Automated report generation using R and Python docutils

Abhijit Dasgupta
Consultant, NIH
Consultant, Tranzxn Inc
statbandit.wordpress.com
@webbedfeet
DC useR group co-leader (with Marck Vaisman)

July 23, 2010

Motivation

- Used automated report generation using LaTeX /Sweave $\rightarrow$ PDF
- Used descriptive tables from `Hmisc` very regularly

Motivation

- Used automated report generation using LaTeX /Sweave $\rightarrow$ PDF
- Used descriptive tables from `Hmisc` very regularly
- Many clinical collaborators work in MS Word 
- Translating LaTeX to Word *well* is a bear

Motivation

- Used automated report generation using LaTeX /Sweave $\rightarrow$ PDF
- Used descriptive tables from `Hmisc` very regularly
- Many clinical collaborators work in MS Word 
- Translating LaTeX to Word *well* is a bear
  - ... but see http://biostat.mc.vanderbilt.edu/wiki/Main/SweaveConvert for some great ideas
  - PDF2Word actually is fantastic

Motivation

- Used automated report generation using LaTeX /Sweave $\rightarrow$ PDF
- Used descriptive tables from `Hmisc` very regularly
- Many clinical collaborators work in MS Word 
- Translating LaTeX to Word *well* is a bear
  - ... but see http://biostat.mc.vanderbilt.edu/wiki/Main/SweaveConvert for some great ideas
  - PDF2Word actually is fantastic
- Looking for **direct** translation of "weaved" document to Word (or rather OOo)

ReStructured Text

- Formatted ASCII text file
  - Headlines by underlining
  - Includes lists, basic formatting, tables, figure inclusion
  - Fairly intuitive
- Basic source format for all Python documentation
- Mature translators to LaTeX, HTML, DocBook, PDF and ODT in Python `docutils` package and addons
- Same source file can be translated into multiple formats efficiently

Initial thoughts

- Main result types for weaving are tables and figures
  - Tables can be passed through print.char.matrix (`Hmisc`) into exactly the right format for reSt
  - Some tweaks would be necessary to make `summary.formula.reverse` nice.
  - Leverage utlity functions to format model output (personal, `arm`, others) through print.char.matrix to output reSt
  - Figures are no problem (inserts link).

Initial thoughts

- Main result types for weaving are tables and figures
  - Tables can be passed through print.char.matrix (`Hmisc`) into exactly the right format for reSt
  - Some tweaks would be necessary to make `summary.formula.reverse` nice.
  - Leverage utlity functions to format model output (personal, `arm`, others) through print.char.matrix to output reSt
  - Figures are no problem (inserts link).
- Weaving
  - Didn't want to create *weave function for this (lazy)
  - `brew` (J. Horner, Vanderbilt) provides a easy method for weaving

Initial thoughts

- Main result types for weaving are tables and figures
  - Tables can be passed through print.char.matrix (`Hmisc`) into exactly the right format for reSt
  - Some tweaks would be necessary to make `summary.formula.reverse` nice.
  - Leverage utlity functions to format model output (personal, `arm`, others) through print.char.matrix to output reSt
  - Figures are no problem (inserts link).
- Weaving
  - Didn't want to create *weave function for this (lazy)
  - `brew` (J. Horner, Vanderbilt) provides a easy method for weaving
- Onward!!!
  - Abstract to useR2010
  - Development

6 weeks later

- 2nd DC useR meeting
  - Out for beeR & liquoR at goRdon bieRsch
  - Hear about **ascii** library for first time
  - **ascii** committed to CRAN about 9 months before my idea germinated (David Hajage)

6 weeks later

- 2nd DC useR meeting
  - Out for beeR & liquoR at goRdon bieRsch
  - Hear about **ascii** library for first time
  - **ascii** committed to CRAN about 9 months before my idea germinated (David Hajage)
- **Curses, Foiled Again!!!**

6 weeks later

- 2nd DC useR meeting
  - Out for beeR & liquoR at goRdon bieRsch
  - Hear about **ascii** library for first time
  - **ascii** committed to CRAN about 9 months before my idea germinated (David Hajage)
- **Curses, Foiled Again!!!**

  - **ascii** provides about 85% of what I wanted.
  - Provides an *RweaveReST* function
  - Some custom tweaks needed (`summary.formula.reverse` again) for my optimal flavor.

Moral of this anecdote

- Current descriptives and cataloging of CRAN makes useful packages sometimes hard to find $\Rightarrow$ wasted effort
- If you have an idea, there is a good chance that someone has contributed some code towards that idea $\Rightarrow$ potentially saved effort
- Some loss function at play for developing new ideas.
- Maybe StackOverflow and similar sites can help minimize the loss.

ReStructured Text

- Easy-to-read, WYSIWYG plaintext markup syntax and parser system
- Source syntax for Python documentation
- Mature translators in Python Docutils (docutils.sourceforge.net)
- Good introductory tutorial at http://docutils.sourceforge.net/docs/user/rst/quickstart.html
- Can format tables, link to graphics, and have intuitive text formatting methods.

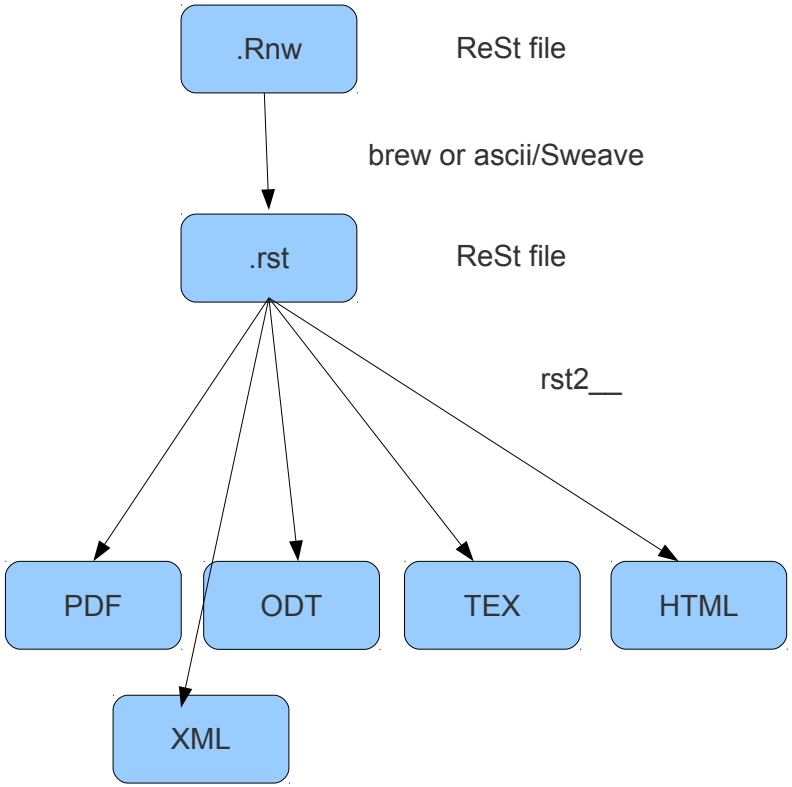
ReStructured Text

- Translators available for LaTeX, HTML, Docbook, XML, OpenOffice.org ODT, PDF (without LaTeX) and other addons.
- Can provide custom stylesheets or CSS to customize appearance.
- One source to rule them all!!!

Table: Docutils Translators

Translator	Output
rst2html	HTML
rst2pdf	PDF (through Python Reportlab)
rst2latex	LaTeX
rst2newlatex	
rst2xml	XML
rst2odt	OpenOffice ODT

Workflow


.Rnw
ReSt file
brew or ascii/Sweave
.rst
ReSt file
rst2__
PDF
ODT
TEX
HTML
XML

Example

Conclusions

- Restructured text is easy to learn and use
- Substantial flexibility in output formats from a common source
- Some limitations in formatting details
- Certainly less powerful than LaTeX, but still quite effective.
- Using `brew` or `ascii` allows weaving code and text.
- Defaults give decent formats, but additional stylesheets can be added.
- Some custom tweaks to make formatting optimal on a personal level (E-mail me if you would like to see them)

statbandit.wordpress.com
Twitter: @webbedfeet
DC useR group (August meeting coming up)