Research note

Filling Information Management Gaps of Forest Dynamics Plot Databases Using Ecological Metadata Language and a Scientific Workflow System

Chau-Chin Lin,[1] Chi-Wen Hsiao,[1] Sheng-Shan Lu,[1]
Wen-Liang Chiou,[2] Li-Wan Chang,[3] Meei-Ru Jeng[3,4]

【Summary】

The study of forest dynamics plots began in the 1970s. A great deal of plot data accumulated through censuses at each site. The wealth of these datasets presents challenges in information management for researchers. Therefore, a purposely designed workshop for forest dynamics plot data management was held on June 15~19, 2009 at Lienhuachih Research Center of the Taiwan Forestry Research Institute. This paper describes preliminary results of the workshop that produced an integrated information management system. This framework attempts to facilitate the effective use of fully documented data archives for data discovery, access, retrieval, analysis, and integration. Results from our work include the Fushan and Leinhuachih databases based on the Center for Tropical Forest Science (CTFS) structure in a local MySQL server, an authentication interface, a metadata query web page, and 2 workflows.

Key words: forest ecology, Fushan, Lienhuachih, metadata, LTER.

Lin CC, Hsiao CW, Lu SS, Chiou WL, Chang LW, Jeng MR. 2010. Filling information management gaps of forest dynamics plot databases using Ecological Metadata Language and a scientific workflow system. Taiwan J For Sci 25(1):97-105.

[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] Botanical Garden Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Taipei 10066, Taiwan. 林業試驗所植物園組，10066台北市南海路53號。

[3] Technical 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。

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

Received August 2009, Accepted December 2009. 2009年8月送審 2009年12月通過。

研究簡報

以生態後設資料語言及科學工作流程系統 輔助森林動態樣區資訊管理

林朝欽[1)] 蕭其文[1)] 陸聲山[1)] 邱文良[2)] 張勵婉[3)] 鄭美如[3,4)]

摘　要

森林動態樣區研究起於1970年代，迄今各樣地與各次調查已累積相當豐富的原始資料，對研究人員而言這些龐大的資料集也呈現出資料管理的挑戰。為協助森林動態樣區資訊管理，2009年6月15~19日期間在林業試驗所蓮華池研究中心舉辦了一場研習會，本文為研習會中經討論後設計的一個森林動態樣區整合性的資訊管理架構及測試之初步結果。此架構主要功能是為協助森林動態樣區資料完整的紀錄與倉儲，並能在倉儲後有效的提供查詢、取用與分析整合。依此架構，我們將福山及蓮華池森林動態樣區資料庫按美國熱帶森林科學中心(Center for Tropical Forest Science)的規範重建在本地端，此資料庫需認證方能取得原始資料，但可以查詢到此兩動態樣區的其他基本資料，這些基本資料是透過生態後設資料語言(Ecological Metadata Language)所建立的，最後我們以福山及蓮華池之資料，使用科學工作流程系統測試了整個資訊管理系統。

關鍵詞：森林生態、福山、蓮華池、後設資料、長期生態研究。

林朝欽、蕭其文、陸聲山、邱文良、張勵婉、鄭美如。2010。以生態後設資料語言及科學工作流程系統輔助森林動態樣區資訊管理。台灣林業科學25(1):97-105。

The study of forest dynamics plots (FDPs) began in the 1970s. The first 50-ha FDP was established on Barro Colorado Island (BCI) in Panama by the Smithsonian Tropical Research Institute in 1980. The Center for Tropical Forest Science (CTFS) has led the study of FDPs, and is committed to the study of tropical and temperate forest function and diversity (Condit 1995). The network is unified by common plot structures and scientific methodologies. Currently, a plot requires 25~50 ha in area and a standardized inventory method. All free-standing trees and saplings with a diameter at breast height (DBH) of at least 1 cm are tagged, measured, identified to species, and resurveyed every 5 yr. This methodology assures strict comparability between sites and allows the development of general models for the dynamics of forests. The inventory gathers demographic information on individual tree species, which aims to provide long-term information on forest composition and dynamics so that changes can be predicted (Condit 1995). In addition, data provided by the plots also serve as a control of undisturbed forests for studies of anthropological impacts on and management of forests. Many scientific papers have been published using data from different plots to study factors involved in population regulation and the maintenance of diversity, changes in species and in climate, and models of demographics of individual species (Bakker et al. 1996, Rees et al. 2001). The results from these plots have obviously expanded the original focus on community ecology, the maintenance of diversity, and tropical forests to include forest management, interdisciplinary research, and broader forest types (Condit 1995).

A great deal of plot data have accumulated through censuses at each site. The wealth of these datasets presents challenges in information management for researchers. For example, to fulfill the potential of databases, they need to meet several requirements. First, the data need to be stored in a way suitable for long-term survival. Second, the data need to be readily accessible. Third, the data need be supported by clear descriptions of the context in which the study was undertaken (Magnuson 1990, Le Due et al. 2007, McIntosh et al. 2007). A database system designed by the CTFS attempts to meet these requirements. This database system consists of standardized databases, a hierarchical series of web forms for data entry and data uploading, and Hypertext Preprocessor (PHP) scripts that allows the user to check and extract data from the database. Each member site is provided with a user name and password. Data sharing is controlled at the owner site's discretion. However, for the purpose of information management, the existing system is not versatile enough. It lacks metadata of the plot which easily allows data discovery. In addition, researchers working on CTFS plots have been frustrated by the lack of user-friendly analytical tools to analyze and visualize the data, and cross-site analysis has been limited despite the standardization of methodology. Although the CTFS website provides analytical tools for scientists to use, these tools are built with R project language, and scientists face the formidable task of learning some programming in R.

Establishing an information management system for long-term ecological research datasets has been ongoing and a high priority for the East Asia Pacific Region of the International Long Term Ecological Research (EAP-ILTER) member networks (Lin et al. 2006). The community of information managers has participated in a series of training workshops for scientists and information managers in this region to promote its development. A common and compatible information management system is a requisite for data to be efficiently and effectively shared, exchanged, and synthesized. Several forest dynamics research projects in the EAP-ILTER are also a subset of the CTFS network. EAP-ILTER information managers recognized that ILTER scientists working on large forest plots could benefit from existing information management expertise and information management systems. Collaboration by EAP-ILTER information managers with their US counterparts to improve ecological research through using information, identifying information management system requirements, and building capacity for using data through workshops since 2004 has been very successful. A workshop for FDP information management was designed and held on June 15~19, 2009 at the Lienhuachih Research Center of the Taiwan Forestry Research Institute (TFRI).

The objectives of this workshop were to determine obstacles that researchers have encountered with data application and build a system to resolve the identified problems. We grouped FDP researchers and information managers from Taiwan, Japan, Malaysia, and the US during the workshop. In each group, ecologists were encouraged to describe their needs. Then information managers proposed designs and a possible prototype solution accordingly.

This paper describes the preliminary results of the workshop that produced an integrated information management framework. Figure 1 displays the conceptual framework of our information management system for FDP databases. Three components are included in this system.

The first component is data storage. This component includes 2 database types. The 1st type of database is the census database which is a relational database using MySQL (CTFS MySQL database, Fig. 1). This database is for validated and well-documented data archives of plots and is password-protected using the Lightweight Directory Access Protocol (LDAP) which serves as the authentication control. The other type of database is the metadata database which is a schema-independent eXtensible Markup Language (XML) database using PostgreSQL. This database is based on Ecological Metadata Language (EML) schemas. EML is a metadata standard developed by and for the ecology discipline (Fegraus et al. 2005). The database structure is a storage subsystem of Metacat (short for metadata catalog) framework designed by the National Center for Ecological Analysis and Synthesis (NCEAS) at the Univ. of California,

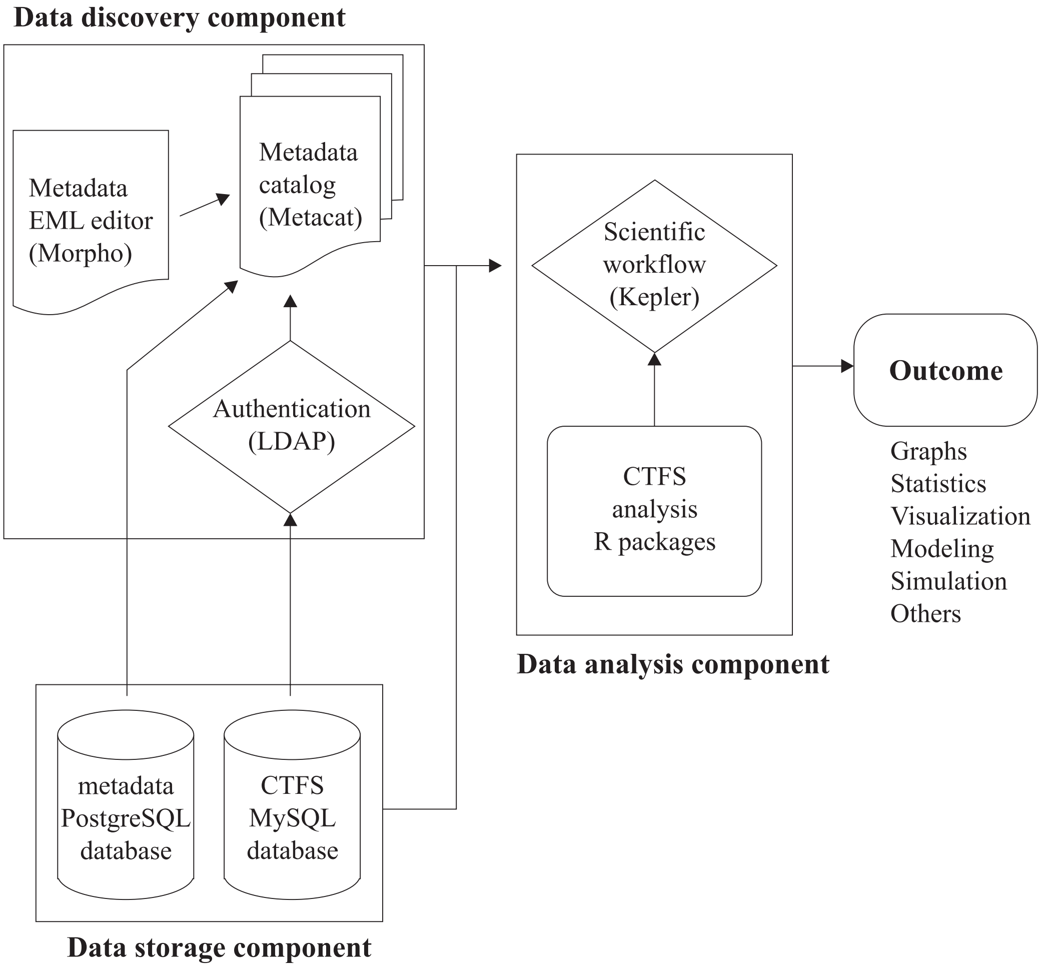


Fig. 1. Framework of the forest dynamics plot information management system. The MySQL database and Metacat are linked by an authentication interface. The system shows that metadata consist of general information for users, and the collected dataset is open to users. However, the database has a password-protected entry system.

Santa Barbara, CA (Jones et al. 2001, Fegraus et al. 2005). It is a hybrid relational database by storing XML data with arbitrary schemas. This approach permits structured, path-based queries of metadata.

The 2nd component is data discovery which handles metadata editing and querying. We choose Morpho as the metadata editor. This is one of the tools developed by the Knowledge Network for Biocomplexity (KNB) Project of the US (http://knb.ecoinformatics.org) (Higgins et al. 2002). Morpho is a cross-platform which can be used on Windows, Linux, and Mac computers. It allows researchers to describe their data using comprehensive and flexible metadata specifications, and to share their data publicly. Users can store their metadata either locally or on a remote server. In addition to Morpho, a data catalog web interface which is another subsystem of the Metacat framework, was included. It is a simple but powerful querying interface to assist in locating useful datasets registered within the Metacat storage subsystem. The Metacat uses LDAP as the authentication control for those datasets not open to the public.

The 3rd component is data analysis. We used Ecogrid, which is a collection of distributed ecological, biodiversity, and environmental data and analytic resources, to provide a uniform and simple programming interface to access data and metadata. Kepler is a community-driven, open-source project, and a particular scientific workflow system (Ludäscher et al. 2005). We chose Kepler as the uniform tool to allow scientists to design, execute, and monitor analytic procedures with minimal effort. Kepler links metadata databases to execute data acquisition, integration, transformation, synthesis, and archival information (Michener et al. 2007).

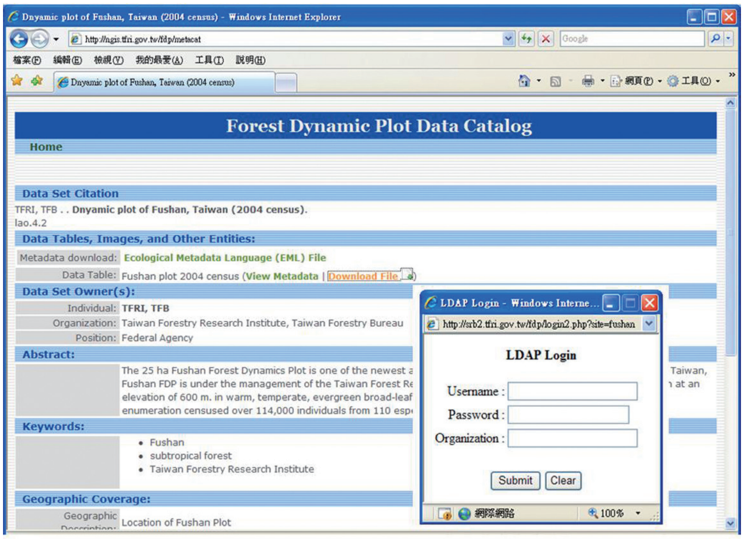
Our framework was tested using data from 2 forest dynamics plot data of Taiwan, Fushan and Leinhuachih. Fushan is located in northeastern Taiwan. The plot is 25 ha in a subtropical evergreen forest. This plot has been censused twice since 2004. We use data from the 1st censuses of both plots. Leinhuachih is located in central Taiwan. It is also a 25-ha plot in a subtropical evergreen forest and was censused only once in 2008. The 2 databases were created by Dr. Yu Yun Chen of the National Center for Theoretical Sciences, Mathematics Division, Hsinchu, Taiwan using the CTFS database system structure and stored on the CTFS server at Harvard Univ., Boston, MA. They provided backup files for restoring the databases, and these files are on a TFRI server.

Results from our work include Fushan and Leinhuachih census databases, an LDAP authentication interface, a metadata query web page, and 2 workflows designed and tested using data stored in the TFRI server.

Both the Fushan and Lienhuachin census databases consist of 34 original linked tables and other additional tables which are created for storing results of querying or exporting. The complete census data of each plot were merged into a table called dftemp. It can be exported as an ASCII file to provide analysis or integration. There are 166,591 records for the Fushan plot and 203,313 for the Lienhuachih plot. Briefly, tables of the database based on the CTFS database structure can be grouped into 8 catalogs: data entry, measurement and remeasurement, tree location, site and plot information, taxonomic information, and logs of changes. The measurement and remeasurement tables are linked to most other tables. They contain all tree survey data. There are site and plot tables corresponding to each individual plot. The tree location tables relate to all trees in a plot which are labeled with tree tags. The taxonomic information tables consist of codes for the family, genus, and species name of each tree with the primary key of species identification for each tree. The log of text change and measurement change tables records all modifications after the first upload of census data. These 2 databases are password-protected by an authentication server and serve only as data storage in our framework (Fig. 2).

Two EML standard metadata documents for the Fushan and Lienhuachih plots were created. The metadata we created are stored in a Postgresql database and can be searched and retrieved by a query interface (Fig. 2). All metadata include 5 information sections: a general description, geographic description, temporal description, taxonomic classification, census methods, and description of the data tables. The general description documents the purpose of data collection and original intentions such as the title, abstract of the project, and plot. The geographic, temporal, and taxonomic sections describe the place, time, and species or organism information. The data table section is the most useful part of the EML document. It contains the definition of each field in the CTFS database. This section contains details that would be useful to users with little prior knowledge of the dataset. Additionally, this sort of detailed information on the attributes facilitates future analyses and exportation of dataset patterns.

The framework was tested using the restored databases of Fushan and Lienhuachih. We retrieved 2-ha data from each database for the test. Two scientific workflows were created using Kepler. The first workflow was simple statistics to compare the diameter ranges (Fig. 3). Using Ecogrid, data of the 2 plots were directly retrieved from the MySQL databases on the TFRI server. Once the 2 datasets were downloaded, EML documents provide field names to be selected for further calculation. In this example, we chose diam-


Dnyamic plot of Fushan, Taiwan (2004 census) - Windows Internet Explorer
http://ngis.tfri.gov.tw/fdp/metacat
Google
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
Dnyamic plot of Fushan, Taiwan (2004 census)
網頁(P) 工具(O)
Home
LDAP Login - Windows Interne...
http://srb2.tfri.gov.tw/fdp/login2.php?site=fushan
Submit
Clear
網際網路
100%

Forest Dynamic Plot Data Catalog

Data Set Citation

TFRI, TFB . . **Dnyamic plot of Fushan, Taiwan (2004 census).**
lao.4.2

Data Tables, Images, and Other Entities:

Metadata download: Ecological Metadata Language (EML) File
Data Table: Fushan plot 2004 census (View Metadata | Download File)

Data Set Owner(s):

Individual: **TFRI, TFB**
Organization: Taiwan Forestry Research Institute, Taiwan Forestry Bureau
Position: Federal Agency

Abstract:

The 25 ha Fushan Forest Dynamics Plot is one of the newest a Taiwan,
Fushan FDP is under the management of the Taiwan Forest Re ı at an
elevation of 600 m. in warm, temperate, evergreen broad-leaf
enumeration censused over 114,000 individuals from 110 espe

Keywords:

- Fushan
- subtropical forest
- Taiwan Forestry Research Institute

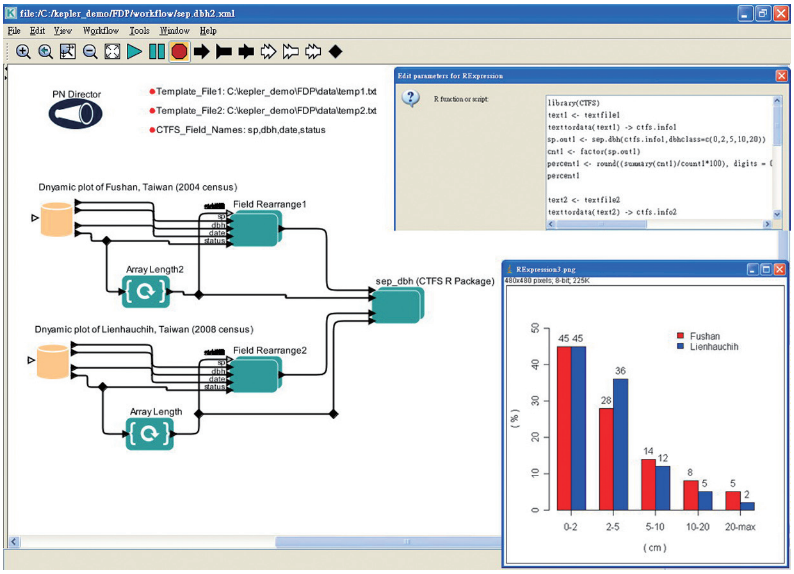
Geographic Coverage:

Geographic Description: Location of Fushan Plot

LDAP Login

Username :
Password :
Organization :

Fig. 2. Databases of Fushan and Lienhuachih are password protected by an authentication server powered by LDAP. When users browse the data catalog and try to download raw data of the census, the system will produce a pop-up window for the user name, password, and organization for authentication.


file:/C:/kepler_demo/FDP/workflow/sep.dbh2.xml
File Edit View Workflow Tools Window Help
PN Director
Dnyamic plot of Fushan, Taiwan (2004 census)
Field Rearrange1
sp
dbh
date
status
Array Length2
sep_dbh (CTFS R Package)
Dnyamic plot of Lienhauchih, Taiwan (2008 census)
Field Rearrange2
Array Length
Edit parameters for RExpression
R function or script:
REXpression3.png
480x480 pixels; 8-bit; 225K
(%)
0
10
20
30
40
50
45 45
36
28
14
12
8
5
5
2
0-2
2-5
5-10
10-20
20-max
(cm)
Fushan
Lienhauchih

●Template_File1: C:\kepler_demo\FDP\data\temp1.txt

●Template_File2: C:\kepler_demo\FDP\data\temp2.txt

●CTFS_Field_Names: sp,dbh,date,status

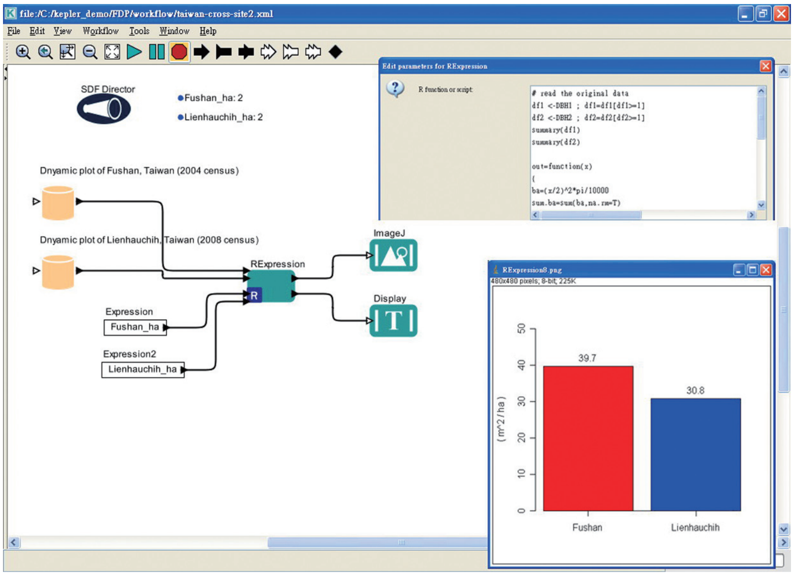
```
library(CTFS)
text1 <- textfile1
texttordata(text1) -> ctfs.info1
sp.out1 <- sep.dbh(ctfs.info1,dbhclass=c(0,2,5,10,20))
cnt1 <- factor(sp.out1)
percent1 <- round((summary(cnt1)/count1*100), digits = (
percent1

text2 <- textfile2
texttordata(text2) -> ctfs.info2
```

Fig. 3. Scientific workflow designed for a simple comparison of the size class distributions of the Fushan and Lienhuachih plots.

eter at breast height (DBH) as the variable to place in the data array and sorted the data into 5 size classes. In total, 4 actors of Kepler were used to process the calculation. The result of this procedure was a bar chart of size classes using a graph display actor of Kepler. The result shows that most trees censused in the 2 plots were < 5 cm in DBH (Fig. 3). The 2nd workflow compared abundance differences between the Fushan and Lienhuachih plots. We used the abundance.spp function of the CTFS R package in the workflow. We selected 2 actors to compute abundances. The result showed that Fushan's 2004 census had a higher abundance than Lienhuachih's 2008 census (Fig. 4). The 2 workflows showed that the framework can be used to fill in gaps in the current CTFS database system: the lack of a data discovery procedure and a friendly analytical interface.

This was a unique workshop in Taiwan that gathered researchers and information managers to work together for the first time for discussion, examination and design to fulfill common needs in using FDP datasets. The results showed that more communication between researchers and information managers is needed. While information managers can help solve technical issues that are barriers to analysis, the issues of formal data sharing and access policy for other plots still remain to be solved. The debate of open sourcing ecological data was recently discussed (Cassey and Blackburn 2006, Parr 2007). As FDP-based studies become increasingly broad in different forest types, data management issues such as data sharing and repeating complex analyses easily should be examined. Tools developed by the ecoinformatics community and the rise of open source software and collaborative content building have challenged the old model 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


file:/C:/kepler_demo/FDP/workflow/taiwan-cross-site2.xml
File Edit View Workflow Tools Window Help
SDF Director
●Fushan_ha: 2
●Lienhauchih_ha: 2
Dnyamic plot of Fushan, Taiwan (2004 census)
Dnyamic plot of Lienhauchih, Taiwan (2008 census)
RExpression
R
Expression
Fushan_ha
Expression2
Lienhauchih_ha
ImageJ
Display
T
Edit parameters for RExpression
R function or script:
REXpression8.png
480x480 pixels; 8-bit; 225K
(m^2 / ha)
0
10
20
30
40
50
39.7
30.8
Fushan
Lienhauchih

```
# read the original data
df1 <-DBH1 ; df1=df1[df1>=1]
df2 <-DBH2 ; df2=df2[df2>=1]
summary(df1)
summary(df2)

out=function(x)
{
ba=(x/2)^2*pi/10000
sum.ba=sum(ba,na.rm=T)
```

Fig. 4. Scientific workflow to compare abundances of the Fushan and Lienhuachih plots using the CTFS abundance R statistical package. The R codes are embedded in the workflow which runs R code in the background.

et al. 2009). Therefore, development of a data policy for forest dynamics plots in Taiwan should be seriously considered.

We conclude that the framework prototype developed in this workshop should be useful to the forest dynamics research community. Although the functions of this framework have not immediately solved metadata and data-sharing problems, it provides a collaborative way to link CTFS databases without conflicting with the protection of data use rights.

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. Data collection at the Fushan and Lienhuachih forest dynamics plots was supported by the Taiwan Forestry Research Institute and Taiwan Forestry Bureau. We are thankful for the grants and many scientists involved allowing these datasets to become a reality. Finally, we thank Dr. Jeen Lian Hwong and colleagues at the Lienhuachih Research Center for providing logistical support during the workshop in June 2009.

LITERATURE CITED

Bakker JP, Olff H, Willems JH, Zobel M. 1996. Why do we need permanent plots in the study of long-term vegetation dynamics? J Veg Sci 7:147-56.

Cassey P, Blackburn TM. 2006. Reproducibility and repeatability in ecology. Bioscience 56(12):958-9.

Condit R. 1995. Research in large, long-term tropical forest plots. Trends Ecol Evol 10(1): 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(3):158-68.

Higgins D, Berkley C, Jones MB. 2002. Managing heterogeneous ecological data using Morpho. Proceedings of the 14th international conference on Scientific and Statistical Database Management (SSDBM'02); 2002 July 24-26; Edinburgh, Scotland; Los Alamitos, CA: IEEE Computer Society. 8 p.

Jones MB, Berkley C, Bojilova J, Schildhauer M. 2001. Managing scientific metadata. IEEE Internet Comput 5(5):59-68.

Le Duc MG, Yang L, Marrs RH. 2007. A database application for long-term ecological field experiments. J Veg Sci 18(4):509-16.

Lin CC, Porter JH, Lu SS. 2006. A metadata-based framework for multilingual ecological information management. Taiwan J For Sci 21(3):377-82.

Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, Lee EA, Tao J, Zhao Y. 2006. Scientific workflow management and the Kepler system. Concurr Comput Pract Ex. 18(10):1039-65.

Magnuson JJ. 1990. Long-term ecological research and the invisible present. Bioscience 40(7):495-501.

McIntosh ACS, Cushing JB, Nadkarni NM, Zeman L. 2007. Database design for ecologists: composing core entities with observations. Ecol Info 2(3):224-36.

Michener WK, et al. 2007. A knowledge environment for the biodiversity and ecological sciences. J Intell Info Syst 29(1):111-26.

Parr CS. 2007. Open sourcing ecological data. Bioscience 57(4):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.

Rees M, Condit R, Crawley M, Pacala S, Tilman D. 2001. Long-term studies of vegetation dynamics. Science 293:650-5.