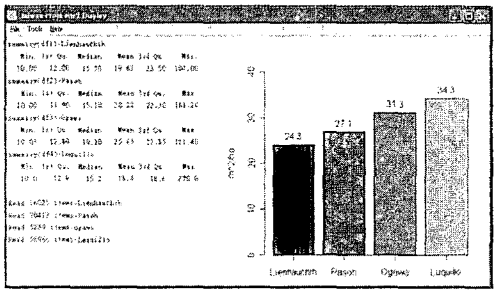

file:/C:/Kepler-1.0.0/demos/FDP/taiwan%20cross-site-final.xml
File Edit View Workflow Tools Window Help
SDF Director
●Lienhauchih_ha: 2
●Pasoh_ha: 10
●Ogawa_ha: 6
●Luquillo_ha: 16
Lienhauchih, Taiwan (2008 census)
Pason, Malaysia (1990 Census)
Ogawa , Japan(1993 census)
Luquillo, Puerdo Rico(2000 Census
RExpression
R
ImageJ
Display
Expression3
Lienhauchih_ha
Expression
Pasoh_ha
Expression2
Ogawa_ha
Expression4
Luquillo_ha
execution finished.

Fig. 2. Simple statistical calculation of basal area and comparison of 4 sites.

stored partial data from the other 3 plots. Once the 4 datasets were downloaded, EML documents provided field names to be selected for further calculation. In this example, diameter at breast height (dbh) was converted to a standard unit (cm) and used in an R-script to calculate basal area. In total, 7 Kepler actors and 4 Kepler parameters were used to implement the workflow. The result of this procedure was a bar chart of basal area at each site generated using a graph display actor from Kepler. The result showed 2 kinds of information (Fig. 3). Text results summarized each site's dbh and data records, and a graph result showed cross-site comparisons of basal areas.

The second workflow compared species abundances at the Luquillo plot among different censuses (Fig. 4). We used the function, abundance.spp, of the CTFS R package (http://cran.us.r-project.org/src/contrib/Archive/CTFS/) in the workflow, and selected 2 Kepler actors to compute the abundance. The result showed that 3 censuses of the Luquillo plot had'almost the same basal area (Fig. 5).

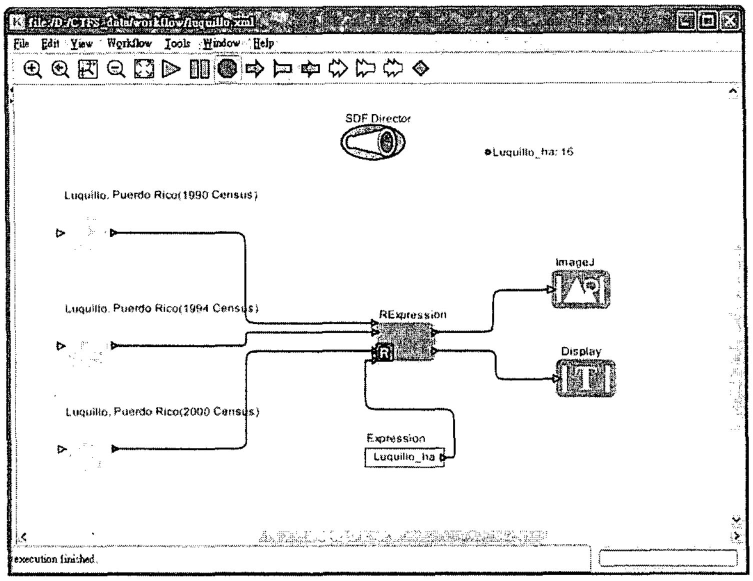


```
summary(d1)-Lienhauchih
  Min.  1st Qu.  Median   Mean 3rd Qu.    Max.
 10.00   12.00   15.30   19.63   23.50  104.00
summary(d2)-Pasoh
  Min.  1st Qu.  Median   Mean 3rd Qu.    Max
 10.00   11.90   15.10   20.22   22.30  [illegible]
summary(d3)-Ogawa
  Min.  1st Qu.  Median   Mean 3rd Qu.    Max
 10.[illegible]  12.40  16.10  22.63  [illegible]  121.40
summary(d4)-Luquillo
  Min.  1st Qu.  Median   Mean 3rd Qu.    Max.
  10.0    12.9    15.2    18.4   [illegible]  [illegible]

Read 16025 items-Lienhauchih
Read [illegible] items-Pasoh
Read 5259 items-ogawa
Read [illegible] items-Luquillo
```

Fig. 3. Results of the first workflow shows cross-site comparisons of tree diameters through text output and comparisons of mean basal areas through a graphical output. The items listed correspond to the basal area in each plot.

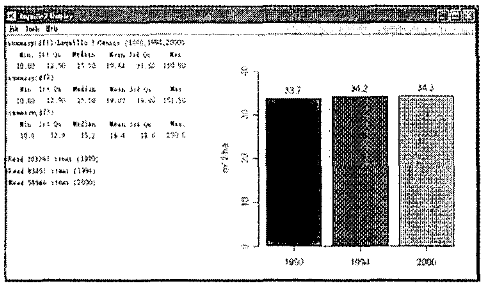
The 2 workflows showed that the framework could be used to fill in gaps of the current CTFS database system: lack of both a data-discovery procedure and a friendly analytical interface.

The third workflow compared tree abundances in different size classes. This workflow attempted to create graphs and statistical output that would facilitate analyses of size-dependent abundances in forests. The work-


file:/D:/CTFS_data/workflow/luquillo.xml
File Edit View Workflow Tools Window Help
SDF Director
●Luquillo_ha: 16
Luquillo, Puerdo Rico(1990 Census)
Luquillo, Puerdo Rico(1994 Census)
Luquillo, Puerdo Rico(2000 Census)
RExpression
ImageJ
Display
Expression
Luquillo_ha

```
execution finished.
```

Fig. 4. The second workflow calculation of tree abundances on the Luquillo plot over 3 censuses.



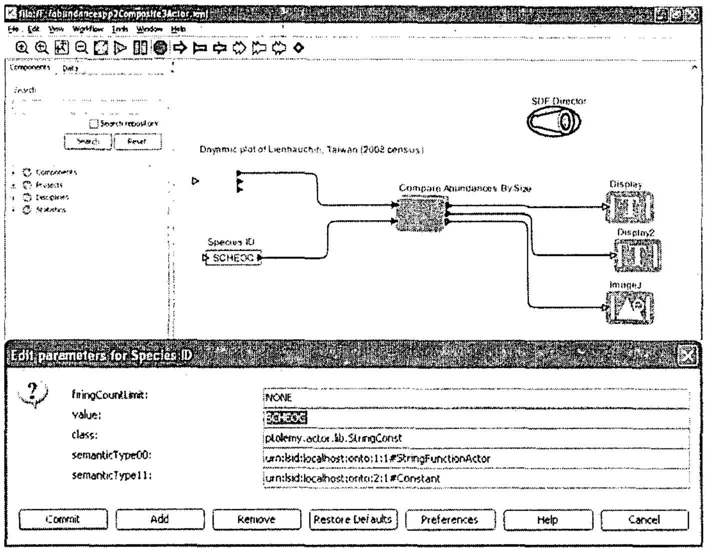
```
summary(d[1]) Luquillo 3 Census (1990,1994,2000)
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
  10.00   12.50   15.50   19.42   21.50   190.50
summary(d[2])
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
  10.00   12.50   15.50   19.07   19.50   191.50
summary(d[3])
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
   10.0    12.0    15.2    19.4    18.6    [illegible]

Read [illegible] items (1990)
Read [illegible] items (1994)
Read [illegible] items (2000)
```

Fig. 5. Result of the second workflow showing comparisons of basal areas of 3 censuses at the Luquillo site.

flow took as its input forest plot data from a metadata-driven web service, transformed the data into forms suitable for use in internal R-statistical language programs, performed statistical analyses using those R programs, and produced textual and graphical results. The workflow encapsulated operations to make high-level functionality available to the user without exposing them to unnecessary complexity. The main workflow included the data, a parameter that allows a user to select a tree species for analysis, a composite actor ("compare abundance by size") that performs the actual analysis, and text and graphical displays for the resulting statistical output and graph (Fig. 6). Expansion of the "compare abundance by size" actor showed the major functions that comprise the analysis (Fig. 7). They included preparation of the data for processing using R programs, the R program itself which performed the necessary aggregations using the CTFS abundance.spp function, and a section that further processed the output from the R analysis to produce a graph and a data file for use in other analyses. Fig. 7 also displays the "data preparation" actor and reveals how headers are added and delimiters are altered to prepare the raw data for subsequent analyses. Text and graphical results from the third workflow are displayed in Fig. 8.

DISCUSSION

This unique workshop of the EAP-ILTER combined both IMgrs and scientists to work together to discuss and examine of their



firingCountLimit: NONE
value: SCHEOC
class: ptolemy.actor.lib.StringConst
semanticType00: urn:lsid:localhost:onto:1:1#StringFunctionActor
semanticType11: urn:lsid:localhost:onto:2:1#Constant

Fig. 6. Main workflow including the data, a parameter that allows a user to select a tree species for analysis, a composite actor "compare abundance by size" that performs the actual analysis, and text and graphical displays for the resulting statistical output.

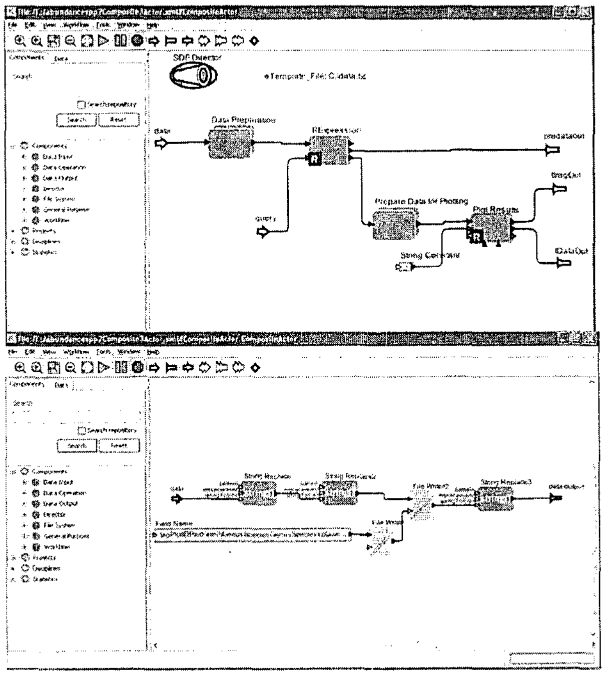
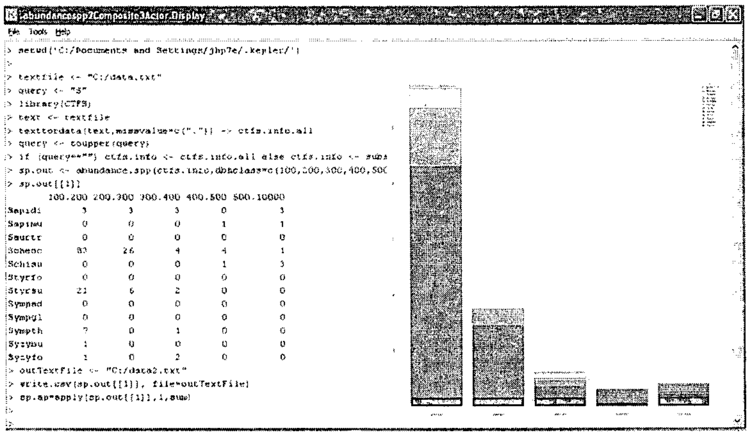


Fig. 7. Expansion of the "compare abundance by size" actor revealing the hidden, major functions that comprise the analysis. The upper picture details the preparation of the data for processing and plotting using R programs. The lower picture further details the "data preparation" actor and shows how headers are added and delimiters are altered to prepare the raw data for subsequent analyses.



```
> setwd('C:/Documents and Settings/jhp7e/.kepler/')
>
> textfile <- "C:/data.txt"
> query <- "S"
> library(CTFS)
> text <- textfile
> texttodata(text,missvalue=c(".")) -> ctfs.info.all
> query <- toupper(query)
> if (query=="") ctfs.info <- ctfs.info.all else ctfs.info <- subs
> sp.out <- abundance.spp(ctfs.info,dbhclass=c(100,200,300,400,50C
> sp.out[[1]]
        100.200 200.300 300.400 400.500 500.10000
Sapidi        3       3       3       0         3
Sapimu        0       0       0       1         1
Sauctr        0       0       0       0         0
Scheoc       83      26       4       4         1
Schisu        0       0       0       1         3
Styrfo        0       0       0       0         0
Styrsu       21       6       2       0         0
Sympad        0       0       0       0         0
Sympgl        0       0       0       0         0
Sympth        7       0       1       0         0
Syzybu        1       0       0       0         0
Syzyfo        1       0       2       0         0
> outTextFile <- "C:/data2.txt"
> write.csv(sp.out[[1]], file=outTextFile)
> sp.ap=apply(sp.out[[1]],1,sum)
>
>
```

Fig. 8. Output display of the workflow showing the R programs analytical results (left text), and graph (right picture). The graph displays the abundances of several species by diameter size classes.

common needs. Regional scientists are finding that information management is an integral part of research and not a discrete issue. Anyone who has been involved with a project that brings scientists and information technology (IT) specialists together knows it can be tough to move efforts forward (Pennington 2010). Domain scientists might not be able to envision how new technologies can assist their work, and IT specialists may have difficulty communicating in nontechnical terms. We believe that the successful outcome of our workshop, as illustrated by the efficient tools and applications developed, was due to the brainstorming exercise that allowed the group to think creatively together to find the overlap between scientist and IMgr needs. We hope that this workshop will serve as a model for future workshops on other topics that include both scientists and IMgrs, and that this model will positively influence more-traditional ways of conducting science.

Although IMgrs can help resolve technical issues that are barriers to analysis, another issue that was repeatedly mentioned during the workshop remains. As in many areas of science (Anonymous 2009, Nelson 2009), there is no formal data sharing and access policy for CTFS data. The understood policy is that each research site controls access to its own data, and has the choice of web publication to share the data or not. At this time, only Barro Colorado Island CTFS data and Luquillo LTER data are openly accessible. The wish of scientists at this meeting is that CTFS data be shared in at least a limited way so that cross-site analyses can be done. The scientists expressed the desire to always have control of their data, but to be able to expose single columns of data, or a single census, for public use. Development of a data policy for CTFS that reflects the wishes of the scientists is needed.

Open-sourcing ecological data has been the topic of much recent discussion (Cassey and Blackburn 2006, Parr 2007). As Forest Dynamics Plot (FDP)-based studies become increasingly broad across different forest types, the data management issues, such as data sharing and ease of repeatability of complex analyses, will need to be addressed. Tools developed by the ecoinformatics community and the rise of open- source software coupled with collaborative content building have challenged the old models of intellectual property and notions about the best ways to foster creativity, progress, and quality (Parr 2007). In the future, open-source approaches to data sharing will become a new form of science (Penev et al. 2009). Therefore, development of a data policy for forest dynamics plots of ILTER should be seriously considered.

This workshop laid the foundation for the next workshop, in which we will use the tools built in this workshop to analyze data from other countries and develop more- complex workflows. Scientists of FDPs are becoming more interested in focusing on workflow applications after seeing them demonstrated during the workshop and recognizing their value as mechanisms to automate, document, and make repeatable scientific processes (Ludäscher et al. 2010). Workflows have emerged as a paradigm for representing and managing complex distributed computations, and they can be used to accelerate the pace of scientific progress (Gil et al. 2007). We expect the next workshop to bring together domain scientists and IMgrs again to discuss more CTFS-related computational and visualization needs and the challenges of linking these tasks together as workflows. We anticipate that workflows developed in the next workshop will be integrated with the package of tools available from CTFS, reducing the user's need for skills with the R package.

We conclude that the framework prototyped in this workshop should be useful to the forest dynamics research community through the establishment of mutualistic relationships between scientists and IMgrs. Although the functions of this framework have not immediately resolved all metadata and data-sharing problems, it provides a collaborative way to link CTFS databases. This successful workshop provides a valuable foundation for future interactions between science and technology specialists in the EAP ILTER.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Hen-biau King and Dr. William Chang for supporting information management activities between Taiwan and US LTER since 2004. We also thank Dr. Jeen Lian Hwong and colleagues of the Lienhuachih Research Center and plot scientists for participating and supporting logistics during the workshop in June 2009.

LITERATURE CITED

Altintas I, Berkley C, Jaeger E, Jones M, Ludascher B, Mock S. 2004. Kepler: an extensible system for design and execution of scientific workflows. Proceedings of the 16th International Conference on Scientific and Statistical Database Management. 21-23 June, Santorini Island, Greece. IEEE Press. p 423-4.

Anonymous. 2009. Data's shameful neglect. Nature 461:145.

Ashton PS, Boscolo M, Liu J, LaFrankie JV. 1999. A global programme in interdisciplinary forest research: the CTFS perspective. J Trop For Sci 11:180-204.

Berkley C, Jones M, Bojilova J, Higgins D. 2001. Metacat: a schema-independent XML database system. Proceedings of the 13th International Conference on Scientific and Statistical Database Management. 18-20 July, George Mason University, Virginia, USA. p 171-9.

Cassey P, Blackburn, TM. 2006. Reproducibility and repeatability in ecology. Bioscience 56:958-9.

Condit R. 1995. Research in large, long-term tropical forest plots. Trends Ecol Evol 10:18-22.

Fegraus E, Andelman S, Jones, MB, Schildhauer MP. 2005. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. ESA Bull 86:158-68.

Gil Y, Deelman E, Ellisman M, Fahringer T, Fox G, Gannon D, et al. 2007. Examining the challenges of scientific workflows. IEEE Comp 40:24-32.

Higgins D, Berkley C, Jones M. 2002. Managing heterogeneous ecological data using Morpho. Proceedings of the 14th International Conference on Scientific and Statistical Database Management. 24-26 July, Edinburgh, Scotland and Los Alamitos, CA: IEEE Computer Society. p 8.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comp Graph Stat 5:299-314.

Jones M, Berkley C, Bojilov AJ, Schildhauer M. 2001. Managing scientific metadata. IEEE Int Comp 5:59-68.

Lin CC. 2008. Ecological information management system for an East and Pacific International Long Term Ecological Research Network. *FORTROP II* International Workshop: Tropical Forestry Change in a Changing World, 17-20 Nov 2008. Bangkok: Kasetsart Univ. p 64.

Lin CC, Porter JH, Hsiao CW, Lu SS, Jeng MR. 2008. Setting an EML-based data management system for automating analysis of field sensors data. Taiwan J For Sci 23:279-85.

Lin CC, Porter JH, Lu SS. 2006. A metadata-based framework for multilingual ecological information management. Taiwan J For Sci 21:377-82.

Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, et al. 2006. Scientific workflow management and the Kepler system. Concurrency and computation: practice and experience. Spec Issue Sci Workflows 18:1039-65.

Ludäscher B, Altintas I, Bowers S, Cummings J, Critchlow T, Deelman E, et al. 2010. Scientific process automation and workflow management. Chap. 13. In Scientific data management: challenges, technology, and deployment, Shoshan A, Rotem D, editors. Boca Raton, FL: Chapman & Hall/CRC Computational Science. p 467-508.

Michener WK. 2000. Metadata. In: Ecological data: design, management and processing. Michener WK, Brunt JW, editors. Blackwell Science. p 92-116.

Michener WK, Brunt JW, Vanderbilt KL. 2002. Ecological informatics: a long-term ecological research perspective. CODATA 2002:1-6.

Nelson B. 2009. Data sharing: empty archives. Nature 461:160-3.

Parr CS. 2007. Open sourcing ecological data. Bioscience 57:309-10.

Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C. 2009. Publication and dissemination of datasets in taxonomy: ZooKeys working example. ZooKeys 11:1-8.

Pennington D. 2010. Enabling science and technology research teams: a bread making metaphor. Educ. Available at http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/EnablingScienceandTechnologyRe/199387.

Zhang J, Altintas I, Tao J, Liu X, Pennington DD, Michener WK. 2006. Integrating data grid and web services for E-science applications: a case study of exploring species distributions. Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (e-Science'06). Washington, DC: IEEE Computer Society. p 31.