

Pandoc: Ένας ελβετικός σουγιάς μετατροπής αρχείων

Απόστολος Συρόπουλος

28ης Οκτωβρίου 366

671 00 Ξάνθη

H/T: *asyropoulos at yahoo dot com*

Η δυνατότητα μετατροπής αρχείων επισήμανσης από μία μορφή σε κάποια άλλη είναι πολύ σημαντική. Το pandoc είναι μια εφαρμογή γραμμένη στη συναρτησιακή γλώσσα προγραμματισμού Haskell η οποία κάνει πολύ εύκολη την μετατροπή αρχείων επισήμανσης. Επιπλέον, επιτρέπει και τη διαχείριση βιβλιογραφικών δεδομένων.

Pandoc: A Swiss Army knife for file conversion, by Apostolos Syropoulos —The ability to convert markup files from one format to another is very important. Pandoc is an application written in the functional programming language Haskell which makes very easy the conversion of markup files. Moreover, it allows for the management of bibliographic data.

1 Γενικά για το pandoc

Το πρόγραμμα pandoc¹, που σχεδιάστηκε και υλοποιήθηκε από τον John MacFarlane, είναι μια εφαρμογή γραμμένη στη συναρτησιακή γλώσσα προγραμματισμού Haskell². Το pandoc είναι ένα εργαλείο που κάνει πολύ εύκολη τη μετατροπή αρχείων επισήμανσης από μία (υποστηριζόμενη) μορφή σε κάποια άλλη (υποστηριζόμενη μορφή). Επιπλέον επιτρέπει την επεξεργασία βιβλιογραφικών βάσεων κάνοντας έτσι το έργο της μετατροπής σχεδόν τετρημένη υπόθεση. Προφανώς, η ιδιαιτερότητα του προγράμματος βρίσκεται στο γεγονός πως μπορεί να διαχειριστεί πολλές μορφές αρχείων ενώ συνήθως παρόμοιες εφαρμογές μπορούν να κάνουν μετατροπή από μία μορφή σε μία άλλη μορφή. Πιο συγκεκριμένα στον Πίνακα 1 φαίνονται τα είδη αρχείων που αναγνωρίζει και που μπορεί να δημιουργήσει το πρόγραμμα pandoc.

Όπως αναφέρθηκε, το πρόγραμμα pandoc είναι γραμμένο σε Haskell το οποίο σημαίνει πως αν η υπολογιστική πλατφόρμα που χρησιμοποιείτε δεν είναι κάποια έκδοση των Windows ή το MacOS, τότε θα πρέπει να δημιουργήσετε εσείς το πρόγραμμα χρησιμοποιώντας την εντολή

¹<http://johnmacfarlane.net/pandoc/>

²<http://www.haskell.org/>

Αρχεία εισόδου	Αρχεία εξόδου
docbook	asciidoc
haddock	beamer, texinfo και latex
html	context
json	docbook
latex	docx και rtf
markdown	dzslides
markdown_github	epub και epub3
markdown_mmd	fb2
markdown_phpextra	html και html5
markdown_strict	json
mediawiki	man
native	markdown
opml	markdown_github
rst	markdown_mmd
textile	markdown_phpextra
	markdown_strict
	mediawiki
	native
	odt
	opendocument
	opml
	org
	pdf
	plain
	revealjs
	rst
	s5
	slideous
	slidy
	textile

Πίνακας 1: Είδη αρχείων που αναγνωρίζει και που μπορεί να δημιουργήσει το pandoc.

```
$ cabal install pandoc
```

Αυτό βέβαια προϋποθέτει πως έχετε εγκαταστήσει μια πολύ πρόσφατη έκδοση του Glasgow Haskell Compilation System (ghc³). Για τις περισσότερες πλατφόρμες το ghc είναι διαθέσιμο ως πακέτο συστήματος. Για κάποιες δεν υπάρχει ούτε ως πακέτο, οπότε θα πρέπει να κάνετε μόνοι σας την εγκατάσταση.

Θα πρέπει να τονιστεί πως το pandoc είναι ένα πρόγραμμα που τρέχει από την γραμμή εντολών (για τους χρήστες των Windows) ή από ένα τερματικό (για τους χρήστες ενός λειτουργικού συστήματος Unix ή τύπου Unix). Επειδή δεν υπάρχει κάποιο γραφικό περιβάλλον εργασίας, θα μπορούσε κάλλιστα αυτό να δημιουργη-

³<http://www.haskell.org/ghc/>

Μορφή	Επέκταση ονόματος αρχείου
MODS	.mods
Bib $\LaTeX$	.bib
Bib $\TeX$	.bibtex
RIS	.ris
EndNote	.enl
EndNote XML	.xml
ISI	.wos
MEDLINE	.medline
Copac	.copac
JSON citeproc	.json

Πίνακας 2: Είδη αρχείων που αναγνωρίζει και που μπορεί να δημιουργήσει το pandoc-citeproc.

Θεί είτε από κάποιο φοιτητή που ψάχνει να κάνει μια ενδιαφέρουσα διπλωματική εργασία, είτε από κάποιον που δεν ξέρει πώς να σκοτώσει τον ελεύθερο χρόνο του.

2 Βασική χρήση του pandoc

Πριν από λίγο καιρό μαζί με έναν φίλο είχαμε γράψει ένα άρθρο χρησιμοποιώντας φυσικά το $\LaTeX$. Θελήσαμε να δημοσιεύσουμε αυτό το άρθρο και τελικά βρήκαμε ένα περιοδικό το οποίο ενδιαφέρεται για άρθρα με περιεχόμενο όπως του δικού μας. Το κακό ήταν πως το περιοδικό ζητούσε να υποβάλλουμε το άρθρο σε μορφή docx. Μολονότι μίλησα με τους υπεύθυνους του περιοδικού και τους είπα πως αν γίνει κρίση και αυτή είναι θετική, θα μετέτρεπα το άρθρο άμεσα στη ζητούμενη μορφή, αυτοί επέμεναν πως έπρεπε να υποβάλω το άρθρο σε μορφή .docx! Μάλιστα μου προτείναν να χρησιμοποιήσω ένα εργαλείο που μετατρέπει αρχεία $\LaTeX$ σε μορφή .rtf, ώστε στη συνέχεια να εισαγάγω το παραγόμενο αρχείο στο Word 2010. (Για κάποιο λόγο, η έκδοση 2007 δεν ήταν αρκετή!) Εκείνες τις ημέρες είχα εγκαταστήσει το pandoc, το οποίο είχα σκοπό να χρησιμοποιήσω για κάποια άλλη δουλειά, και έτσι σκέφτηκα να χρησιμοποιήσω αυτό το πρόγραμμα.

Αρχικά χρησιμοποίησα την παρακάτω εντολή:

```
$ pandoc -s teach_script.tex -o teach_script.docx
```

Το αποτέλεσμα ήταν καλό αλλά η βιβλιογραφία δεν υπήρχε στο τέλος του παραγόμενου αρχείου! Μετά από λίγο ψάξιμο, ανακάλυψα πως υπήρχε ένα δεύτερο εργαλείο που ονομάζεται pandoc-citeproc. Αυτό το εργαλείο μπορεί να διαχειριστεί βιβλιογραφικές βάσεις σε διάφορες μορφές (βλ. Πίνακας 2).

Με αυτό το δεδομένο, διαπίστωσα πως η παρακάτω εντολή δημιουργούσε τη βιβλιογραφία στο τέλος του παραγόμενου αρχείου:

```
$ pandoc -s teach_script.tex --bibliography=teach-script.bib \
-o teach_script.docx
```

Όμως η βιβλιογραφία εμφανιζόταν σε μία μορφή που δεν ήταν ακριβώς η μορφή που ζητούσε το περιοδικό... Η μόνη λύση έμοιαζε να είναι η διόρθωση του αρχείου

με το μάτι αλλά, ευτυχώς, μετά από λίγο ψάξιμο, διαπίστωσα πως είναι δυνατόν να καθορίσουμε και τον τρόπο με τον οποίο θα εμφανίζεται η βιβλιογραφία. Απλά θα πρέπει να σημειώσουμε το όνομα ενός αρχείου στυλ⁴ με τον τρόπο που φαίνεται παρακάτω:

```
$ pandoc -s teach_script.tex --bibliography=teach-script.bib \
--csl=acm-sigchi-proceedings.csl -o teach_script.docx
```

Αυτή τη φορά η βιβλιογραφία βγήκε σωστά αλλά οι αναφορές δεν ήταν εντελώς σωστές. Αυτές, δυστυχώς, έπρεπε να τις διορθώσω με το χέρι. Ένα άλλο πρόβλημα που αντιμετώπισα ήταν οι πίνακες: Το pandoc απλά πετάει στο τελικό αρχείο τον κώδικα $\LaTeX$ του πίνακα. Το pandoc δεν κατανοεί όλες τις εντολές του $\LaTeX$, αλλά μπορεί και επεξεργάζεται μια χαρά ένα βασικό υποσύνολο των εντολών του $\LaTeX$.

Παραπάνω αναφέρθηκε πως το pandoc μπορεί να παράγει αρχεία html αλλά και αρχεία html5. Το ερώτημα είναι: Πώς μπορούμε να δηλώσουμε τι μορφή αρχείου θέλουμε να παραχθεί; Απλά χρησιμοποιούμε τον *διακόπτη* γραμμής εντολών `-t` και δίπλα το είδος του αρχείου που θέλουμε να παραχθεί. Για παράδειγμα, να πώς παράγουμε αρχείο html5:

```
$ pandoc -s -t html5 teach_script.tex \
--bibliography=teach-script.bib \
--csl=acm-sigchi-proceedings.csl -o teach_script.html
```

3 Με τα μαθηματικά, τι γίνεται;

Είναι γεγονός πως η εμφάνιση μαθηματικού κειμένου είναι μια ιδιαίτερα δύσκολη υπόθεση. Φυσικά το $\LaTeX$ και το $\XeLaTeX$ παράγει μαθηματικό κείμενο αξεπέραστης ποιότητας, αλλά σε πολλές περιπτώσεις αυτό δεν είναι αρκετό. Για παράδειγμα, αν θέλουμε να φτιάξουμε ένα e-book σε μορφή ePUB, τότε θα πρέπει το μαθηματικό κείμενο να σημειωθεί σε MathML ή, στην χειρότερη περίπτωση, να εισαχθεί στη μορφή εικόνας. Το pandoc επιτρέπει την εύκολη μετατροπή μαθηματικού κειμένου που έχει σημειωθεί σε μορφή $\LaTeX$ σε μορφή MathML. Φυσικά και υπάρχουν περιορισμοί καθώς δεν αναγνωρίζονται όλα τα περιβάλλοντα μαθηματικού κειμένου, αλλά αυτό είναι αναμενόμενο καθώς κάποια, όπως το `multline`, δεν έχουν το αντίστοιχο τους στην HTML. Ας δούμε ένα απλό παράδειγμα.

Ο παρακάτω κώδικας περιέχει δύο εξισώσεις και κείμενο στα ελληνικά. Υποθέτουμε πως το αρχείο `test.tex` στο οποίο θα αποθηκευτεί ο κώδικας αυτός θα είναι κωδικοποιημένο σε μορφή UTF-8.

```
\documentclass[a4paper]{article}
\usepackage{amsmath}
\begin{document}
Αυτή είναι μια απλή εξίσωση σε «εμφάνιση αράδας»
 $E=mc^2$  και μία σε εμφάνιση «βιτρίνας»:
```

⁴Μπορείτε να διαλέξετε ένα από το δικτυακό τόπο <http://zotero.org/styles> και να το κατεβάσετε για να το χρησιμοποιήσετε με το pandoc.

Pandoc: Ένας ελβετικός σουγιάς μετατροπής αρχείων

11

```
\begin{displaymath}
Q=\iiint_{\Omega}\rho\,dV
\end{displaymath}
\end{document}
```

Για να δημιουργήσουμε ένα αρχείο html5 από αυτόν τον κώδικα, θα πρέπει να χρησιμοποιήσουμε μια εντολή όπως αυτή που ακολουθεί:

```
$ pandoc -s -t html5 test.tex --mathml -o test.html
```

Το αποτέλεσμα της μετατροπής είναι το ακόλουθο:

Αυτή είναι μια απλή εξίσωση σε «εμφάνιση αράδας» $E = mc^2$ και μία σε εμφάνιση «βιτρίνας»:

$$Q = \iiint_{\Omega} \rho dV$$

Σημειώνω πως χρησιμοποιήθηκε ο Firefox 24.0 για την προβολή του παραγόμενου αρχείου html5. Δυστυχώς οι άλλοι κοινοί φυλλομετρητές έχουν περιορισμένες δυνατότητες προβολής κώδικα MathML. Αυτό με τη σειρά του μπορεί να δημιουργήσει προβλήματα στη σωστή εμφάνιση μιας σελίδας html. Μια απλή λύση είναι να χρησιμοποιήσουμε τις δυνατότητες που παρέχει το MathJax, το οποίο είναι ένα πρόγραμμα ανοικτού κώδικα σε JavaScript που στοχεύει στην ποιοτική εμφάνιση μαθηματικού περιεχομένου σε όλους τους φυλλομετρητές. Αν λοιπόν σκοπεύουμε να παράγουμε αρχεία html με μαθηματικό κείμενο που να προβάλλεται μέσω του MathJax, θα πρέπει να χρησιμοποιήσετε μια εντολή όπως η ακόλουθη:

```
$ pandoc -s -t html5 test.tex --mathjax -o test.html
```

Το παραγόμενο αρχείο html θα περιέχει μερικές αράδες όπως περίπου οι παρακάτω:

```
<script src="http://cdn.mathjax.org/mathjax/latest/MathJax.js?
config=TeX-AMS-MML_HTMLorMML" type="text/javascript">
MathJax.Hub.Queue(["Typeset",MathJax.Hub]);
</script>
```

Προφανώς έχει αλλαχθεί η εμφάνιση του κώδικα html για να χωρέσει στη σελίδα του περιοδικού. Ο κώδικας αυτός δηλώνει πως η σελίδα μας θα φορτώσει τον αντίστοιχο κώδικα JavaScript από το δικτυακό τόπο που παρέχει όλες τις εκδόσεις του MathJax. Αν έχουμε κατεβάσει τον κώδικα του MathJax, τότε μπορούμε να βάλουμε μετά τον διακόπτη `--mathjax` το σύμβολο `=` και μετά τη διεύθυνση στην οποία βρίσκεται ο κώδικας του MathJax. Προφανώς είναι καλύτερα να χρησιμοποιείτε την τελευταία έκδοση του κώδικα και άρα να χρησιμοποιείτε τον σχετικό διακόπτη χωρίς κάποια τιμή.

4 Επίλογος

Αν και το κείμενο είναι σχετικά σύντομο πιστεύω πως είναι τόσο όσο χρειάζεται ώστε κάποιος ν' αρχίσει να ψάχνει τις δυνατότητες του pandoc. Άλλωστε η τεκμηρίωση στο δικτυακό τόπο του προγράμματος είναι υπερεπαρκής. Οπότε βρείτε το δικτυακό τόπο του pandoc με τη χρήση της αγαπημένης σας μηχανής αναζήτησης και αρχίστε τις μετατροπές!