Available online at www.sciencedirect.com
ScienceDirect
Dendrochronologia 28 (2010) 131–134
DENDROCHRONOLOGIA
www.elsevier.de/dendro

TECHNICAL NOTE

A new approach to dendrochronological data management

Peter W. Brewer[a,*], Kit Sturgeon[b], Lucas Madar[c], Sturt W. Manning[a]

[a]*Cornell Tree-Ring Laboratory, B-48 Goldwin Smith Hall, Cornell University, Ithaca, NY 14853, USA*
[b]*The Epison Group LLC, 971 Taughannock Blvd, Ithaca, NY 14850, USA*
[c]*Portland State University, Portland, OR 97207, USA*

Received 19 November 2008; accepted 4 March 2009

Abstract

The procedures undertaken during the construction of chronologies often result in large quantities of data, typically in many associated files. The management of these data files, especially maintaining the relationships between processed and raw data, becomes increasingly difficult as a collection grows. To maintain a high level of accountability it is necessary to ensure that any derived or processed data can be replicated by independent researchers. Due primarily to practical limitations, some authors publish only subsets of the processed and/or raw data that constitute a chronology, making the process of independent scientific scrutiny difficult and on occasions impossible. There are associated problems when master chronologies are utilised for the purpose of dating specimens. Maintaining the relationships between chronologies and specimens is extremely important for accountability purposes, especially when errors are detected or revisions required. In these circumstances it is highly desirable to locate all specimens that have been dated with the original chronology to ascertain the implications of changes. This paper describes a new relational database design that addresses these problems and which has been implemented in the Corina dendrochronology application and web service. The new Corina system enables the data and analyses for all chronology building and cross-dating processes to be stored and documented, enabling scrutiny at every stage.

Keywords: Relational database; Data standards; Data integrity; Corina; Web service

Introduction

The chronology construction process in dendrochronological research includes many stages of data processing, normalisation and analysis. Many dendrochronological software packages have been released and successfully used over the years to perform these tasks including: DPL (Holmes, 1983, 2001), CATRAS (Aniol, 1983), dplR (Bunn, 2008), TSAP (Rinn, 2008) and PAST (Knibbe, 2008). These applications produce significant quantities of data, and whilst great efforts and advances have been afforded on their analytical functionality, the problems of data management are often left to the end user. Comprehensive data management is essential to enable independent scrutiny and replication of results, without which research can be called into question.

Data management problems

The current file-based nature of data storage within most existing dendro software tends to be either via a collection of file system folders of varying complexity or through a bespoke database running alongside the file system. The latter solution, being greatly preferable over

*Corresponding author.
E-mail address: p.brewer@cornell.edu (P.W. Brewer).

1125-7865/$ - see front matter © 2009 Elsevier GmbH. All rights reserved.
doi:10.1016/j.dendro.2009.03.003

the former, still suffers from significant data integrity problems. The underlying dendro data are exposed as simple files on the user's computer and these are too readily moved, renamed or deleted.

Data management problems are exacerbated by data processing. For example, when using existing dendro software, performing standard analyses such as the indexing of raw tree ring measurement series to remove age/growth relationships requires a user to choose between overwriting the original raw data file or doubling the number of files they have to manage. The issues become more pronounced when processed files are combined into chronologies, as the user is then required to keep track of which files have been cross-dated and combined.

Chronologies are mathematical abstracts of raw tree-ring width data, and when they are presented in isolation from the original ring measurements they are difficult to validate independently. The community should be able to ask the following questions of the dataset:

1. What method of indexing was performed on each series?
2. How many trees were used to build the chronology?
3. How many years of overlap are there between each tree?
4. What criteria were used to select the cross-date between series?

Defending the integrity of particular chronologies and the discipline as a whole requires robust validation of data, especially in the face of sharp criticism. Ideally, there should be comprehensive accountability for published chronologies enabling independent researchers to replicate the entire data-processing procedure. The current file-based data storage used by most researchers makes such transparency and accountability extremely difficult and, in many cases, impossible.

Design requirements

Large dendro databases such as the International Tree-Ring Databank (Grissino-Mayer and Fritts, 1997) and the WSL Dendro Database (Schmatz et al., 2001) have been used by the community for many years. These are, however, designed for data archiving and collaboration purposes rather than day-to-day data management in a particular laboratory. To address the issues outlined in the previous section a relational database management system (RDBMS) is required that must be:

1. capable of storing ring-width measurements directly without reference to external files;
2. capable of storing all the associated information (metadata) that researchers require for their research;
3. capable of storing or tracking analyses performed on raw ring-width data;
4. able to trace a chronology back to the original raw ring-width measurements and wood specimens.

Such a database would need to be either built as part of the software package performing the analyses or accessible as a plug-in. This in itself presents a challenge to the community considering the widespread use of legacy applications built in the 1980s and the 1990s for which there is limited or no current development. In the case of legacy commercial applications, the implementation of new features is impossible as the community does not have access to the original source code. The community must therefore concentrate either on freely accessible open source projects or request that the commercial companies still active in the dendro community implement these features on our behalf.

Implementing a solution in Corina

Corina is an open source application developed by the Malcolm and Carolyn Wiener Laboratory for Aegean and Near Eastern Dendrochronology, Cornell University, and released under the GNU General Public License.[1] Corina is a desktop application that runs on most popular operating systems and includes functionality for measuring, storing, analysing and presenting tree-ring data. Like most dendrochronological software, Corina originally utilised file-based data storage, but in order to address the key data management issues outlined above, we have implemented a web-services based, relational database management system. The database is implemented using the PostGreSQL software with both Java and PostGIS extensions and is accessed through a PHP web service. Both the desktop client and server software are available for download from the Cornell Tree-Ring Laboratory website[2] for installation by other laboratories around the world.

Corina database

The key concept of our database design is that of the virtual measurement series (v-series). The relationships between the database tables that implement the v-series concept are illustrated in Fig. 1. Whilst current software workflows mean a researcher has different versions of files on their computer from each step of an analysis, the only data supplied by the user is the raw measurement series. All other data are derived directly from calculations performed on this raw data. Corina reflects this fact by storing only raw ring-width data along with the algorithms and any required parameters that have been used in derived series.

[1] http://www.gnu.org/copyleft/gpl.html
[2] http://dendro.cornell.edu/corina/