Set my (pdf)pages free

David Walden

Experienced (La)TeX users know that they can do many things with these systems. However, new users, for instance only having learned enough to typeset a thesis, may not think of some of the other possibilities. Below I describe one (admittedly trivial) use of LaTeX for something other than typesetting a document.

With fair frequency I receive PDF files from which I wish to extract pages or images but cannot (my collaborators may not know their word processors are creating protected files). Maybe if I knew more about such security settings, I could undo the protection in other ways. However, I do know that the following tiny LaTeX program has always "set my PDF pages free" in the way I wanted. The file for the following program is named `select-pages.tex`.

```
\documentclass{article}
\usepackage{pdfpages}
\begin{document}
\includepdf[pages=1-8]% omit for all pages
  {name-of-file-to-be-freed.pdf}
\end{document}
```

I put a copy of the file in the directory with the PDF I want to set free, and then change the file name in the `\includepdf` command to the name of the file I want to unlock. I compile this little LaTeX program, and rename the result (which initially is `select-pages.pdf`) to be whatever I want it to be. Now I have a file which is no longer protected. (I don't know why this works, but it does.)

All the work is done by the `pdfpages` package (`ctan.org/pkg/pdfpages`). In the above example, pages 1 to 8 of the original document are processed into the output file. If the optional argument in square brackets is left out, the entire input document is processed into the output file. Other options for the `pages` parameter are available, and the `pdfpages` package has lots of other options; read about it at the above noted url.

Once the desired "free" pages are in the new file, I have found I can now extract pages and copy images which Acrobat and other applications on my Windows computer previously would not let me touch except to read.

This is one minuscule example of how (La)TeX can do miscellaneous things for you. *TUGboat* has published many articles on using (La)TeX as a more general purpose computing tool than typesetting alone, and no doubt would welcome more.

◇ David Walden
`walden-family.com/texland`

Automatic generation of herbarium labels from spreadsheet data using LaTeX

R. Sean Thackurdeen and Boris Veytsman

Abstract

LaTeX, being a programmable language, has advanced capabilities for automatic generation of documents. While these capabilities are often considered the realm of advanced users, they are also attractive for entry-level users. The latter can use them to learn about LaTeX while performing a typesetting task. The goal of this tutorial is to describe a method to typeset herbarium labels using data stored in a `.csv` file. This example is especially relevant for the botanical research community, where labels must be generated from standardized data sets to annotate physical plant collections.

1 Botanical primer

Botanical vouchers are the foundation of the study of the evolutionary history of plants, known as systematics, and the study of their classification, known as taxonomy. They are the ontological basis on which botanical theories and hypotheses of evolution are made. Additionally, the study of specii and their niches (ecology), and the study of their distributions across temporal and spatial scales (biogeography) are allied sciences which draw from these instances of recorded plant life.

Botanical vouchers are composed of two components: 1) a specimen, and 2) a label. The specimen commonly features fertile plant parts as well as other distinguishing characteristics, such as leaf arrangement, developmental variation, etc. When combined with DNA evidence, it is used for classification and identification of a plant. The label presents information grounding a specimen in physical space. It is the written manifestation of the specimen's identification, and includes collection information, geolocality data, and information about the habitat where the specimen was collected, in addition to other information about the specimen not apparent on the sheet.

Although not directly related to the present topic, readers may also be interested in the two articles by Joseph Hogg previously published in *TUGboat* (vol. 26, no. 1 and vol. 35, no. 2) on botanical typesetting: `http://tug.org/TUGboat/Contents/listauthor.html#Hogg,Joseph`.

2 Workflow

While botanists generally proceed by the adage, "by their fruits ye shall know them", it can be more apt to say that "botanists make labels".

```
Mj Gp,Scientific Name,Family,Genus,Specific Epithet,Taxon Rank,Infraspecific epithet,Scientific Name
     Authorship,,IdentifiedBy,Date Identified,Identification Remarks,,Identification Qualifier,,Event Date,
     Collector,Associated Collectors,collectorNumberPrefix,collectorNumber,collectorNumberSuffix,habitat,habit
     ,country,stateProvince,island,locality,localitySecurity,localitySecurityReason,geodeticDatum,
     decimalLatitude,decimalLongitude,elevation (m.),,duplicates,numberLabels,preparations, ,tripNumber,
     shippingPermit,shippingBox,shippingNote,,vernacularName1,languageName1,notesVernacularName1,
     vernacularName2,vernacularLanguage2,notesVernacularName2,plantUse1,plantUseCategory1,plantUse2,
     plantUseCategory2,informationWitheld,sourceNames,interviewers,interviewDate,,enteredBy
Angiosperm,Stachytarpheta jamaicensis (L.) Vahl,Verbenaceae,Stachytarpheta,jamaicensis,sp.,,(L.) Vahl,,,,,,,,"
     June 7, 2014",Gregory M. Plunkett,"Michael Balick, Kate Armstrong, Sean Thackurdeen, Jean-Pascal Wahe,
     Presley Dovo & Joshua Andrew",,2783,,Growing in open area along roadside of disturbed secondary forest.,"
     Herb to subshrub, 0.5 m tall, flowers purple.",Vanuatu,Tafea ,Tanna,"West Tanna, just east of Lenakel,
     along track to Letakran Village, along creek.",,,WGS84,-19.52803,169.2813,44,,6,6,"DNA, digital image
     ",,,,,,,,,,,,,,,,,,,,,,
Angiosperm,Ophioglossum reticulatum var. reticulatum L.,Ophioglossaceae,Ophioglossum,reticulatum,var.,
     reticulatum,L.,,Gregory M. Plunkett,6/25/2014,,,!,,"June 25, 2014",Gregory M. Plunkett,"Tom Ranker,
     Chanel Sam, Jean-Pascal Wahe, Sean Thackurdeen, Kate Armstrong, Laurence Ramon, Frazer Alo, Alexis Tupun,
      David Kapwia & Joseph Dabauh.",,2910,,,Terrestrial fern growing in dense forest.,Vanuatu,Tafea ,Tanna,"
     Southwest Tanna, along trail from Yenhup to Mount Tukosmera.",,,WGS84,-19.588028,169.366611,559,,6,6,"DNA
     , digital image",,,,,,,,,,,,,,,,,,,,,,
```

Figure 1: A three-line botanical `.csv` file (indented line breaks are editorial).

In the field, notes of a plant collection are made on weather resistant paper. At a moment's rest in the field, or back at home, data is provisionally transferred to a spreadsheet. The columns of the spreadsheet usually correspond to the database schema of the archival repository. Most often the schema adheres to the biodiversity standard known as Darwin Core (DwC). From this spreadsheet labels can be made using Microsoft Word's mail merge capabilities. Alternately, data can be uploaded to an intermediary (e.g., Filemaker) or to an archival repository that features reporting capabilities.

Producing labels directly from a field sheet allows a greater flexibility, since the labels can be generated anywhere a user has access to a computer with the requisite software. Unfortunately, the common work flow described above is unreliable. Once a spreadsheet is merged in Microsoft Word, any additional edits produce cascading effects which drastically alter the formatting of the document. The changes require an unnecessary amount of time and tedious effort. This issue can be alleviated through a reporting template used in intermediary repositories, but these systems are less flexible. A portable, field-ready and reliable solution is required to help botanists to avoid loss of time and to help them to make labels. Additionally, a system that is free and open source may be important for the botanists and collections managers in countries where herbaria lack extensive resources.

3 Tutorial

The tutorial which follows is a sequential, step by step, explanation of the LaTeX code which structures the document. As the tutorial proceeds, lines of code are grouped according to similarity of function. They are presented as blocks. While snippets of code are explained in relation to their function in the given example, possible alternatives are rarely explained. More detailed explanations of the options are better found LaTeX tutorials, of which there are many. We use the `.csv` file shown in Figure 1 for our examples.

First, a standard article class is called specifying the font size. The `geometry` package is used to specify the margins, and the `graphicx` and `datatool` packages are called to import images and spreadsheet values, respectively. The `datatool` package, authored by Nicola Talbot, is the key to this tutorial. It allows us to manipulate and typeset data stored in `.csv` files using LaTeX commands. Other approaches to automated document generation often rely on multiple programming languages to generate LaTeX code.

```
\documentclass[12pt]{article}
\usepackage{datatool,graphicx}
\usepackage[right=.2in, left=.2in,
  top=.2in,bottom=.2in,
  columnsep=.5in]{geometry}
```

The next portion of the preamble is a function designed to convert latitudes and longitudes in decimal degrees to degrees-minutes-seconds representation, for example, -19.588028 latitude to S $19^\circ 31' 40''$.

The Darwin Core data standard for biodiversity data specifies decimal-degrees as the accepted standard for geographic data. However, this format is less reader friendly and thus less aesthetically pleasing. The following function transforms and typesets the GPS data so as to satisfy readers and data handlers alike, using datatool (`\DTL...`) commands.

```
% #1- negative suffix,
% #2 - positive suffix,
% #3 - lat/lon
\newcommand{\latlontodeg}[3]{%
 \DTLifnumlt{#3}{0}{#1}{#2}~%
 \DTLabs{\TMPlatlon}{#3}%
 \DTLtrunc{\TMPdeg}{\TMPlatlon}{0}%
 \DTLsub{\TMPlatlon}{\TMPlatlon}{\TMPdeg}%
 \DTLmul{\TMPlatlon}{\TMPlatlon}{60}%
 \DTLtrunc{\TMPmin}{\TMPlatlon}{0}%
 \DTLsub{\TMPlatlon}{\TMPlatlon}{\TMPmin}%
 \DTLmul{\TMPlatlon}{\TMPlatlon}{60}%
 \DTLtrunc{\TMPsec}{\TMPlatlon}{0}%
 $\TMPdeg^\circ\TMPmin'\TMPsec''$}
```

The next code block begins the document environment. Immediately, a datatool command is used to load the spreadsheet data, at which point the working database is named and the file, residing in the same directory, is specified.

```
\begin{document}
\DTLloadrawdb{labels}{labelExample.csv}
```

Herbarium labels are customarily 4″ in width, and approximately 4″ in length, varying with the amount of data recorded. One often prints 4 labels to a US letter page. To typeset multiple labels on a commonly available US letter sheet, we switch to two column layout. Two columns of a portrait US letter, with appropriate margins, produces the desired 4 inches width of herbarium labels.

```
\twocolumn
```

There are two parts to the datatool formula which will generate the labels: assignments and commands. The first will designate identifiers for each row in the spreadsheet. Once the database is defined, a working name is assigned to the `.csv` column name that is to be typeset. Below is an abbreviated version of the code. Note the spaces and capitalization on the right side of the equation, which refer to column names in your `.csv` file.

```
\DTLforeach{labels}{%
  \Family=Family,
  \Genus=Genus,
  \Specie=Specific Epithet,
  \Authorship=Scientific Name Authorship}
```

Thus far we have defined the document size, its margins, created a multi-column environment, called a function and set assignments. Now, we begin the task of organizing the static and dynamic elements of the label.

While the two column typesetting allows text to flow from the base of one column to the beginning of the next one, we do not want an individual label to be continued on the next column or page. The text of a given label must be manipulated as a block. To create this environment we put the label inside a `minipage`:

```
{\noindent
 \begin{minipage}{1.0\linewidth}%
 \raggedright
  \setlength{\parskip}{.5\baselineskip}%
  \raisebox{-.5\height}
```

Next is a set of instructions to typeset a header. Many herbarium labels simply include a title which indicates flora of which the specimen is a part. In this case, an additional header with logos and herbarium codes is used. This can easily be customized as needed.

```
{\includegraphics[height=.8cm]{nybgLogo}}%
 \hfill
 \parbox[t]{5cm}{\centering\scshape\tiny
   New York Botanical Garden: NY\\
   Vanuatu National Herbarium: PVNH}%
 \hfill
 \raisebox{-.5\height}{%
 \includegraphics[height=1cm]{pvnhLogo}}%
 \par
 {\centering \bfseries\itshape\large
      The Flora of Vanuatu\par}%
```

Now we typeset the previously assigned elements. This is the heart of the approach. Here is an abbreviated example:

```
\DTLifnullorempty{\Family}{}{%
  \hfill(\Family)}
\DTLifnullorempty{\Genus}{}{%
  \textit{\bfseries\Genus}}
\DTLifnullorempty{\Specie}{}{%
  \textit{\bfseries\Specie}}
\DTLifnullorempty{\Authorship}{}{%
  \Authorship}
```

We use `\DTLifnullorempty` to typeset the data. The program reads the assignment in the first set of braces. If the column does not have data, we skip it, hence the second set of blank braces. If it does have some data, we typeset it as instructed in the third set of braces with the additional formatting instructions.

The benefit of using the `\DTLifnullorempty` command is that static elements of the text can be nested in the third set of braces. This text will then be typeset, but only if the column is present. Thus we do not include ugly placeholders for missing information, as in some other approaches.

The last part of the label is an acknowledgment. This is currently set to be included on every label. Alternatively, it could be easily incorporated into the `\DTLifnullorempty` command and therefore be included only on specimen labels so designated. After the acknowledgment there is a command ending the label block, and a command to include space between each label. Then we close the document:

```
 \centering\itshape\small
 A collaboration of NYBG and PVNH,
 funded by The~Christensen~Fund,
 The~National~Geographic~Society,
 and
 The~Critical~Ecosystem~Partnership~Fund.

\end{minipage}%
  \vspace{1cm}\par}
\end{document}
```

A page of example labels is given in Figure 2.

4 Notes on implementation

Many of the users adapting this code for use will stem from the natural history research community. Thus, it is worth mentioning a caveat about compiling documents from code. Computer programming languages, such as LaTeX, are exact and precise. If there are invalid characters in your data set, and you are unaware, you are sure to find out when you receive an error message (likely inscrutable) upon compiling. Similarly, if there are incorrect characters in your column mapping, an error message will result. While there are numerous possible errors, here are few tips to help you along.

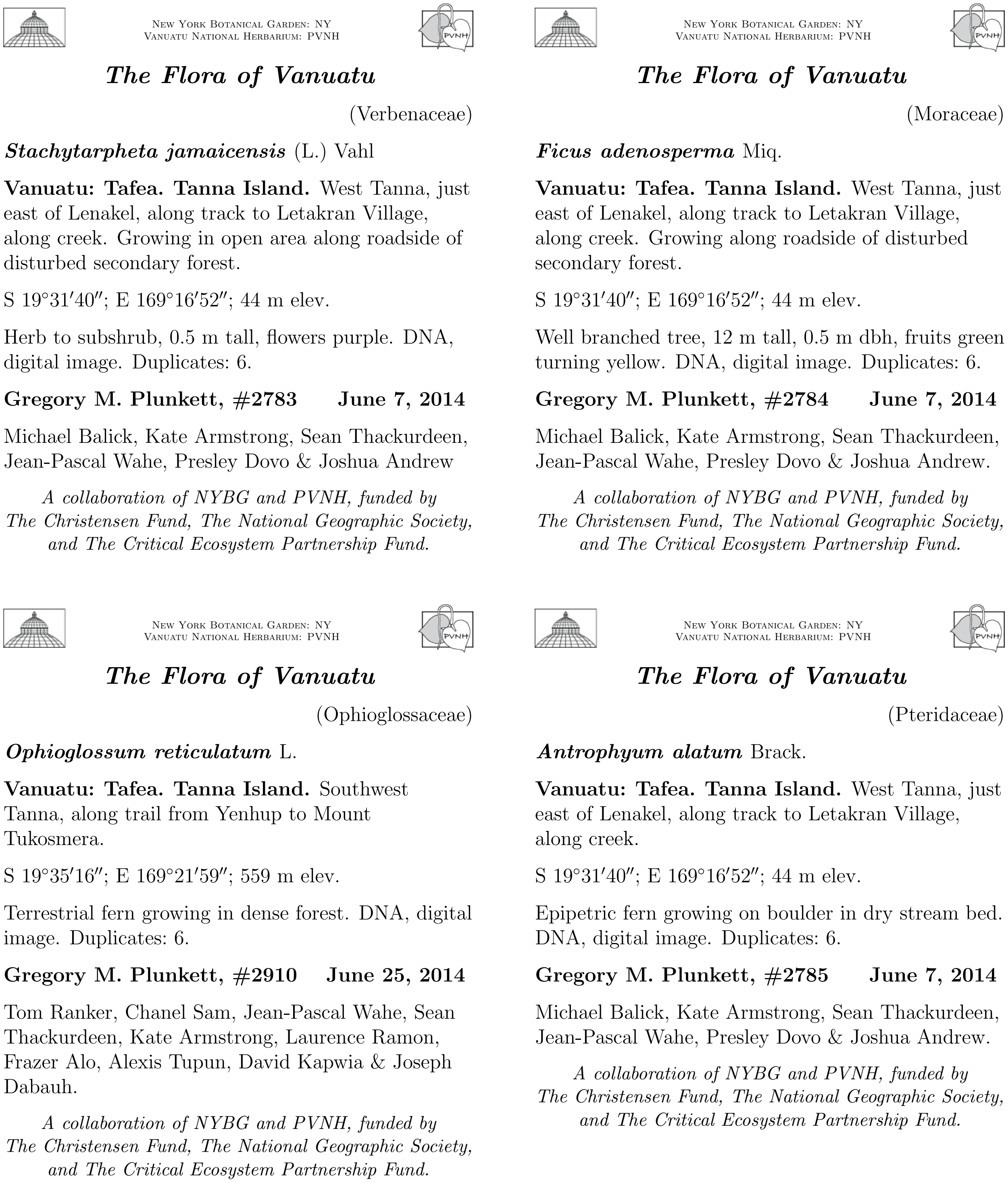
- Compile in batches:
  - 100 rows returns a quick and sufficiently large output, while creating a smaller dataset to troubleshoot.
  - If your data set is greater than 100 rows, use the `split` command line tool to generate parts. The resulting files have no headers and these headers can be specified as an option to the `\DTLloaddb` command.
- Consider both `\DTLloaddb` & `\DTLloadrawdb`:
  - `\DTLloaddb` requires LaTeX special characters to be treated as such in your `.csv` file. A positive benefit to this is that you can format specific text within cells using LaTeX syntax (e.g., taxon names in habitat descriptions).
  - `\DTLloadrawdb` is needed when LaTeX special characters are present in the file. This command will automatically convert those special characters to the required LaTeX format.
- Clean code & clean data:
  Issues compiling are probably due to either a syntax error in the code or invalid characters in your data set.
  - Be mindful of stray spaces, generally, and capitalization when defining the column mapping.
  - Be mindful of the input encoding and the text encoding.

5 Compile your own

A package including an example `.csv` file, LaTeX template, and output PDF will be posted to CTAN. This template will additionally be published to the online LaTeX compilers *Overleaf* and *ShareLaTeX*. Should you have questions regarding the template, feel free to reach out to the authors. Contact information is provided below.

◇ R. Sean Thackurdeen
Institute of Economic Botany
New York Botanical Garden
Bronx, NY 10458 USA
sthackurdeen (at) nybg dot org
http://thackur.org/

◇ Boris Veytsman
Systems Biology School and
Computational Materials
Science Center
MS 6A2
George Mason University
Fairfax, VA 22030 USA
borisv (at) lk dot net
http://borisv.lk.net/



NEW YORK BOTANICAL GARDEN: NY
VANUATU NATIONAL HERBARIUM: PVNH

The Flora of Vanuatu

(Verbenaceae)

Stachytarpheta jamaicensis (L.) Vahl

Vanuatu: Tafea. Tanna Island. West Tanna, just east of Lenakel, along track to Letakran Village, along creek. Growing in open area along roadside of disturbed secondary forest.

S 19°31′40″; E 169°16′52″; 44 m elev.

Herb to subshrub, 0.5 m tall, flowers purple. DNA, digital image. Duplicates: 6.

Gregory M. Plunkett, #2783 June 7, 2014

Michael Balick, Kate Armstrong, Sean Thackurdeen, Jean-Pascal Wahe, Presley Dovo & Joshua Andrew

A collaboration of NYBG and PVNH, funded by The Christensen Fund, The National Geographic Society, and The Critical Ecosystem Partnership Fund.

NEW YORK BOTANICAL GARDEN: NY
VANUATU NATIONAL HERBARIUM: PVNH

The Flora of Vanuatu

(Moraceae)

Ficus adenosperma Miq.

Vanuatu: Tafea. Tanna Island. West Tanna, just east of Lenakel, along track to Letakran Village, along creek. Growing along roadside of disturbed secondary forest.

S 19°31′40″; E 169°16′52″; 44 m elev.

Well branched tree, 12 m tall, 0.5 m dbh, fruits green turning yellow. DNA, digital image. Duplicates: 6.

Gregory M. Plunkett, #2784 June 7, 2014

Michael Balick, Kate Armstrong, Sean Thackurdeen, Jean-Pascal Wahe, Presley Dovo & Joshua Andrew.

A collaboration of NYBG and PVNH, funded by The Christensen Fund, The National Geographic Society, and The Critical Ecosystem Partnership Fund.

NEW YORK BOTANICAL GARDEN: NY
VANUATU NATIONAL HERBARIUM: PVNH

The Flora of Vanuatu

(Ophioglossaceae)

Ophioglossum reticulatum L.

Vanuatu: Tafea. Tanna Island. Southwest Tanna, along trail from Yenhup to Mount Tukosmera.

S 19°35′16″; E 169°21′59″; 559 m elev.

Terrestrial fern growing in dense forest. DNA, digital image. Duplicates: 6.

Gregory M. Plunkett, #2910 June 25, 2014

Tom Ranker, Chanel Sam, Jean-Pascal Wahe, Sean Thackurdeen, Kate Armstrong, Laurence Ramon, Frazer Alo, Alexis Tupun, David Kapwia & Joseph Dabauh.

A collaboration of NYBG and PVNH, funded by The Christensen Fund, The National Geographic Society, and The Critical Ecosystem Partnership Fund.

NEW YORK BOTANICAL GARDEN: NY
VANUATU NATIONAL HERBARIUM: PVNH

The Flora of Vanuatu

(Pteridaceae)

Antrophyum alatum Brack.

Vanuatu: Tafea. Tanna Island. West Tanna, just east of Lenakel, along track to Letakran Village, along creek.

S 19°31′40″; E 169°16′52″; 44 m elev.

Epipetric fern growing on boulder in dry stream bed. DNA, digital image. Duplicates: 6.

Gregory M. Plunkett, #2785 June 7, 2014

Michael Balick, Kate Armstrong, Sean Thackurdeen, Jean-Pascal Wahe, Presley Dovo & Joshua Andrew.

A collaboration of NYBG and PVNH, funded by The Christensen Fund, The National Geographic Society, and The Critical Ecosystem Partnership Fund.

Figure 2: A page of example labels