

## SOFTWARE REPORT

# TRICYCLE: A UNIVERSAL CONVERSION TOOL FOR DIGITAL TREE-RING DATA

PETER W. BREWER<sup>1\*</sup>, DANIEL MURPHY<sup>1</sup>, and ESTHER JANSMA<sup>2,3</sup>

<sup>1</sup>Malcolm and Carolyn Wiener Laboratory for Aegean and Near Eastern Dendrochronology, Cornell University, Ithaca, NY 14853, USA

<sup>2</sup>Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

<sup>3</sup>Cultural Heritage Agency, Rijksdienst voor het Cultureel Erfgoed (RCE), Amersfoort, The Netherlands

### ABSTRACT

There are at least 21 dendro-data formats used in dendrochronology laboratories around the world. Many of these formats are read by a limited number of programs, thereby inhibiting collaboration, limiting critical review of analyses, and risking the long-term accessibility of datasets. Some of the older formats are supported by a single program and are falling into disuse, opening the risk for data to become obsolete and unreadable. These formats also have a variety of flaws, including but not limited to no accurate method for denoting measuring units, little or no metadata support, lack of support for variables other than whole ring widths (*e.g.* earlywood/latewood widths, ratios and density). The proposed long-term solution is the adoption of a universal data standard such as the Tree-Ring Data Standard (TRiDaS). In the short and medium term, however, a tool is required that is capable of converting not only back and forth to this standard, but between any of the existing formats in use today. Such a tool is also required to provide continued access to data archived in obscure formats. This paper describes TRiCYCLE, a new application that does just this. TRiCYCLE is an open-source, cross-platform, desktop application for the conversion of the most commonly used data formats. Two open source Java libraries upon which TRiCYCLE depends are also described. These libraries can be used by developers to implement support for all data formats within their own applications.

*Keywords:* TRiDaS, data standard, file format, dendrochronology, Java, data sharing.

### INTRODUCTION

Dendrochronologists have used computers to assist with the measurement and crossdating of tree rings since the 1970s. In the decades since then, a wide variety of computer programs have been written that rely upon many different data formats to store ring-width data, with each format exhibiting its own features, quirks and limitations. The plethora of formats in use today inhibits data transparency and accountability, limits collaboration, and hinders the development of new and innovative software tools.

The central unit of data in dendrochronology is the ring-width measurement. Most efforts at data

sharing focus on transferring these raw measurement values. However, researchers are increasingly realizing the need to share the metadata associated with these raw measurement values as well. It has long been customary to include basic information such as species and site name, but as dendrochronologists diversify into sub-disciplines, more detailed information such as GPS location, elevation, slope angle/aspect, soil type and tree height is routinely recorded. Many sub-disciplines (including dendroarchaeology, architectural dendrochronology and paleoecology) routinely work with wood samples that do not include bark, sapwood and/or pith, therefore additional metadata about the completeness of samples is vital during analysis. A number of dendrochronology applications include

---

\*Corresponding author: p.brewer@cornell.edu

**Table 1.** List of the 21 dendro-data formats supported by TRiCYCLE. The table highlights whether TRiCYCLE can read and/or write each format, and also indicates whether the format unambiguously supports absolutely dated, relatively dated and undated series.

Format	Read	Write	Absolute Dating	Relative Dating	Undated Series
Belfast Apple	✓	✓			✓
Belfast Archive	✓		✓		
Besançon	✓	✓	✓		✓
CATRAS	✓	✓	✓		✓
Comma Separated Values (CSV)	✓	✓	✓		
Corina legacy	✓	✓	✓	✓	
DendroDB	✓		✓		
Excel	✓	✓	✓		
Heidelberg	✓	✓	✓	✓	✓
Nottingham	✓	✓			✓
ODF spreadsheet	✓	✓	✓		
Oxford	✓	✓	✓	✓	✓
PAST4	✓	✓	✓		
Sheffield	✓	✓	✓	✓	
Topham	✓	✓			✓
TRiDaS	✓	✓	✓	✓	✓
TRIMS	✓	✓			✓
Tucson	✓	✓	✓		
Tucson Compact	✓	✓	✓		
VFormat	✓	✓	✓		
WinDendro	✓		✓		

the ability to store rich metadata directly with the raw measurements, but these rely upon proprietary data files. Accurately converting these rich metadata fields therefore adds an additional layer of complexity.

The long history of using computers in dendrochronology research inevitably means that there are a number of older programs that are no longer developed or supported. This increases the risk that data formats become obsolete and vast quantities of information become permanently inaccessible.

From a programmer's perspective, the variety of data formats is also an obstacle for the development of innovative new tools for data manipulation. Most programmers understandably choose to support only one or two data formats (most often the Tucson decadal format). Examples include COFECHA (Holmes 1983), ARSTAN (Cook and Holmes 1996) and the dpl-R (Bunn 2008). Even then, handling the various peculiarities of the format requires considerable effort and results in some programs reading and writing files that are rejected by other programs that claim to use the same format. Simply providing the user

with feedback on why a file is deemed invalid can require considerable programming effort. As such, many programs simply crash or provide generic error messages when faced with a file in an unexpected format. This typically leaves the user confused and frustrated, especially when formatting errors are subtle such as additional white space characters, or worse still caused by differences in hidden control characters such as line-feeds, which are often caused by different computing platforms.

## EXISTING DATA FORMATS

A total of 21 data formats have been identified as being of importance to the dendrochronology community (see Table 1). A number of these (*e.g.* TRIMS, Belfast Apple, Belfast Archive) are not commonly used today, but nevertheless, there are many thousands of data files in these formats archived in laboratories around the world. It is important to include support for such formats in the converter to ensure accessibility of data for years to come.

For most of these formats there is little or no documentation describing how the format works.

A number of the formats are still actively supported by their original designers but others have slipped out of use and/or their developers are no longer available for consultation, the initial indication of a format falling into obscurity. The first stage of this research has therefore been the collation of all the information that can be gleaned regarding these formats. A PDF document containing this information is available from the journal website as Supplementary Material and is also included as an appendix to the TRiCYCLE manual.

The identified formats vary in many important ways:

**Data variables stored.** Although all these formats can store whole ring-width measurements, many can store earlywood and latewood widths and some in addition can store various wood-density variables. A number of formats explicitly record the type of data that the files hold (*e.g.* PAST4, VFormat, Sheffield). Others (most notably the Tucson formats) were designed for storing ring-width data, and are now used for storing other variables. In these circumstances there is often little or nothing to indicate what type of data a file holds.

**Units of measurement.** Some formats explicitly note the units of measurement within the files (*e.g.* Heidelberg), whereas others record units by convention (*e.g.* Sheffield). Any converter tool must therefore be capable of detecting and handling unit conversion.

**Raw and chronology data.** Certain formats are designed for storing only raw measurement data (*e.g.* TRIMS, Belfast Apple) whereas others can also store processed chronology data (*e.g.* Tucson, Heidelberg and PAST4).

**Single or multiple series.** Some formats require a single file for each data series, whereas others can store a suite of series and/or variables. The conversion of a multi-series data file into a single-series data format will necessarily result in multiple files.

**Text or binary format.** The vast majority of dendrochronological data formats are text-based files that can be read by standard text-editor programs. However, the CATRAS format stores data in binary format and therefore requires specialist software to read it.

**Metadata.** One of the most important differences between the formats is the inclusion or exclusion of metadata. Some formats are completely deficient in this respect (*e.g.* TRIMS, Topham, Belfast Apple), whereas others include mostly free-text comments (*e.g.* Tucson).

**Standardization.** The formats that do include metadata vary in whether they standardize this information. For instance most Heidelberg fields are free-text, allowing users to enter any value in any language, whereas others like Sheffield and VFormat restrict users to a number of predefined options.

**Calendar.** Formats differ in the way they handle years. Most are based on the Gregorian calendar and include support for the BC/AD transition. Some however use the concept of an Astronomical calendar whereby the year zero is included which means years BC are offset by one year.

## DATA STANDARD

An obvious solution to the problems of data sharing would be the development and adoption of a universal dendro-data format. Since 2006, work has been progressing to this end, resulting in the release of the Tree Ring Data Standard (TRiDaS) in October 2008 (Jansma *et al.* 2010). TRiDaS is an XML-based data standard for recording dendrochronological data and metadata. More than 80 dendrochronologists, computer scientists and specialists from research disciplines that rely on dendrochronology have so far contributed to its development, including dendroarchaeologists, art and architecture historians, ecologists, geologists and climatologists. The standard is therefore capable of recording the wide variety of metadata required by these different fields. TRiDaS builds upon other established standards, such as GML (Geographical Markup Language) for the recording of locality information. The extensible nature of XML (Extensible Markup Language) also means that TRiDaS can evolve to accommodate the changing needs of dendrochronologists over time.

TRiDaS has the potential to replace the many existing data formats with a single unifying

format. However, at present the majority of the tools used by dendrochronologists rely upon the traditional data formats. Although it is hoped that TRiDaS will be adopted as a universal data standard within the community, in the intervening time a single conversion tool capable of converting between any combination of the existing formats is clearly desirable and indeed essential to enable such a transition. Such a tool is also essential to ensure data archived in old formats remains accessible.

Although there are a number of conversion tools already available, they do not support TRiDaS and typically convert from and to a limited number of formats (for example Grissino-Mayer's CONVERT5 and Holmes' YUX (Holmes 2001)). Existing converter tools typically support the conversion of data and do not provide a means for converting any associated metadata. A universal conversion tool has previously not been possible because this would require a routine for every combination of formats ( $n \times (n - 1)$ ). This is impractical for even a modest number of formats. For example, a converter that supports ten formats would require 90 routines. The converter would not be scalable, as the burden would become ever greater as support for more formats was added.

## REQUIREMENTS

A tool is required that can read and write the file formats listed in Table 1, enabling users to seamlessly convert data between formats. This tool should be able to read all available data and metadata from these formats. In circumstances where the data are ambiguous, the tool should intelligently assume the most likely meaning of the data while at the same time warning the user of its assumptions. When writing out data, the resulting file should be deemed valid by the original software that was written to handle such files.

The tool should be made available as a traditional desktop application that can be used by individual researchers running any popular operating system. It should also be made available in the form of a library that can be easily integrated into third-party applications so that

programmers can write new applications without dealing with the complexities of reading data. Any programmer making use of the library will therefore have immediate support for the full suite of data formats. The architecture of the tool should be such that additional formats can also be added quickly and efficiently.

The TRiCYCLE application and associated libraries described in this article fulfill all of these requirements. TRiCYCLE is an open-source desktop application available for all major operating systems including Microsoft Windows, Mac OSX and Linux. It is released under the Apache 2 open-source license, which means it can be used by anyone including commercial users (see the full license in the application for further details). The open-source license and modular architecture mean that the underlying libraries that read, write and convert dendro-data files can be used programmatically by developers within their own applications.

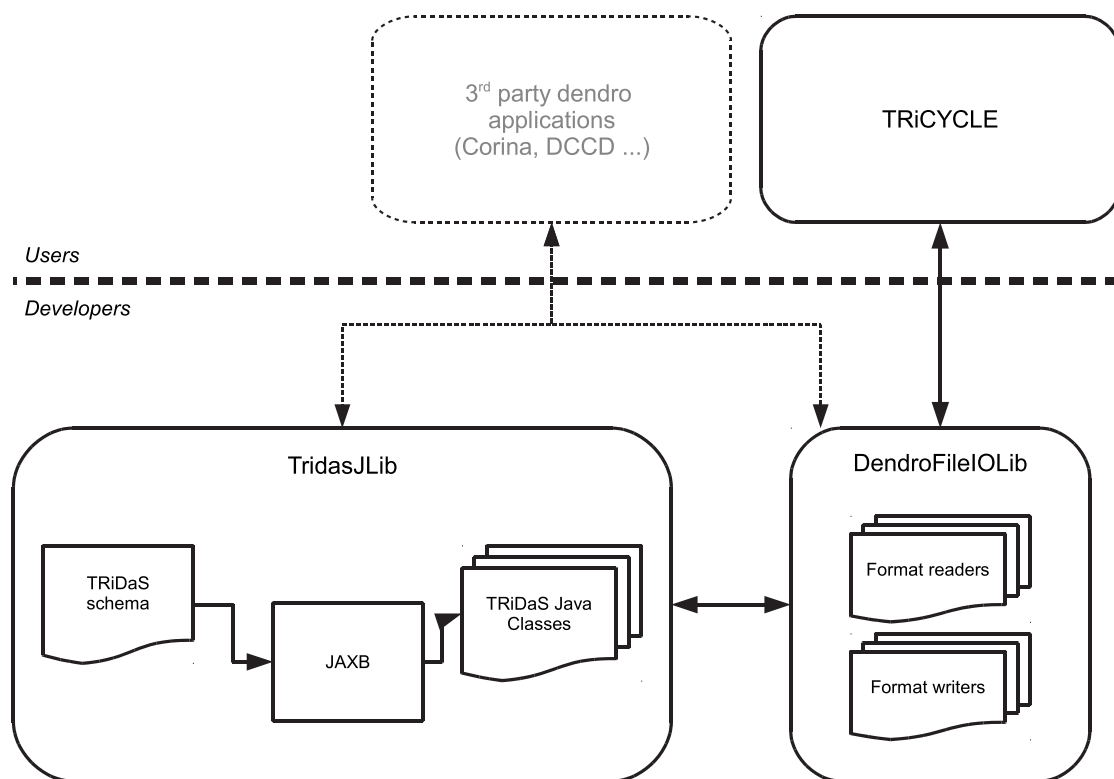
## SOFTWARE ARCHITECTURE

The key to solving the problem of writing a universal data converter has been the development of TRiDaS described by Jansma *et al.* (2010). The wide-ranging ability of TRiDaS to represent dendro-data and metadata accurately means that it is perfectly suited to act as an intermediate format. This means that in TRiCYCLE, only one reader and writer routine is required for each data format. Each reader is written to extract all the data and metadata available from a particular format and convert it into the TRiDaS data model. Conversely, each writer is designed to write out legacy format files from this same data model.

There are three distinct products that work together to produce the converter system: TridasJLib, DendroFileIOlib and the TRiCYCLE desktop application itself. The relationships between these packages are described below and are illustrated in Figure 1.

### TridasJLib

TridasJLib is a library of Java classes representing the TRiDaS data model along with



**Figure 1.** The relationship between the TRiCYCLE application and the TridasJLib and DendroFileIOLib libraries. Dendro applications can utilize the libraries to read and write dendro data as well as use the TRiDaS classes to manage and represent data internally. The components above the line are applicable to end users, whereas the components below are relevant to developers.

classes that are able to marshal and unmarshal TRiDaS-compliant data to and from TRiDaS XML files. Underpinning the TridasJLib is the TRiDaS XSD (XML Schema Definition). The TRiDaS XSD is a complete description of the TRiDaS standard including the names of all entities, fields and enumerations, information on which fields are mandatory and in what circumstances, and details about how these components fit together. The TridasJLib is largely an interpretation of the TRiDaS XSD into Java. Traditionally this interpretation would have been done manually, but the process of converting a data model from an XSD to Java classes is both time consuming and error prone. The TridasJLib is therefore produced automatically using JAXB (Java Architecture for XML Binding). This interprets the TRiDaS XSD automatically, and therefore as the TRiDaS schema

evolves, TridasJLib can be updated easily to reflect any changes.

### DendroFileIOLib

DendroFileIOLib is where the actual data conversion takes place. The library contains a reader and a writer for each supported dendro data format. Each reader contains the logic for converting data from a specific format into the TridasJLib Java class representations of the TRiDaS data model. Conversely each writer contains the logic for converting TridasJLib representations of TRiDaS projects into specific dendro data files. The library also contains infrastructure that is shared between all readers and writers, such as a conversion warning system that enables the comprehensive description of any problems and ambiguities encountered, as well as

a mechanism to report assumptions that need to be made for successful conversion.

## TRICYCLE

The final package presented here is the desktop application that allows users to easily utilize the DendroFileIOLib to convert dendro-data files. It is a graphical application that collects the information needed from the user (*e.g.* input files, output format) and then calls the Dendro-FileIOLib to do the conversion. It is designed to be intuitive to use and follows the normal installation and usage that users expect from a modern software application.

### USING TRICYCLE

Standard installation packages in eight languages (English, Dutch, French, Spanish, German, Polish, Turkish and Greek) are available for Mac OSX, Windows and Linux from the TRiDaS and DCCD (Digital Collaboratory for Cultural Dendrochronology in the Low Countries) websites ([www.tridas.org](http://www.tridas.org) and [www.dendrochronology.eu](http://www.dendrochronology.eu)). Further translations can be made available with the assistance of native speakers. Once installed, the application asks permission to collect anonymous usage statistics to assist with future development. It also periodically checks the [tridas.org](http://www.tridas.org) website for updates. Both these features can be disabled in the options menu if desired.

Once launched, the user is required to select one or more files to convert. This can be done via the file menu, by pressing the browse button, or by dragging files onto the application from the operating systems file manager. The user then needs to specify the format of these files from the pull down menu. If the user is unsure of the file type, then the 'Identify format' tool in the help menu can be used.

Once the input files and format have been defined, the user should then switch to the 'convert' page, where they can select the output format they require. After the user has pressed the 'convert' button, the results of the conversion are summarized in the table below.

Files that have converted successfully with no errors or warnings are indicated with a green tick

followed by the names of the output file or files (see Figure 2). Files that fail to convert (perhaps because of an invalid input file or because the requested output format is incapable of storing the type of data variable stored in the input file) are indicated with a red cross and an explanation of the error. Files that have been converted successfully but for which there are warnings are indicated with an orange exclamation sign. Warnings can be associated with the reader or writer operation and may be applicable to the whole input data file or just a single series within the input file if it is of a multi-series type. The warnings are displayed in a tree format to show the context of each warning.

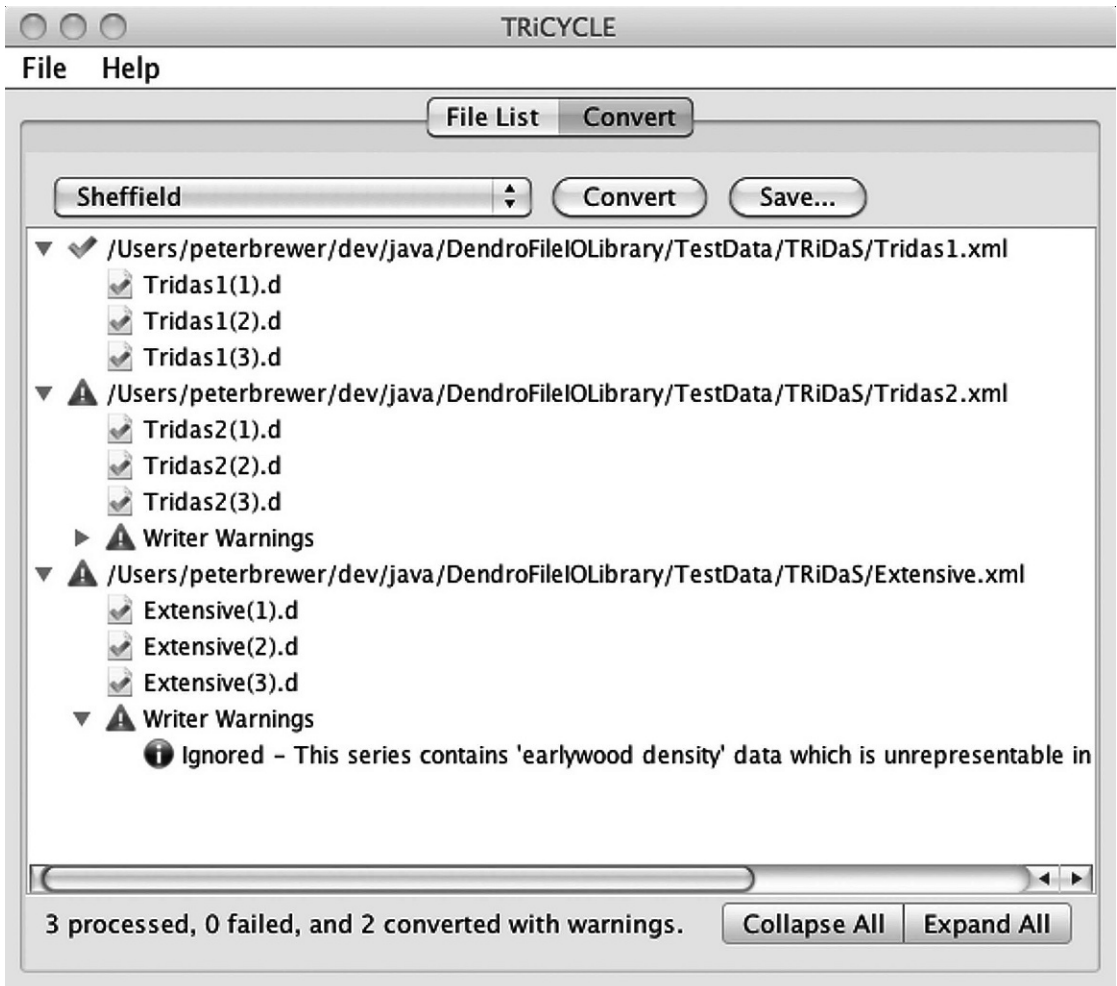
The user can preview files that have been successfully converted, by highlighting the file in the results table and then pressing the 'preview' button. Once the user is satisfied with the results, the save button can be pressed to permanently store the output files to disk.

### USE BY OTHER APPLICATIONS

The libraries associated with TRiCYCLE have been designed to be used programmatically in other applications. The flexibility of both the TridasJLib and DendroFileIOLib is illustrated by their successful incorporation into two quite different applications: the Corina dendrochronology desktop application and the DCCD web repository.

#### Corina

Corina is an open-source desktop application for dendro-measurement (including support for Velmex and Lintab platforms), analysis and data management (Brewer *et al.* 2010). It has been developed at the Malcolm and Carolyn Wiener Laboratory for Aegean and Near Eastern Dendrochronology at Cornell University. Data curation and management are possible because of the TRiDaS-enabled database server architecture that allows multiple users running the Corina client to access data simultaneously from a centralized lab repository. The implementation of the Dendro-FileIOLib means that users can upload legacy



**Figure 2.** A screen shot of TRiCYCLE showing the results of the conversion of three TRiDaS format data files into Sheffield D-Format files. Note that all three files were successfully converted, but the second two both include warnings related to the Sheffield format writer. The warning for the third file has been expanded and shows that the original data file contains earlywood density data that cannot be represented in a Sheffield format file. Also note that each input file has been converted into multiple output files because each TRiDaS file contains multiple data series and the Sheffield format requires that each file should contain just one series.

data files to the Corina server, and download existing data from the database in any one of the twenty-one supported formats.

## DCCD

The DCCD (Jansma 2010) is a web-based data infrastructure and repository of cultural dendrochronology based in the Netherlands. It contains all dendrochronological measurement series and descriptive and interpretive metadata now managed in laboratories in the Netherlands

(6025 BC–present), as well as selections of data from laboratories in Belgium, Germany, France and Poland. Scientists in Austria, Denmark, Ireland, Latvia, Lithuania, Poland, Slovenia, Spain and the UK recently selected the DCCD as their future vehicle for collaborative research (International meeting “Towards a European Research Infrastructure for Dendrochronology,” 14–15 December 2009, RCE, Amersfoort). The implementation of the DendroFileIOLib has been critical to the success of DCCD as it has provided users with the ability to upload data directly to the

repository from the legacy formats that researchers are familiar with and use on a day-to-day basis.

### Third-Party Applications

For developers interested in using the libraries in their own applications, there are source code packages available for download, and latest code developments are available from the open access Sourceforge repository (<http://tridas.sf.net>). The source code packages include API documentation, example code and license information.

The TRiCYCLE libraries are all written in Java, as are the Corina and DCCD applications that utilize them. Clearly, the libraries will be of most interest to Java programmers; however there are a number of techniques for providing language bindings to the libraries in other programming languages. If the dendro community desires to have access to these libraries in other languages then this could be the focus of the next stage of development.

### LIMITATIONS

Clearly, if the original format has limited data and/or metadata capabilities, then the corresponding output file will also contain accordingly limited information, even if the output format is capable of storing much more. TRiCYCLE provides a method for converting the available data from one format to another. It is therefore not directly suited to users hoping to augment the metadata of their existing collections.

For a large number of conversion types, the process will inevitably result in the loss of some information. For example, a round-trip conversion from a rich data format A to a simplistic data format B and back to format A again, will result in a file with less information than was initially provided. The extent to which data is lost is entirely dependent on what formats are used. Although TRiCYCLE provides detailed information regarding errors and assumptions made during the conversion process, it does not list the precise details of information lost. TRiCYCLE therefore does not remove the necessity for users to understand the limitations of the formats that they are using.

Perhaps TRiCYCLE's biggest limitation is its inability to understand *ad hoc* naming conventions and methods used within particular laboratories. Faced with the certain limitations of the data formats, it is typical for laboratories to resort to localized conventions, especially with regards file names and series codes, to keep track of data files. For instance a Tucson file may be named 'ABC-15-A.tuc' referring to the first core (A) of tree 15 from site ABC. When converted to a Heidelberg file, TRiCYCLE will be unable to extract the site, tree and core codes into the separate fields provided by the format as this naming convention is non-standard. The user will therefore have to manually edit the output file to make this information clear.

### FUTURE

The inclusion of the TRiCYCLE libraries in Corina and DCCD show how useful modular open-source technology can be, but there are many other applications that could also make use of TRiCYCLE. Perhaps the most obvious is the International Tree-Ring Data Bank—ITRDB (Grissino-Mayer and Fritts 1997). The TRiCYCLE libraries have the potential to be installed on the ITRDB server to enable users to download the data in any of the supported formats. The libraries could also provide a method for users to access the metadata available in the ITRDB in a more efficient and standardized way. At the moment, data in the ITRDB is stored as a large collection of Tucson files associated with a database containing simple metadata. Although technically the Tucson format can store the metadata available within the database directly within the files, the variable nature of the Tucson format means that this is often done in a non-standard way. With the TRiCYCLE libraries in place, it would be possible to ensure the standardized metadata within the ITRDB could be output consistently by using one of the more extensive formats. This would be very beneficial for users with software capable of utilizing such metadata-rich files.

In the longer term, TRiCYCLE offers the starting point for a substantial expansion of the capabilities of the ITRDB. By building upon the



Tree-Ring Data Standard and using TRiCYCLE to deliver the data, the ITRDB database could be extended to enable the inclusion of many more metadata fields. As TRiDaS has very few mandatory fields, this would not be a burden to data contributors as they could continue to provide the limited metadata already required by the existing data submission procedure. It would, however, give contributors the opportunity to provide much more detailed information if these data were available. An expanded ITRDB would provide many exciting new opportunities for large-scale ‘meta-analyses’ of tree-ring data that are currently not possible with the existing system.

Within individual laboratories perhaps the most useful development directions would be those that enable the integration of TRiCYCLE into existing workflows. For dendrochronologists who rely upon dedicated commercial dendrochronology software such as TSAP-Win and PAST4, this will require cooperation with the commercial developers, and we suggest subscribers contact these companies with their requests. For users of scripting languages and libraries such as Matlab, Python and R, TRiCYCLE could be integrated by the community providing users with direct access to data in many data formats. Perhaps more importantly though, this would be the next step in providing the user community with better access to the TRiDaS data model with all the benefits that this will bring. We would welcome the opportunity to work with others in the community to make this happen.

## CONCLUSIONS

The TRiCYCLE application provides users with a much-needed tool to assist with the open sharing of dendro-data and metadata. It also offers the opportunity to maintain a single package that can read older data formats that are in danger of becoming obsolete. It is hoped that its simple interface and multi-lingual packaging will make it accessible to the widest possible audience.

The co-launch of the stand-alone libraries that provide the core functionality of TRiCYCLE is hoped to be the first step in a new modular,

open-source and object-oriented approach to application development in the dendrochronology community. We believe that by sharing development resources the community will be better able to develop innovative tools, especially for the newer sub-disciplines that are reliant on rich metadata (for example dendrogeomorphology, cultural dendrochronology, dendropyrology and dendrochemistry).

Both TRiCYCLE and the associated libraries will continue to be developed and updated. We therefore welcome assistance from programmers who would like to contribute to their development, and especially those wishing to implement support for additional formats. Assistance is also warmly welcomed from non-programmers in the form of translation, testing, feature requests and user support.

## ACKNOWLEDGMENTS

Funding for the development of TRiCYCLE has been provided by The Netherlands Organization for Scientific Research (NWO) section Humanities through the DCCD project and through the various patrons of the Malcolm and Carolyn Wiener Laboratory for Aegean and Near Eastern Dendrochronology.

We would like to thank the numerous contributors to the open-source libraries used by TRiCYCLE and its associated libraries. We would also like to thank Roland Aniol, Rémi Brageu, Aoife Daly, Marta Domínguez Delmás, Pascale Fraiture, Henri Grissino-Mayer, Kristof Haneca, Patrick Hoffsummer, Bernhard Knibbe, George Lambert, Rowin van Lanen, Lars-Åke Larsson, Catherine Lavier, Hans-Hubert Leuschner, Martin Munro, Ian Tyers and Ronald Visser for helping us to understand aspects of the implemented data formats and for testing the conversion routines. Finally, we would like to thank two anonymous reviewers for their comments on an earlier version of this manuscript.

## REFERENCES CITED

- Brewer, P., K. Sturgeon, L. Madar, and S. W. Manning. 2010. A new approach to dendrochronological data management. *Dendrochronologia* 28:131–134.

- Bunn, A., 2008. A dendrochronology program library in R (dplR). *Dendrochronologia* 26:115–124.
- Cook, E., and R. Holmes, 1996. Guide for computer program ARSTAN. In *The International Tree-Ring Data Bank Program Library Version 2.0 User's Manual*, edited by R. Holmes, and H. Fritts, pp. 75–87. University of Arizona, Tucson.
- Grissino-Mayer, H., and H. Fritts, 1997. The international tree-ring data bank: An enhanced global database serving the global scientific community. *The Holocene* 7:235–238.
- Holmes, R., 1983. Computer-assisted quality control in tree-ring dating and measurement. *Tree-Ring Bulletin* 43:69–78.
- Holmes, R., 2001. *Dendrochronology Program Library (DPL)*. The University of Arizona, Tucson.
- Jansma, E., 2010. Preserving tree-ring data: a repository for the Low Countries. In *Driven by Data, Exploring the Research Horizon*, edited by M. de Groot, and M. Wittenberg, pp. 29–33. Pallas Publications, Amsterdam University Press.
- Jansma, E., P. Brewer, and I. Zandhuis, 2010. TRiDaS 1.1: The tree-ring data standard. *Dendrochronologia* 28:99–130.

*Received 2 December 2010; accepted 5 March 2011.*

**Supplementary Material** is available at <http://www.treeringsociety.org/TRBTRR/TRBTRR.htm>