

TeXfrWpxml Group

John D. Baker

<https://github.com/bakerjd99/jacks/blob/master/texfrwpxml/TeXfrWpxml.ijs>

SHA-256: 6c4fadccb208ad1f01042e40ea14afc31effa30fabe7e5418b940be9917ed19e

April 10, 2023

Contents

TeXfrWpxml Overview	2
TeXfrWpxml Interface	2
TeXfrWpxml Notes	2
TeXfrWpxml Source Code	4
=: Index	56

TeXfrWpxml Overview

TeXfrWpxml is a J script that that extracts all published blog postings from WordPress export XML files and converts them, using [pandoc](#), to L^AT_EX and markdown files that are then used to build [PDF](#), [EPUB](#), [MOBI](#), and [AZW3](#) versions of a blog.

Using TeXfrWpxml is described in a series of blog posts:

1. <https://analyzethedatanotthedrive1.org/2012/02/11/wordpress-to-latex-with-pandoc-and-j-prerequisites-part-1/>
2. <https://analyzethedatanotthedrive1.org/2012/02/18/wordpress-to-latex-with-pandoc-and-j-latex-directories-part-2-2/>
3. <https://analyzethedatanotthedrive1.org/2012/02/25/wordpress-to-latex-with-pandoc-and-j-using-texfrwpxml-ijs-part-3/>

TeXfrWpxml Interface

BlogHashes	[12]	<i>update blog hashes</i>
FixBaddown	[13]	<i>attempt to convert *.baddown files to *.markddown</i>
LatexFrWordpress	[14]	<i>experimental conversion of Wordpress XML to LaTeX</i>
MainMarkdown	[16]	<i>assembles *.markdown files in EPUB directory in a master file</i>
MarkdownFrLatex	[18]	<i>converts edited LaTeX post files to image free markdown</i>
countYearposts	[25]	<i>adds post counts to table of content lines</i>

TeXfrWpxml Notes

TeXfrWpxml is at the mercy of WordPress's XML export format. This format has changed, without warning, many times during the last decade. Most changes do not impact TeXfrWpxml but in the last two years I've hit changes that do. This is a sign that TeXfrWpxml is nearing the end of it's life. I've been thinking about modifying or replacing the program, but if I take such a step, I will rethink my use of WordPress. The service is not entirely friendly to technical material, and is lagging well behind competitors in support of programmer friendly features like defining your our source code highlighting. For now I will note the changes that force me to tweak TeXfrWpxml if I see too many to forgive I will revisit this issue.

1. 20apr01 CDATA now used for blog post titles.

TeXfrWpxml Source Code

```

NB. *TeXfrWpxml s-- LaTeX source from WordPress export XML.
NB.
NB. verbatim: see the following
NB.
NB. source (this) script
NB. https://github.com/bakerjd99/jacks/blob/master/texfrwpxml/TeXfrWpxml.ijs
NB.
NB. https://github.com/bakerjd99/jacks/blob/master/texfrwpxml/wordpressstolatemwith2374.pdf
NB. https://github.com/bakerjd99/jacks/blob/master/texfrwpxml/wordpressstolatemwith2456.pdf
NB. https://github.com/bakerjd99/jacks/blob/master/texfrwpxml/wordpressstolatemwith2456.pdf
NB.
NB. interface word(s):
NB. -----
NB. BlogHashes      - update blog hashes
NB. FixBaddown      - attempt to convert *.baddown files to *.markddown
NB. LatexFrWordpress - experimental conversion of Wordpress XML to LaTeX
NB. MainMarkdown    - assembles *.markdown files in EPUB directory in a master file
NB. MarkdownFrLatex - converts edited LaTeX post files to image free markdown
NB. countYearposts  - adds post counts to table of content lines
NB.
NB. author:  John D. Baker
NB. created: 2012feb10
NB. -----
NB. 12feb24 (MarkdownFrLatex) added
NB. 12feb27 (FixBaddown, MainMarkdown) added

```

NB. 12feb29 (blogings) added
NB. 12jun28 (sortonpublishdate) added - publish order not always post id
NB. 12oct08 changed (texFrhtml) to process pandoc highlighted source
NB. 13dec20 save copy in GitHub (jacks) repository
NB. 15may06 (BlogHashes) added
NB. 17may13 (LATEXFIGURETEMPLATES) added
NB. 17sep29 use J 8.06 sha hash functions - removes need for external dll
NB. 20jul11 (BlogHashes) adjusted to track xhtml version
NB. 20sep12 (mdfootnotes) added to prefix markdown footnotes with post number
NB. 22mar29 (filenamesFrtid) adjusted to extract titles from new <!CDATA['s
NB. 22jul19 (countYearposts) added
NB. 23apr10 switch (BlogHashes) to sha-256

```
require 'task'
coclass 'TeXfrWpxml'
```

*NB.*dependents*

NB. declared global here to avoid confusing

NB. following HTML and LaTeX names with J names

NB. ()=: EPUBAMBLE EPUBFILE EPUBFRWPDIR HTMLREPS LATEXFIGURETEMPLATES LSTLISTINGHDR LSTLISTINGEND*

NB. ()=: MARKDOWNFILE TEXPREAMBLE TEXFRWPDIR TEXINCLUSIONS TEXSECTIONTITLE TEXWRAPFIGURE TEXROOTFILE*

NB. profile & require words ()=. IFIOS UNAME*

NB. dll/so is machine/os specific - assumes jqd 8.02 or later is installed

NB. OPENSSSL=: ;(IFIOS + (;'Win Linux Android Darwin') i. <UNAME) { 'libeay32.dll '; (2 \$ <'libssl.so ');
>..>(2 \$ <'libssl.dylib ')

```

NB. call dll
NB. cd=: 15!:0
NB. sslsha1=: (OPENSSL , ' SHA1 > + x *c x *c')&cd
NB.*enddependents

EPUBAMBLE=: 0 : 0
% Analyze the Data not the Drive1
% John D. Baker

)

NB. name of generated EPUB file
EPUBFILE=: 'bm.epub'

NB. root EPUB from LaTeX directory
EPUBFRWPDIR=: 'c:/pd/blog/wp2epub/'

NB. XML CDATA encoding and replacement for special characters
NB. stored in this form to hide the characters from web browsers
HTMLREPS=: 35 38 108 116 59 35 60 35 38 103 116 59 35 62 35 38 113 117 111 116 59 35 34 35 38 97 109 112 5
>...>9 35 38
HTMLREPS=: HTMLREPS{a.

NB. default lstlisting source block header
LSTLISTINGHDR=: 0 : 0
\begin{lstlisting}[frame=single,framerule=0pt,label=lst:~~~LSTLABEL~~~,

```

```

caption={source caption}]
)

NB. default lstlisting source block terminator
LSTLISTINGEND=: '\end{lstlisting}'

NB. name of generated *.markdown file
MARKDOWNFILE=: 'bm.markdown'

NB. name of LaTeX preamble file
TEXPREAMBLE=: 'bmamble.tex'

NB. root LaTeX from WordPress XML directory
TEXFRWPDIR=: 'c:/pd/blog/wp2latex/'

NB. immediate graphics subdirectory name, typically: inclusions
TEXINCLUSIONS=: 'inclusions'

NB. main LaTeX root file
TEXROOTFILE=: 'bm.tex'

NB. LaTeX post section title template
TEXSECTIONTITLE=: 0 : 0
\subsection*{\href{~~~POSTURL~~~}{~~~TITLETEXT~~~}}
\addcontentsline{toc}{subsection}{~~~TITLETEXT~~~}
)

```

NB. LaTeX wrapped figure template(s)

```
TEXWRAPFIGURE=: 0 : 0
%\captionsetup[floatingfigure]{labelformat=empty}
%\begin{figure}[htbp]
%\begin{floatingfigure}[l]{0.25\textwidth}
%\centering
%\includegraphics[width=0.23\textwidth]{~~~IMGRAPHICFILE~~~}
%\caption{~~~IMCAPTION~~~}
%\label{fig:~~~IMLABEL~~~}
%\end{floatingfigure}
%\end{figure}
)
```

```
LATEXFIGURETEMPLATES=: 0 : 0
```

```
% standard floating figure
% \captionsetup[figure]{labelformat=empty}
% \begin{figure}[htbp]
% \centering
% \href{}{\includegraphics[width=0.50\textwidth]{} }
% \caption{}
% \label{fig:???x0}
% \end{figure}
```

```
% captions beside figure
% \captionsetup[figure]{labelformat=empty}
% \begin{SCfigure}
```



```
% \centering
% \href{}\includegraphics[width=0.40\textwidth]{}
% \caption{}
% \label{fig:???x0}
% \end{SCfigure}

% wrapped figure - outer size > inner size
% \captionsetup[floatingfigure]{labelformat=empty}
% \begin{floatingfigure}[l]{0.23\textwidth}
% \centering
% \href{}\includegraphics[width=0.22\textwidth]{}
% \caption{}
% \label{fig:???x0}
% \end{floatingfigure}
)
```

*NB.*end-header*

NB. file extension given to tex files that do not convert to markdown
 BADDOWNEXT=: '.baddown'

NB. title marker must be an alphabetic string that is untouched by LaTeX
 BEGINTITLE=: 'BEWPTITLE'

NB. WordPress source code addon begin/end delimiters
 BESOURCEDELS=: <;._1 '|[sourcecode |[/sourcecode]'

NB. pandoc highlighted <pre> source block begin/end delimiters

BESOURCEPREDELS=: <;._1 '|<pre class="sourceCode|</pre>'

NB. carriage return character

CR=: 13{a.

NB. carriage return line feed character pair

CRLF=: 13 10{a.

NB. maximum length of alpha only part of file name

FILETITLELEN=: 20

NB. HTML file extension

HTMLTEXT=: '.html'

NB. interface words (IFACEWORDSTeXfrWpxml) group

IFACEWORDSTeXfrWpxml=: <;._1 ' BlogHashes FixBaddown LatexFrWordpress MainMarkdown MarkdownFrLatex countYea
>..>rposts'

NB. substitute for WordPress \$latex ... \$ blocks - must be untouched by latex

LATEXFRAGMARK=: 'LLLATEXFRAGGG'

NB. line feed character

LF=: 10{a.

NB. extension of markdown text files

MARKDOWNEXT=: '.markdown'

NB. markdown post/section prefix

MDSECTIONPFX=: ''

NB. pandoc shell command prefix

PANDOC CMD=: 'pandoc -o '

NB. root words (ROOTWORDSTeXfrWpxml) group

ROOTWORDSTeXfrWpxml=: <;._1 ' BlogHashes FixBaddown IFACEWORDSTeXfrWpxml LatexFrWordpress MainMarkdown Mark
>..>downFrLatex ROOTWORDSTeXfrWpxml SetTeXfrWpxmlPaths blogings countYearposts postfiles posttex showpass uedp
>..>osts'

NB. placeholder substitute for WordPress source blocks - must be untouched by LaTeX

SOURCEBLOCKMARK=: 'SSSOURCEBLOCKEEE'

NB. placeholder substitute for pandoc highlighted <pre> source blocks

SOURCEPREMARK=: 'SSSSOURCEPREBEEE'

*NB. temporary *.tex file - choose name to avoid clashes*

TEMPTExFILE=: 'temp.tex'

NB. LaTeX file extension

TEXEXT=: '.tex'

NB. temporary HTML file

TFWTEMPHTML=: 'temp.html'

NB. wget shell command prefix

WGETCMD=: 'wget --no-clobber --no-check-certificate --output-document='

BlogHashes=: 3 : 0

*NB.*BlogHashes v-- update blog hashes.*

NB.

NB. monad: BlogHashes uuIgnore

texpath=. 'c:/pd/blog/wp2latex/'

hash=. ctl ;"1 ' ' ,&.> sha256dir texpath,'*.tex'

NB. hash=. ctl ;"1 ' ' ,&.> sha256dir texpath,'.pdf'*

(toJ hash) write texpath,'bmpdfsha256.txt'

mdpath=. 'c:/pd/blog/wp2epub/'

hash=. ctl ;"1 ' ' ,&.> sha256dir mdpath,'*.epub'

NB. hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir mdpath,'.azw3'*

NB. hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir mdpath,'.mobi'*

hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir mdpath,'*.markdown'

(toJ hash) write mdpath,'bmepubsha256.txt'

xhtmlpath=. 'c:/pd/blog/wp2epub/xhtmll/'

hash=. ctl ;"1 ' ' ,&.> sha256dir xhtmlpath,'*.xhtml'

hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir xhtmlpath,'*.css'

```

hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir xhtmlpath,'*.ncx'
hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir xhtmlpath,'*.jpg'
(toJ hash) write xhtmlpath,'bmexhtmlsha256.txt'

onedrvpath=. 'C:/Users/baker/OneDrive/Public/pdocs/'
hash=. ctl ;"1 ' ' ,&.> sha256dir onedrvpath,'*.pdf'
hash=. hash, LF, ctl ;"1 ' ' ,&.> sha256dir onedrvpath,'*.epub'
(toJ hash) write onedrvpath,'bmsha256.txt'
)

FixBaddown=: 3 : 0

NB.*FixBaddown v-- attempt to convert *.baddown files to *.markddown
NB.
NB. monad:  FixBaddown uuIgnore
NB. dyad:    clDirectory FixBaddown uuIgnore

EPUBFRWPDIR FixBaddown y
:
epubdir=. x

NB. collect any *.baddown files
if. #files=. 0 {"1 (1!:0) EPUBFRWPDIR,'*',BADDOWNEXT do.

    files=. sortonid (<epubdir) ,&.> files
    outinext=. MARKDOWNEXT;TEXEXT
    fixed=. ''
    for_file. files do.

```

```

tex=. rmLatexGraphics read file=. ;file
texfile=. (t slash jpathsep epubdir),TEMPTEXFILE
(utf8 tex) write texfile
mdown=. outinext pandoc texfile
if. 0=#allwhitetrims mdown do.
  smoutput 'no markdown again -> ',file
else.
  fixed=. fixed,<file [ ferase file
  mdown write epubdir,(justfile@winpathsep file),MARKDOWNEXT
end.
outinext cleartemps texfile
end.
1;fixed
else.
  1;'no *',BADDOWNEXT,' file(s)'
end.
)

```

LatexFrWordpress=: 3 : 0

*NB.*LatexFrWordpress v-- experimental conversion of Wordpress XML*

NB. to LaTeX.

NB.

NB. monad: (iaRc;blcl) =. LatexFrWordpress clPathFileXML

NB.

NB. NB. window/linux

NB. LatexFrWordpress 'c:/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'

NB. LatexFrWordpress '/home/john/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'

NB.

NB. dyad: (iaRc;blcl) =. (clRoot;clPreamble;clDir;clIncl) LatexFrWordpress clPathFileXML

NB. LaTeX file & directory defaults

```
(TEXROOTFILE;TEXPREAMBLE;TEXFRWPDIR;TEXINCLUSIONS) LatexFrWordpress y
:
'texroot texpreamble texdir texincl'=. x
```

NB. must have a root tex file

```
if. -.fexist texdir,texroot do. 0;'missing or invalid LaTeX root file' return. end.
```

NB. read wordpress xml

```
if. fexist y do. xml=. read y else. 0;'missing or invalid XML export file' return. end.
```

NB. new published posts

```
if. #newposts=. (texdir;TEXEXT) prunePtable ptableFrwpxml xml do.
```

```
newposts=. sortposts newposts
titles=. texdir tfwTitles 1 {"1 newposts
'title post mismatch' assert (#titles) = #newposts
```

```
predir=. texpreamble;texdir
for_post. newposts do.
  smoutput ;post_index{titles
  pdat=. (post_index{titles),(2 3 4{post),<cdatatext;5{post
  tex=. predir texFrhtml pdat
```

```

    NB. append common figure code to each post
    tex=. tex,CRLF,LATEXFIGURETEMPLATES

    tex write texdir,(;0{post),TEXEXT
end.

NB. adjust root tex file to reference new posts
NB. no additions if files already referenced
tex=. read root=. texdir,texroot
mask=. -(0 {"1 newposts) 1&e.@E.&> <tex
newposts=. mask#newposts
titles=. mask#titles
tex=. tex inputposts newposts
tex write root

    NB. result titles of new posts
    1;titles
else.
    1;'no new posts'
end.
)

MainMarkdown=: 3 : 0

NB.*MainMarkdown v-- assembles *.markdown files in EPUB directory in a master
NB. file.
NB.
NB. monad: bl =. MainMarkdown clPathFile

```



```

NB.
NB.   MainMarkdown 'c:/pd/blog/wordpress/analyzethedatanotthedrivel.wordpress.xml'
NB.
NB. dyad:   bl =. (clMdownfile;clDirectory;clAmble) MainMarkdown clPathFile

(MARKDOWNFILE;EPUBFRWPDIR;EPUBAMBLE;MDSECTIONPFX) MainMarkdown y
:
'mdownfile epubdir epubamble mdsecpfx'=. x

NB. read wordpress xml - valid posts
if. fexist y do. xml=. read y else. 0;'missing or invalid XML export file' return. end.

pfiles=. 0 {"1 (1!:0) epubdir,'*',MARKDOWNNEXT

NB. keep only post markdown files
ptable=. ptableFrwpxml xml
posts=. (0{"1 ptable) ,&.> <MARKDOWNNEXT
pfiles=. pfiles -. pfiles -. posts

NB. sort files by publish date
pfiles=. ptable sortonpublishdate pfiles
files=. (<epubdir) ,&.> pfiles

NB. NOTE: sometimes posts are written long before they are published
NB. sort files by trailing post id
NB. files=. sortonid files

```

```
NB. mash posts together - affix date
epubamble=. (allwhitetrim epubamble),LF,('% ',timestamp ''),2#LF
posts=. ; (<mdsecpfx) ,&.> (allwhitetrim&.> read&.> files) ,&.> <2#LF
posts=. utf8 toHOST epubamble,(2#LF),posts
posts write file=. epubdir,mdownfile
1;((":#files),' post(s)');file
)
```

```
MarkdownFrLatex=: 3 : 0
```

```
NB.*MarkdownFrLatex v-- converts edited LaTeX post files to image
NB. free markdown.
```

```
NB.
```

```
NB. This verb converts edited *.tex files into *.markdown which
NB. are then used to build an EPUB. The markdown requires a small
NB. bit of editing, mostly to cleanup the odd LaTeX fragment that
NB. Pandoc does not convert. This is nowhere near the chore that
NB. editing the WordPress CDATA HTML to *.tex is and has the nice
NB. feature of preserving all the corrections made to the *.tex
NB. files.
```

```
NB.
```

```
NB. monad: bl =. MarkdownFrLatex clPathFileXML
```

```
NB.
```

```
NB. MarkdownFrLatex 'c:/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'
```

```
NB.
```

```
NB. dyad: bl =. blcl MarkdownFrLatex clPathFileXML
```

```
(MARKDOWNFILE;EPUBFRWPPDIR;TEXFRWPPDIR) MarkdownFrLatex y
```

```

:
'markfile epubdir texdir'=. x

NB. read wordpress xml - defines post order
if. fexist y do. xml=. read y else. 0; 'missing or invalid XML export file' return. end.

NB. require 'regex' !(*)=. rxutf8
NB. turn off utf8 for regex - required for J 9.04
rxp=. rxutf8 0

NB. posts without markdown versions
if. #newposts=. (epubdir; MARKDOWNEXT) prunePtable ptableFrwpxml xml do.

  cvtitles=. ''
  newposts=. sortposts newposts
  outinext=. MARKDOWNEXT; TEXEXT
  for_post. newposts do.
    'post title'=. 0 1{post
    if. fexist tex=. texdir, post, TEXEXT do.
      tex=. rmLatexGraphics read tex
      texfile=. (t slash jpathsep epubdir), TEMPTEXFILE
      (utf8 tex) write texfile
      mdown=. outinext pandoc texfile
      NB. if pandoc cannot convert to markdown it returns nothing
      if. 0=#allwhitetrims mdown do.
        smoutput 'no markdown -> ', title
        NB. save the original file - tweaking is necessary

```

```

        (utf8 tex) write epubdir,post,BADDOWNEXT
    else.
        mdown=. post mdfootnotes mdown
        mdown write epubdir,post,MARKDOWNEXT
    end.
    outinext cleartemps texfile
    cvtitle=. cvtitles,<title
else.
    smoutput 'skipping missing *.tex file -> ',post,TEXEXT
end.
end.
1;cvtitles

else.
    1;'no new posts'
end.
)

```

```
SetTeXfrWpxmlPaths=: 3 : 0
```

*NB.*SetTeXfrWpxmlPaths v-- sets OS dependent paths.*
NB.
NB. Customize the path and file settings in this verb
NB. to match your locations.
NB.
NB. monad: SetTeXfrWpxmlPaths uuIgnore

NB. system nouns !()=. IFWIN IFUNIX*

```

if.      IFWIN  do.
  TEXFRWPDIR=: 'c:/pd/blog/wp2latex/'
  EPUBFRWPDIR=: 'c:/pd/blog/wp2epub/'
elseif. IFUNIX do.
  TEXFRWPDIR=: '/home/john/pd/blog/wp2latex/'
  EPUBFRWPDIR=: '/home/john/pd/blog/wp2epub/'
elseif.do.
  'not on supported OS' assert 0
end.

```

NB. TeX root, preamble, inclusions subdirectory name

```

TEXROOTFILE=: 'bm.tex'
TEXPREAMBLE=: 'bmamble.tex'
TEXINCLUSIONS=: 'inclusions'

```

NB. EPUB eBook file

```

EPUBFILE=: 'bm.epub'

```

NB. standardize document directory paths

```

TEXFRWPDIR=: tlash jpathsep TEXFRWPDIR
EPUBFRWPDIR=: tlash jpathsep EPUBFRWPDIR
)

```

NB. retains string (y) after last occurrence of (x)

```

afterlaststr=: ] }.~ #@[ + 1&(i:~)@([ E. ])

```

NB. retains string after first occurrence of (x)

```

afterstr=: ] }.~ #@[ + 1&(i.~)@([ E. ])

```

NB. trims all leading and trailing white space

```
allwhitetrims=: ] #~ [: -. [: (*./\ . +. */.\) ] e. (9 10 13 32{a.})"
```

NB. signal with optional message

```
assert=: 0 0" _ $ 13!:8^:((0: e. ])^(12"_))
```

NB. extracts text of xml attribute

```
attrvalue=: '""' _ beforestr ([ , '=' _) afterstr '>' _ beforestr ]
```

NB. retains string (y) before last occurrence of (x)

```
beforelaststr=: ] {~ 1&(i~)@([ E. ])
```

NB. retains string before first occurrence of (x)

```
beforestr=: ] {~ 1&(i~)@([ E. ])
```

```
blogimgs=: 3 : 0
```

*NB.*blogimgs v-- extracts all images referenced in post CDATA.*

NB.

NB. monad: btcl =. blogimgs btclPosts

NB.

NB. blogimgs posts NB. see (ptableFrwpxml)

```
if. 0=#y do. 0 3$' ' return. end.
```

NB. cut CDATA on <img's

```
txt=. ;(cdatatext&.> 5 {"1 y) ,&.> LF
cimg=. ('>'&beforestr) &.> ( '<'<img ' E. txt) <.;.1 txt
```

NB. form table of titles and src urls

```
cimg=. ('title'&attrvalue ; 'src'&attrvalue) &> cimg
```

NB. prefix file name

```
cimg ,.~ ('?'&beforestr)@('/'&afterlaststr) &.> {"1 cimg
)
```

NB. boxes open nouns

```
boxopen=: <^(L. = 0:)
```

NB. extract character list from HTML CDATA

```
cdatatext=: [: ']]>'&beforelaststr '<'[CDATA['&afterstr
```

```
changestr=: 4 : 0
```

*NB.*changestr v-- replaces substrings - see long documentation.*

NB.

NB. dyad: clReps changestr cl

NB.

NB. NB. first character delimits replacements

NB. '/change/becomes/me/ehh' changestr 'blah blah ...'

```
pairs=. 2 {"(1) _2 [\ <.;._1 x      NB. change table
```

```
cnt=._1 [ lim=. # pairs
```

```

while. lim > cnt=.>:cnt do.      NB. process each change pair
  't c'=. cnt { pairs           NB. /target/change
  if. +./b=. t E. y do.         NB. next if no target
    r=. I. b                    NB. target starts
    'l q'=. #&> cnt { pairs      NB. lengths
    p=. r + 0,+/\(<:# r)$ d=. q - 1 NB. change starts
    s=. * d                      NB. reduce < and > to =
    if. s = _1 do.
      b=. 1 #~ # b
      b=. ((1 * # r)$ 1 0 #~ q,l-q) (,r +/ i. l)} b
      y=. b # y
      if. q = 0 do. continue. end. NB. next for deletions
    elseif. s = 1 do.
      y=. y #~ >: d r} b        NB. first target char replicated
    end.
    y=. (c $~ q *# r) (,p +/i. q)} y NB. insert replacements
  end.
end. y                          NB. altered string
)

charsub=: 4 : 0

NB.*charsub v-- single character pair replacements.
NB.
NB. dyad: clPairs charsub cu
NB.
NB.   '-_$ ' charsub '$123 -456 -789'

```



```
'f t'=. ((#x)$0 1)<@,&a./ .x
t {~ f i. y
)
```

```
cleartemps=: 3 : 0
```

```
NB.*cleartemps v-- erase temporary HTML/TEX & TEX/MARKDOWN files
NB.
```

```
NB. monad: cleartemps clPathFile
```

```
NB. dyad: (clOutExt;clInExt) cleartemps clPathfile
```

```
(TEXEXT;HTMLEXT) cleartemps y
```

```
:
```

```
'outext inext'=. x
```

```
(inext,' extension required') assert 1 e. inext E. y
```

```
ferase y;('.'&beforelaststr y),outext
```

```
)
```

```
countYearposts=: 3 : 0
```

```
NB.*countYearposts v-- adds post counts to table of content lines.
```

```
NB.
```

```
NB. NOTE: this verb depends on the layout of (bm.tex).
```

```
NB.
```

```
NB. monad: clTex =. countYearposts clFileTeX
```

```
NB.
```

```
NB. NB. root latex file
```

```

NB. tex=. 'c:\pd\blog\wp2latex\bm.tex'
NB. (toHOST countYearposts tex) write tex

NB. read and cut latex
ctex=. tex <.1~ 1 (0)}'\addcontentsline{toc}{section}' E. tex=. read y

NB. post counts
pcnts=. +/@('input{'&E.)&.> }.ctex

NB. add content lines
atocs=. ('---'&beforelaststr)@('}'&beforelaststr)@(\LF&beforestr)&.> }.ctex

NB. append counts
atocs=. atocs ,&.> (<' --- ') ,&.> (":&.> pcnts) ,&.> <' posts}' ,CRLF

NB. clip old content lines and append new
;(0{ctex) , atocs ,&.> \LF&afterstr&.> }.ctex
)

NB. character table to newline delimited list
ctl=: }.@(:,@ (1&(",1)@(-.@(*./\."1@(&(' '0]]))) # ,@((10{a.)&(",1)@)])

cutincludgraphicsidx=: 3 : 0

NB.*cutincludgraphicsidx v-- cut list into \includgraphics
NB. LaTeX and other
NB.

```

```

NB. monad: (ilIdx ;< blcl) =. cutincludegraphicsidx clTex

('\includegraphics{' ;'}';0) cutpxtidx y
)

cutlatexidx=: 3 : 0

NB.*cutlatexidx v-- cut list into WordPress LaTeX and other.
NB.
NB. monad: (ilIdx ;< blcl) =. cutlatexidx clHtml
NB.
NB. cutlatexidx ' ... yada yada $latex frac{a}{b}$ and so on ... '

('$latex' ; '$';1) cutpxtidx y
)

cutnestidx=: 4 : 0

NB.*cutnestidx v-- cut list into nested runs and other.
NB.
NB. Nested runs are delimited by begin and end tags. This verb is
NB. oriented toward XML parsing where typical begin end tags are
NB. <ul> </ul> and tags with attributes like: <hoo boy="2">
NB. </hoo>
NB.
NB. This verb can process numeric lists but care must be taken to
NB. insure the pad item (1{.0$y) does not match begin and end

```

```

NB. values.
NB.
NB. dyad: (ilIdx ;< blcl) =. (clStart;clEnd) cutnestidx cl
NB.      (ilIdx ;< blnl) =. (nlStart;nlEnd) cutnestidx nl
NB.
NB. xml=. 'yada <ol><li>one</li><ol><li>sub one</li></ol></ol> boo'
NB.      ('<ol';'</ol>') cutnestidx xml
NB.
NB.      88 99 cutnestidx (i.5),88,(10?10),99 88 5 5 5 5 5 99

if. #y do.
  's e'=. ,&.> x          NB. start end lists
  ut=. 1{.0$y             NB. padding
  assert. -.s -: e        NB. they must differ
  assert. -. (s -:ut) +. e -:ut
  sp=. s E. ut=.y,ut       NB. start mask

  NB. quit if no delimiters
  if. -.1 e. sp do. (i.0);<<y return. end.

  ep=. e E. ut            NB. end mask
  assert. (+/sp) = +/ep    NB. basic balance
  dp=. sp + - ep          NB. start end marks
  assert. 0 *./ . <: +/\ dp NB. nested balance
  ep=. I. _1=dp [ sp=. I. 1=dp NB. start end indexes
  ut=. +/\dp -. 0         NB. scanned marks
  dp=. /:~ sp,ep          NB. all indexes

```

```

sp=. (firstones 1<:ut)#dp      NB. starts of nested
ep=. (#e)+(0=ut)#dp          NB. starts of other
dp=. /:~ ~.0,sp,ep           NB. cut starts
ut=. }: 1 dp} (>:#y)#0       NB. cut mask
(dp i. sp);<ut <;.1 y        NB. nest indexes cut list
else.
  (i.0);<<y                  NB. empty arg result
end.
)

cutpxtidx=: 4 : 0

NB.*cutpxtidx v-- cut list into prefix with character terminator
NB. and other.
NB.
NB. monad: (ilIdx ;< blcl) =. (clPfx;caEend;iaPos) cutpxtidx clHtml
NB.
NB. ('$latex';'$';1) cutpxtidx ' ... yada yada $latex frac{a}{b}$ and so on ... '
NB. ('\includegraphics{';'}';0) cutpxtidx ' boo hooo \includegraphics{pictures.png} et cetera'

's e p'=. x NB. start end position
assert. (1=#e) *. 1=#p
if. 1 e. b=. s E. ,y do.
  sp=. I. b
  op=. (0 e. sp) }. 0,sp + >:p&{@I.@(e&=)&> b <;.1 y
  op=. /:~ sp,op -. #y
  (op i. sp) ;< (1 op} b) <;.1 y
else.

```

```

    (i.0);<<y
end.
)

cutstridx=: 4 : 0

NB.*cutstridx v-- cut list into (x) and other.
NB.
NB. dyad: (ilIdx ;< blcl) =. clStr cutstridx cl
NB.
NB.    'CHOP' cutstridx 'CHOP CHOP me up CHOP eh'

if. 1 e. b=. x E. ,y do.
    sp=. I. b
    op=. (0 e. sp) }. 0,sp + #x
    op=. /:~ sp,op -. #y
    (op i. sp) ;< (1 op} b) <;.1 y
else.
    (i.0);<<y
end.
)

NB. boxes UTF8 names
fboxname=: ([: < 8 u: >) ::]

NB. erase files - cl / blcl of path file names
ferase=: 1!:55 ::(_1:)@(fboxname&>)@boxopen

```

```

NB. 1 if file exists 0 otherwise
fexist=: 1:@(1!:(4) ::0:@(fboxname&>))@boxopen

filenamesFrtid=: 3 : 0

NB.*filenamesFrtid v-- form file names from titles and ids.
NB.
NB. monad: blclFilename =. filenamesFrtid btclTitleId
NB.
NB.   wpxml=. read 'c:/pd/blog/wordpress/analyzethedatanotthedrivvel.wordpress.xml'
NB.   posts=. ptableFruwpxml wpxml
NB.   filenamesFrtid 0 1 {"1 posts

NB. NOTE: there's been a change in WordPress (22mar29) XML that
NB. encloses blog post titles in <![CDATA['s cdatatext extracts
NB. the titles - will fail if there are no <![CDATA['s

NB. NOTE: this algorithm for forming file names has proven very
NB. stable over years of use. The generated names are readable,
NB. portable and unique. Sometimes you get things right.

NB. remove all but upper and lowercase alpha and lower case remainder
fn=. (cdatatext&.> 0 {"1 y) tolower@-.&.> <a.-.((65+i.26),97+i.26){a.

NB. take at most FILETITLELEN chars and append unique post id
((FILETITLELEN <. #&> fn) {.&.> fn) ,&.> 1 {"1 y
)

```

NB. 0's all but first 1 in runs of 1's - like (firstone) but differs for nulls
firstones=: > (0: , }:)

NB. size of file in bytes
fsize=: 1!:4 ::(_1:)@(fboxname&>)@boxopen

getNewgraphics=: 3 : 0

*NB.*getNewgraphics v-- downloads graphics files referenced in*
NB. LaTeX.
NB.
NB. monad: ((<blclDown),<blclMissing) =. getNewgraphics clTex

none=. '';
if. y do.

NB. extract any graphics urls
graphic=. '\includegraphics{'
mask=. graphic E. y
if. -.1 e. mask do. none return. end.
urls=. (#graphic) }.&.> mask <|.1 y
urls=. ('?'&beforestr)@('}'&beforestr)&.> urls

NB. download images to inclusions/ directory
ifiles=. (<TEXFRWPDIR,tlslash TEXINCLUSIONS) ,&.> '/'&afterlaststr&.> urls
wcmds=. WGETCMD ,"1 > ifiles ,&.> ' ' ,&.> urls


```

NB. require 'task' !(*)=. shell
skipcnt=. 0 [ dfiles=. mfiles=. ''
for_cmd. wcmds do.
  file=. ;cmd_index{ifiles

  NB. skip files that exist in inclusions/
  if. fexist file do. skipcnt=.:skipcnt continue. end.

  shell cmd [ smoutput 'downloading: ',file
  if. 0<fsize file do. dfiles=. dfiles,<file
  else.
    ferase file NB. clears any 0 byte files
    mfiles=. mfiles,<file [ smoutput 'warning - did not download: ',file
  end.
end.
smoutput (":#dfiles),' downloaded; ',(":#mfiles),' not downloaded; ',(":skipcnt),' skipped'
(<dfiles),<mfiles NB. downloaded & not downloaded

else.
  none
end.
)

htmlParagraphs=: 3 : 0

NB.*htmlParagraphs v-- mark missing html paragraphs.
NB.
NB. WordPress HTML frequently omits paragraph tags <p> </p>.

```

*NB. Missing paragraph tags cause Pandoc to run paragraphs
NB. together in the generated LaTeX. This verb inserts leading
NB. <p> tags in LF delimited runs. The vast majority of such runs
NB. are paragraphs.*

NB.

NB. monad: cl =. htmlParagraphs clHtml

```
if. 1 e. '<p>' E. y do. y
else.
```

NB. cut paragraphs

```
cs=. <.;_2 tlf y -. CR
```

NB. tag or newlines

```
tnl=. ((2#LF)"_)`('<p>'&,)@.(0 < #)
```

```
; tnl&.> cs
```

```
end.
```

```
)
```

```
inputposts=: 4 : 0
```

*NB.*inputposts v-- appends new %\input{file.tex} commands to root*

NB. tex.

NB.

NB. dyad: clTex =. clTex inputposts btclPosts

```
if. #y do.
```

```
bp=. '%</blogposts>'
```

```
new=. ;(<'%\input{' ,&.> (0 {"1 y) ,&.> (<TEXEXT,'} %') ,&.> (3 {"1 y) ,&.> LF
```

```
head=. bp&beforestr x
```

```

    tail=. bp&afterstr x
    head,LF,new,bp,tail
else.
    x
end.
)

```

NB. standarizes J path delimiter to unix/linux forward slash

```
jpathsep=: '/'&(('\' I.@:= ])) }
```

NB. extracts the drive from qualified file names

```
justdrv=: [: }: ] #~ [: +./\ . ' : '&=
```

NB. extracts the extension from qualified file names

```
justext=: '"_`([ #~ [: -. [: +./\ . ' : '&=)@.(' : '&e.)
```

NB. file name from fully qualified file names

```
justfile=: ([ #~ [: -. [: +./\ . ' : '&=)@([ #~ [: -. [: +./\ . e.&' : \')
```

NB. extracts only the path from qualified file names

```
justpath=: [: }: ] #~ ([: -. [: +./\ . ' : '&=) *. [: +./\ . '\ '&=
```

```
lstFrsrcb=: 4 : 0
```

*NB.*lstFrsrcb v-- lstlisting from source block.*

NB.

```

NB. monad:  cl =. clPid lstFrsrcb clSrc

'start end'=. BESOURCEDELS

NB. first line is block header with wp addon parameters
head=. LF&beforestr y
body=. LF&afterstr end&beforelaststr y

NB. revert special CDATA HTML characters
body=. HTMLREPS changestr body

NB. insert label uses post id and scr block cnt to be unique
lstlisting=. ('#~~~LSTLABEL~~~#scr',x) changestr LSTLISTINGHDR

NB. leave original header as latex comment
LF,'% ',head,LF,lstlisting,LF,body,LSTLISTINGEND
)

mdfootnotes=: 4 : 0

NB.*mdfootnotes  v-- prefix numbered pandoc markdown footnotes
NB. with post number.
NB.
NB. When building EPUB documents from many markdown texts it's
NB. important that all footnote labels are unique. The standard
NB. practice of [^1], [^2], ... leads to clashes. This verb
NB. prefixes all numbered pandoc footnotes with the post number.
NB. The post number is a unique blog specific integer assigned by

```

```

NB. WordPress.com. A superior unique key like a secure hash is
NB. overkill here but may be necessary if you are merging posts
NB. across many blogs.
NB.
NB. dyad: cl =. clPostId mdfootnotes clMd
NB.
NB. md=. read 'c:/pd/blog/wp2epub/oscarsnowasmeaning6975.markdown'
NB. 'oscarsnowasmeaning6975' mdfootnotes md

NB. require 'regex' !(*)=. rxall rxmatches rxmerge rxutf8

NB. turn off utf8 for regex - required for J 9.04
rxp=. rxutf8 0

fp=. '\[\\^[0-9]\\]' NB. pandoc footnotes

if. #fn=. >fp rxall y do.
  NB. post number
  pn=. 'x' ,~ x #~ x e. '0123456789'
  NB. relabeled footnotes
  nfn=. <"1 ' ' -."1~ ((0 1 {"1 fn) ,"1 pn) ,"1 > 2 }."1 fn
  nfn (fp rxmatches y) rxmerge y
else.
  y NB. no footnotes
end.
)

pandoc=: 3 : 0

```

*NB.*pandoc v-- shells pandoc to convert HTML->LaTeX & LaTeX->Markdown.*

NB.

NB. monad: cl =. pandoc clFile

NB.

NB. tex=. pandoc 'c:/temp/cdata.html'

NB.

NB. dyad: cl =. (clOutExt;clInExt) pandoc clFile

NB.

NB. markdown=. (MARKDOWNEXT;TEXEXT) pandoc 'c:/temp/post.tex'

(TEXEXT;HTMLEXT) pandoc y

:

'outext inext'=. x

y=. winpathsep y

(inext,' extension required') assert (inext-.'.') -: justext y

file=. justfile y

drv=.]`([@,&':')@.(0 < #) justdrv y

dir=. tlash drv,justpath y

NB. output written to same directory as source

in=. jpathsep y [out=. jpathsep dir,file,outext

(inext,' file must exist') assert fexist in

ferase out

NB. require 'task' !()=. shell*

shell PANDOC CMD,' ',out,' ',in

(outext,' conversion failed') assert fexist out

read out

```

)

postTitleDate=: 3 : 0

NB.*postTitleDate v-- post LaTeX section title code.
NB.
NB. monad: cl =. postTitleDate (clTitle;clDate;clUrl)

'ptitle pdate purl'= . y
reps=. '|~~~TITLETEXT~~~|',allwhitetrims pttitle
reps=. reps,'|~~~POSTURL~~~|',allwhitetrims purl
pttitle=. reps changeSTR TEXSECTIONTITLE
pdate=. 'Posted: ',timestamp ".-' : ' charsub pdate
pttitle,(2#LF),'\noindent\emph{'',pdate,'}',LF,'\vspace{6pt}','2#LF
)

postfiles=: 3 : 0

NB.*postfiles v-- list of post LaTeX files.
NB.
NB. monad: blclTexfiles =. postfiles uuIgnore

NB. system nouns !(*)=. IFWIN IFUNIX
if. IFWIN do. wpxml=. read 'c:/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'
elseif. IFUNIX do. wpxml=. read '/home/john/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'
elseif.do.
  'not on supported os' assert 0

```

```

end.
posts=. ptableFrwpxml wpxml
(<TEFRWPDIR) ,&.> (0 {"1 posts) ,&.> <TEEXT
)

postid=: 3 : 0

NB.*postid v-- test verb that forms (texFrhtml) (y) arguments.
NB.
NB. monad: postid iaPid
NB. dyad: (clStatus;clType) postid iaPid
NB.
NB.    (;:'draft post') postid '' NB. drafts

(;:'publish post') postid y
:
NB. !(*)=. IFWIN IFUNIX posts list nc
if.      IFWIN do. wpxml=. read 'c:/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'
elseif. IFUNIX do. wpxml=. read '/home/john/pd/blog/wordpress/analyzethedatanotthedrive1.wordpress.xml'
elseif.do.
    'not on supported OS' assert 0
end.
posts=: x ptableFrwpxml wpxml
pids=. 2 {"1 posts
if. 0=#y do. list pids return. end.
y=. ":y
if. (<y-.' ') e. pids do.
    pt=. posts {~ pids i. <y NB. post id

```



```
(tfwTitles 1{pt}),(2 3 4{pt}),<cdatatext;5{pt}
else.
  smoutput 'no post with pid: ',y
end.
)
```

```
posttex=: 3 : 0
```

```
NB.*posttex v-- LaTeX code for single post/draft.
NB.
NB. monad: clTex =. posttex iaPid
NB. dyad: clTex=. (clStatus;clType) posttex iaPid
NB.
NB. tex=. (;'draft post') posttex 638
```

```
(;:'publish post') posttex y
:
texFrhtml x postid y
)
```

```
prunePtable=: 3 : 0
```

```
NB.*prunePtable v-- removes post table entries that have
NB. corresponding files.
NB.
NB. monad: btcl =. prunePtable btclPosts
NB.
```

```

NB.  prunePtable posts NB. see (ptableFrwpxml)
NB.
NB. dyad: btcl =. (clDirectory;clExt) prunePtable btclPosts

(TEXTFRWPDIR;TEXEXT) prunePtable y
:
'path ext'=. x
y #~ -.fexist (<path) ,&.> (0 {"1 y) ,&.> <ext
)

ptableFrwpxml=: 3 : 0

NB.*ptableFrwpxml v-- type status table from wordpress xml.
NB.
NB. monad: btcl =. ptableFrwpxml clXml
NB.
NB.  wpxml=. read 'c:/pd/blog/wordpress/analyzethedatanotthedrivvel.wordpress.xml '
NB.  posts=. ptableFrwpxml wpxml
NB.
NB. dyad: btcl =. (clStatus;clType) ptableFrwpxml clXml
NB.
NB.  drafts=. (;:'draft post') ptableFrwpxml wpxml

(;:'publish post') ptableFrwpxml y
:
NB. cut items
cxml=. ('<item>' E. y) <;.1 y

```

NB. item attribute extractors

```
istatus=.  [: '</wp:status>'&beforestr&.> '<wp:status>'&afterstr&.>
itype=.    [: '</wp:post_type>'&beforestr&.> '<wp:post_type>'&afterstr&.>
ipostid=.  [: '</wp:post_id>'&beforestr&.> '<wp:post_id>'&afterstr&.>
ititle=.   [: '</title>'&beforestr&.> '<title>'&afterstr&.>
ilink=.    [: '</link>'&beforestr&.> '<link>'&afterstr&.>
idate=.    [: '</wp:post_date_gmt>'&beforestr&.> '<wp:post_date_gmt>'&afterstr&.>
icontent=. [: '</content:encoded>'&beforestr&.> '<content:encoded>'&afterstr&.>
```

NB. all status + types

```
ppxml=. cxml #~ x -:"1 (istatus ,. itype) cxml
```

NB. return filename, title, id, date, link, content

```
ppxml=. (ititle ,. ipostid ,. idate ,. ilink ,. icontent) ppxml
(filename$Frtid 0 1 {"1 ppxml) ,. ppxml
)
```

NB. reads a file as a list of bytes

```
read=: 1!:1&([`<@.(32&>@.(3!:0)))
```

```
rmLatexGraphics=: 3 : 0
```

*NB.*rmLatexGraphics v-- remove/blank out LaTeX graphics.*

NB.

NB. This verb removes LaTeX comments and graphics environments

*NB. from *.tex. This is done to produce lightweight EPUB and MOBI*

NB. eBook versions that perform well on Kindles, iPhones, iPads

```

NB. and so forth.
NB.
NB. monad: cl =. rmLatexGraphics clTex
NB.
NB. tex=. read 'c:/pd/blog/wp2latex/cowboysandaliansando1698.tex'
NB. rmLatexGraphics tex

rp=. <' '
tex=. <:.2 tlf y -. CR
tex=. ;('%' ~: {. @allwhitetrims> tex) # tex
'ix cs'=. ('\begin{floatingfigure}';'\end{floatingfigure}') cutnestidx tex
tex=. ;rp ix} cs
'ix cs'=. ('\begin{SCfigure}';'\end{SCfigure}') cutnestidx tex
tex=. ;rp ix} cs
'ix cs'=. ('\begin{figure}';'\end{figure}') cutnestidx tex
if. envtex=. ;ix{cs
  b0=. -.1 e. '\begin{minipage}' E. envtex
  b1=. 1 e. '\includegraphics' E. envtex
  b1 *. b0 do. tex=. ;rp ix} cs
end.
NB. clear any remaining caption setup pandoc passes them to .markdown
'ix cs'=. ('\captionsetup';'}';0) cutpntidx tex
tex=. ;rp ix} cs
)

NB. sha-256 hash from bytes: sha256 'hash me again'
sha256=: 3&(128!:6)

```

```
sha256dir=: 3 : 0
```

*NB.*sha256dir v-- compute sha256 hashes for files matching pattern in directory.*

NB.

NB. monad: btcl =. sha256dir clPathRoot

NB.

NB. sha256dir 'c:/pd/blog/wp2latex/.tex'*

```
jfe=. ] #~ [: -. [: +./\.'/'&= NB. just file extension
```

NB. code used before J 8.06

NB. hexadecimal list from integers

NB. hdl=. [: , [: hfd2 a. i.]

NB. standard profile !()=. dir*

NB. (jfe&.> files) ,.~ hdl @ sha1 @ read&.> files=. 1 dir jpathsep y

```
(jfe&.> files) ,.~ sha256 @ read&.> files=. 1 dir jpathsep y
)
```

NB. show and then pass noun

```
showpass=: ] [ 1!:2&2
```

NB. session manager output

```
smoutput=: 0 0 $ 1!:2&2
```

```
sortonid=: 3 : 0
```

*NB.*sortonid v-- sort files by trailing post id - monotonically increasing*
NB.

NB. monad: blcl=. sortonid blclFiles

```
(/: "&> ('/'&afterlaststr&.> y) -.&.> <a. -. '0123456789') { y
)
```

```
sortonpublishdate=: 4 : 0
```

*NB.*sortonpublishdate v-- sort markdown post files by publish date.*
NB.

NB. dyad: blcl =. btclPosts sortonpublishdate blclFiles

NB. posts and dates

```
postdate=. 0 3 {"1 x
```

NB. selected post files without extension

```
pfiles =. '.'&beforelaststr&.> y
```

```
mask=. pfiles e. 0 {"1 postdate
```

```
postdate=. mask # postdate
```

```
files=. 0 {"1 (/: 1 {"1 postdate){postdate
```

```
files ,&.> <MARKDOWNEXT
```

```
)
```

```
sortposts=: 3 : 0
```

```

NB.*sortposts v-- sort posts chronologically.
NB.
NB. monad:  bt =. sortposts blclPosts

(/: ". ' - : ' charsub >3 {"1 y){y
)

texFrhtml=: 3 : 0

NB.*texFrhtml v-- convert WordPress HTML fragments to LaTeX
NB. fragments
NB.
NB. monad:  clTex =. texFrhtml (clTitle;clDate;clPid;clHtml)
NB.
NB.  wpxml=. read 'c:/pd/blog/wordpress/analyzethedatanotthedriviel.wordpress.xml '
NB.  posts=. ptableFrwpxml wpxml
NB.  pt=.    posts {~ (2 {"1 posts) i. <'3303' NB. post id
NB.  tpdup=. (tfwTitles 1{pt),(2 3 4{pt),<cdatatext;5{pt
NB.  tex=.   texFrhtml tpdup NB. title, pid, date, url, post text
NB.
NB. dyad:    clTex =. (clPreamble;clDir) texFrhtml (clTitle;clDate;clPid;clUrl;clHtml)

(TEXPREAMBLE;TEXFRWPDIR) texFrhtml y
:
'texpreamble texdir'=. x

NB. title, id, date, url, html
'ptitle pid pdate url htm'=. y

```

```
if. 0=#htm do. '' return. end.
```

```
cm=. 1&e.@E.
```

```
sblk=. utf8 ,SOURCEBLOCKMARK [ lfrg=. utf8 ,LATEXFRAGMARK [ pblk=. utf8 ,SOURCEPREMARK
```

```
NB. NIMP test is not exhaustive
```

```
'markers must not be substrings' assert -. +./(sblk cm lfrg),(lfrg cm sblk),(sblk cm pblk),lfrg cm pblk
```

```
NB. leave commented warning about HTML tables
```

```
tabwarn=. ('</table>' 1&e.@E. htm)#'%% HTML table in source - edits required'
```

```
NB. hide Pandoc highlighted source code <pre ... </pre> blocks
```

```
'ixpre cspre'=. BESOURCEPREDELS cutnestidx htm
```

```
if. #ixpre do.
```

```
  'source pre marker in .html' assert -. 1 e. pblk E. htm
```

```
  htm=. ; (<pblk) ixpre} cspre
```

```
end.
```

```
NB. hide WordPress [sourcecode ... ] blocks
```

```
'ixsrc cssrc'=. BESOURCEDELS cutnestidx htm
```

```
if. #ixsrc do.
```

```
  'source block marker in .html' assert -. 1 e. sblk E. htm
```

```
  htm=. ; (<sblk) ixsrc} cssrc
```

```
end.
```

```
NB. hide WordPress $latex ... $
```

```
'ixltx csmtx'=. cutlatexidx htm
```

```
if. #ixltx do.
```



```

    'latex fragment marker in .html' assert -. 1 e. lfrg E. htm
    htm=. ; (<lfrg) ixltx} csltx
end.

```

NB. insert missing paragraph tags - will wreck <pre> blocks

```
htm=. htmlParagraphs htm
```

NB. restore hidden <pre> blocks - not LF delimited paragraphs

```

if. #ixpre do.
    'ixltxpre csltxpre'=. pblk cutstridx htm
    '<pre> block source fragment count mismatch' assert (#ixltxpre) = #ixpre
    htm=. ;(ixpre{cspre) ixltxpre} csltxpre
end.

```

NB. convert html to latex

```

htmfile=. (t slash jpathsep texdir),TFWTEMPHTML
(utf8 htm) write htmfile
tex=. pandoc htmfile
cleartemps htmfile

```

NB. download any new referenced graphics

```
gdm=. getNewgraphics tex
```

NB. insert any latex \$latex ... \$ math fragments

```

if. #ixltx do.
    'ixltxrp csltxrp'=. lfrg cutstridx tex
    'latex math fragment count mismatch' assert (#ixltx) = #ixltxrp

```

```

ltx=. ixltx{csltx
NB. reset special HTML characters in LaTeX
ltx=. HTMLREPS&changestr&.> ltx
NB. drop leading $latex and replace with $
ltx=. '$' ,&.> (#'$latex ') }.&.> ltx
tex=. ;ltx ixltxrp} csltxrp
end.

NB. insert \lstlisting versions of source code blocks
if. #ixsrc do.
  'ixsrcrp cssrcrp'=. sblk cutstridx tex
  'source code block count mismatch' assert (#ixsrc) = #ixsrcrp
  pidlbls=. (<pid,'X') ,&.> " :&.> <"0 i.#ixsrc
  src=. pidlbls lstFrsrcb&.> ixsrc{cssrc
  tex=. ;src ixsrcrp} cssrcrp
end.

NB. reduce \includegraphics urls to downloaded image file names
'ixgx csgx'=. cutincludegraphicsidx tex
if. #ixgx do.
  gtxt=. '?'&beforestr&.> (}')&beforestr)@('/'&afterlaststr)&.> ixgx{csgx
  if. 1 e. gmsk=. 0 < #&.> gtxt do.
    gfiles=. gmsk#gtxt
    gtxt=. (<' \includegraphics{'') ,&.> gfiles ,&.> '}'
    ixgx=. gmsk#ixgx
    pidlbls=. (<pid,'X') ,&.> " :&.> <"0 i.#ixgx
    fig=. ((<'|~~~IMGRAPHICFILE~~~|') ,&.> gfiles) changestr&.> <TEXWRAPFIGURE

```

```
fig=. ((<'|~~~IMLABEL~~~|') ,&.> pidlbls) changestr&.> fig
tex=. ;gtxt ixgx} csgx
NB. append commented out figure templates manual edits
NB. will be required to tune the placement and size of graphics
tex=. tex , ;LF ,&.> fig
end.
end.

NB. comment out any residual text pandoc did not convert
'ixnp csnp'=. ('{[]}' ; '{[]}') cutnestidx tex
if. #ixnp do.
  nptx=. (LF, ' ', CR, ' ')&charsub&.> ixnp{csnp
  tex=. ;((<LF, '%') ,&.> nptx ,&.> LF) ixnp} csnp
end.

NB. prefix post title
tex=. (postTitleDate ptitle;pdate;url),tex
tex=. tex,LF,tabwarn

'%\input{' ,texpreamble, '}', (2#LF), tex, (2#LF), '%\end{document}'
)

tfwTitles=: 3 : 0

NB.*tfwTitles v-- LaTeX titles from WordPress XML titles.
NB.
NB. WordPress XML title text may contain numerous HTML special
NB. characters (see HTMLREPS) pandoc converts such characters to
```

```

NB. LaTeX equivalents.
NB.
NB. monad: blclTeXTitles =. tfwTitles blclHtmlTitles
NB.
NB.   wpxml=. read 'c:/pd/blog/wordpress/analyzethedatanotthedriviel.wordpress.xml '
NB.   posts=. ptableFrpwpxml wpxml
NB.   tfwTitles 1 {"1 posts
NB.
NB. dyad: blclTeXTitles =. clDirectory tfwTitles blclHtmlTitles

TEXFRWPDIR tfwTitles y
:
texdir=. x
btitle=. utf8 BEGINTITLE
'title marker occurs in title text' assert -. 1 e. btitle E. ;y
tempfile=. texdir,TFWTEMPHTML
(utf8 toHOST ;(<btitle,' ') ,&.> y ,&.> <2#LF) write tempfile
tex=. pandoc tempfile
cleartemps tempfile
tex=. (LF,' ') charsub tex -. CR
allwhitetrims.> (#btitle) }.&.> (btitle E. tex) <;.1 tex
)

timestamp=: 3 : 0

NB.*timestamp v-- formats timestamp as dd mmm yyyy hr:mn:sc
NB.
NB. monad: cl =. timestamp zu | nlTime

```

```

NB.
NB.   timestamp ''           NB. empty now
NB.   timestamp 2007 9 16    NB. fills missing
NB.   timestamp 1953 7 2 12 33

if. 0 = #y do. w=. 6!:0'' else. w=. y end.
r=. }: $ w
t=. 2 1 0 3 4 5 {"1 [ _6 [\ , 6 {"1 <. w
d=. '+++::' 2 6 11 14 17 }"1 [ 2 4 5 3 3 3 ": t
mth=. _3[\ '   JanFebMarAprMayJunJulAugSepOctNovDec'
d=. ,((1 {"1 t) {" mth) 3 4 5 }"1 d
d=. '0' (I. d=' ') } d
d=. ' ' (I. d='+') } d
(r,20) $ d
)

NB. appends trailing line feed character if necessary
tlf=: ] , ((10{a.)"_ = {:) }. (10{a.)"_

NB. append trailing / character if necessary
t slash=: ] , ('/'"_ = {:) }. '/'"_

NB. converts character strings to CRLF delimiter
toCRLF=: 2&}.@:;@:((13{a.)&,&.>@<;.1@((10{a.)&,)@toJ)

NB. converts character strings to host delimiter
toHOST=: toCRLF

```

```

NB. converts character strings to J delimiter LF
toJ=: ((10{a.) I.@(e.&(13{a.))@]}  ]>@:(#~ -.@((13 10{a.)&E.@,))

tolower=: 3 : 0

NB.*tolower v-- convert to lower case.
NB.
NB. monad: cl =. tolower cl

x=. I. 26 > n=. ((65+i.26){a.) i. t=. ,y
($y) $ ((x{n) { (97+i.26){a.) x)t
)

uedposts=: 3 : 0

NB.*uedposts v-- lists unedited post files.
NB.
NB. monad: uedposts uuIgnore
NB.
NB.    uedposts 0                      NB. unedited files
NB.    (postfiles -. uedposts) 0 NB. edited files

NB. depends on layout of root file: bm.tex
txt=. '%</blogposts>'&beforestr '%<blogposts>'&afterstr read TEXFRWPDIR,TEXROOTFILE
txt=. allwhitetrim&.> a: -.~ <;._2 tlf txt -. CR
txt=. ('}'&beforestr)@('\input{'&afterstr)&.> ('%' = {.&> txt)#txt
(<TEXFRWPDIR) ,&.> txt
)

```

NB. character list to UTF-8

utf8=: 8&u:

NB. to windows \ character in paths

winpathsep=: '\&(('/' I.@:=])})

NB. writes a list of bytes to file

write=: 1!:2]`<@.(32&>@ (3!:0))

NB.POST_TeXfrWpxml TeXfrWpxml post processor

smoutput IFACE=: (0 : 0)

NB. (TeXfrWpxml) interface word(s): 20230410j143000

NB. -----

NB. BlogHashes NB. update blog hashes

NB. FixBaddown NB. attempt to convert *.baddown files to *.markddown

NB. LatexFrWordpress NB. experimental conversion of Wordpress XML to LaTeX

NB. MainMarkdown NB. assembles *.markdown files in EPUB directory in a master file

NB. MarkdownFrLatex NB. converts edited LaTeX post files to image free markdown

NB. countYearposts NB. adds post counts to table of content lines

)

SetTeXfrWpxmlPaths 0

cocurrent 'base'

coinsert 'TeXfrWpxml'

Index

afterlaststr, 21
afterstr, 21
allwhitetrim, 22
assert, 22
attrvalue, 22

BADDOWNEXT, 9
beforelaststr, 22
beforestr, 22
BEGINTITLE, 9
BESOURCEDELS, 9
BESOURCEPREDELS, 10
BlogHashes, 12
blogimgs, 22
boxopen, 23

cdatatext, 23
changestr, 23
charsub, 24
cleartemps, 25
countYearposts, 25
CR, 10
CRLF, 10
ctl, 26
cutincludegraphicsidx, 26
cutlatexidx, 27
cutnestidx, 27

cutpxtidx, 29
cutstridx, 30

EPUBAMBLE, 6
EPUBFILE, 6, 21
EPUBFRWPDIR, 6, 21

fboxname, 30
ferase, 30
fexist, 31
filenamesFrtid, 31
FILETITLELEN, 10
firstones, 32
FixBaddown, 13
fsize, 32

getNewgraphics, 32

HTMLEXT, 10
htmlParagraphs, 33
HTMLREPS, 6

IFACE, 55
IFACEWORDSTeXfrWpxml, 10
inputposts, 34

jpathsep, 35
justdrv, 35

justext, 35
justfile, 35
justpath, 35

LATEXFIGURETEMPLATES, 8
LATEXFRAGMARK, 10
LatexFrWordpress, 14
LF, 10
lstFrsrcb, 35
LSTLISTINGEND, 7
LSTLISTINGHDR, 6

MainMarkdown, 16
MARKDOWNEXT, 11
MARKDOWNFILE, 7
MarkdownFrLatex, 18
mdfootnotes, 36
MDSECTIONPFX, 11

pandoc, 37
PANDOC CMD, 11
postfiles, 39
postid, 40
posts, 40
posttex, 41
postTitleDate, 39
prunePtable, 41

ptableFrwpxml, 42
read, 43
rmLatexGraphics, 43
ROOTWORDSTeXfrWpxml, 11
SetTeXfrWpxmlPaths, 20
sha256, 44
sha256dir, 45
showpass, 45
smoutput, 45
sortonid, 45
sortonpublishdate, 46
sortposts, 46

SOURCEBLOCKMARK, 11
SOURCEPREMARK, 11
TEMPTEXTFILE, 11
TEXEXT, 11
texFrhtml, 47
TEXFRWPDIR, 7, 21
TEXINCLUSIONS, 7, 21
TEXPREAMBLE, 7, 21
TEXROOTFILE, 7, 21
TEXSECTIONTITLE, 7
TEXWRAPFIGURE, 8
TFWTEMPHTML, 12
tfwTitles, 51

timestamp, 52
tlf, 53
tllslash, 53
toCRLF, 53
toHOST, 53
toJ, 54
tolower, 54
uedposts, 54
utf8, 55
WGETCMD, 12
winpathsep, 55
write, 55