FACULTY
OF INFORMATICS
Masaryk University

A M↓ Interpreter for TEX

Project MUNI 33 / 12 2015

https://github.com/witiko/markdown

Vít Novotný

Contents

1. Introduction
   - 1.1 The Case for Lightweight Markup
   - 1.2 Existing Solutions
2. The Markdown Package
   - 2.1 Building a Parser
   - 2.2 Using Markdown from Within LaTeX
3. Conclusion
4. Bibliography

Section 1

Introduction

The Case for Lightweight Markup

What is Wrong with TEX?

1. High Markup to Text Ratio
   - Knuth (1986) is 22 % markup.
   - Downey et al. (2016) is 21 % markup.
2. Zero Sandboxing Support
   - The document you are typesetting may not compile.

     ```
     … a file named \texttt{evil_underscores.tex} …
     ```

   - The document you are typesetting may halt.

     ```
     \def\whiletrue{\whiletrue} \whiletrue
     ```

   - The document you are typesetting may access the system shell.

     ```
     \immediate\write18{sudo rm -rf /}
     ```

3. Steep Learning Curve

The Case for Lightweight Markup

Comparison of LaTeX and Markdown

```
\section{This is a level one heading}
This is a text paragraph with \emph{emphasis}.
\begin{quotation}This paragraph will show as a quote.\end{quotation}
\begin{verbatim}
This is is a source code example.
\end{verbatim}
\begin{itemize}
  \item First item with \alert{strong emphasis}
  \item Second item with a link%
    \footnote{See \url{http://link.com} (Title)}
\end{itemize}
\begin{enumerate}
  \item First item with \verb‘inline code‘.
  \item Second item with an \includegraphics{image.png}
\end{enumerate}
```

The Case for Lightweight Markup

Comparison of LaTeX and Markdown

```
# This is a level one heading
This is a text paragraph with _emphasis_.
> This paragraph will show as a quote.

␣␣␣␣This is is a source code example.

* First item with **strong emphasis**
* Second item with a [link](http://link.com/ "Title")

1. First item with `inline code`.
2. Second item with an ![image](image.png "Title")
```

The Case for Lightweight Markup

How is Markdown Useful?

1. Minimal Markup to Text Ratio
   - Recall: Knuth (1986) and Downey et al. (2016) are ~22 % markup.
   - Gillespie et al. (2016) is 5.5 % markup.
   - Grolemund et al. (2016) is 3.8 % markup.
2. Sandboxing Support
   - A Markdown document converted to TEX will always compile.
   - The document may neither halt nor access the shell.
3. Hybrid Markup Support
   - Markdown was designed to supplement HTML, not replace it.
   - Structurally simple sections can use pure Markdown, complex sections may combine Markdown and the host markup.
4. Mild Learning Curve

Existing Solutions

The Swiss Army Knife of Pandoc

> *If you need to convert files from one markup format into another, Pandoc is your swiss-army knife.*
>
> — MacFarlane (2016b), emphasis mine

- A multi-target publishing software.
- Supports tens of markup languages (Markdown, LaTeX, HTML, XML Docbook) and output formats (ODF, OOXML, PDF).
- The use of Pandoc for the preparation of LaTeX documents has been described by Dominici (2014).

Existing Solutions

What is Wrong with Pandoc?

1. No Way to Change Output Markup

```
# Heading {#link}
This is [a link](#link).
```

↓ ↓

```
\hypertarget{link}{\section{Heading}\label{link}}
This is \protect\hyperlink{link}{a link}.
```

2. Not a Part of TeX Distributions
   - Markdown documents cannot be directly edited at collaborative TeX platforms such as ShareLaTeX or Overleaf.

Existing Solutions

What is Wrong with Pandoc?

3. Half-hybrid, Half-sandboxed
   - The input is heuristically parsed and sanitized:

```
This {will} 2^n \begin{get} s~nitized and \this{will}
not \begin{equation}2^n\end{equation} $2^n$.
```

↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓

```
This \{will\} 2\^{}n \textbackslash{}begin\{get\}
s\textasciitilde{}nitized and \this{will} not
\begin{equation}2^n\end{equation} \(2^n\).
```

   - Malicious input such as

```
\def\shell{18} \immediate\write\shell{sudo rm -rf /}
```

is left alone by Pandoc.

Section 2

The Markdown Package

Building a Parser

Is TEX Up to the Task?

There exist formal language parsers written solely in TEX. These parsers recognize regular (LATEX3 Project, 2016) and context-free LL(1) languages (Carlisle, 2000). Markdown is not context-free:

“There is a literal backtick (‘) here.“

and a parser needs to be able to backtrack over the entire input:

```
[this is not a link](http://link.com/ ”Title”
```

Implementing a recursive-descent parser with backtracking in TEX is possible, but generally a bad idea:

- Difficult to Maintain, Highly Unidiomatic
- Lack of Efficient Data Structures

Building a Parser

Can We Use Lua Instead of TEX?

> *Lua is a powerful, efficient, lightweight, embeddable scripting language. It supports procedural programming, object-oriented programming, functional programming, data-driven programming, and data description.*
>
> — Lua Team (2016)

> *LuaTEX is an extended version of pdfTEX using Lua as an embedded scripting language.*
>
> — LuaTEX Team (2016)

Building a Parser

Can We Use Lua Instead of TEX?

- With LuaTEX, we can directly execute Lua code:

```
1 + 2 = \directlua{ tex.sprint(1 + 2) }
```

- With pdfTEX and other modern TEX engines, we can spawn a shell and execute the Lua code in a separate process:

```
1 + 2 = \newwrite\script
\immediate\openout\script=script.lua
\immediate\write\script{ print(1 + 2) }%
\immediate\closeout\script
\immediate\write18{texlua script.lua > output.tex}%
\input output.tex
```

Building a Parser

The Lunamark Library

- Lunamark (MacFarlane, 2016a) is a Markdown parser for Lua.
- The language is specified using a Parsing Expression Grammar (PEG) via the LPeg C library (with some cheating).
  - PEGs are CFGs with ordered choice; as a corrolary, any PEG is unambiguous. (Ford, 2004) An input word u can be parsed according to a PEG G in linear time relative to the size of u. (Ford, 2002) It is conjectured that not all CFLs are recognized by a PEG.
  - The dependencies of Lunamark were all either compiled into LuaTEX (LPeg, Slnunicode), or unneeded (Cosmo, Alt-getopt).
- The library has been released under the Expat (MIT) License.

Building a Parser

The Lunamark Library

The modified version of Lunamark:

- produces a parse tree rather than presentation markup:

```
# Heading
This is [a link](#link).
```

↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓

```
\markdownRendererHeadingOne{Heading}
This is \markdownRendererLink{a link}{#link}{#link}{}.
```

The Markdown TeX package:

- converts a Markdown document to the parse tree via Lunamark,
- defines the macros and typesets the parse tree using TeX.

Using Markdown from Within LaTeX

The Sandbox and Hybrid Modes

```
\documentclass{article}
\usepackage{markdown}
\begin{document}
\begin{markdown}
  Foo bar \TeX{} $2^n$.
\end{markdown}
\begin{markdown*}{hybrid}
  Foo bar \TeX{} $2^n$.
\end{markdown*}
\end{document}
```

Foo bar \TeX{} 2^n. Foo bar TeX 2^n.

Using Markdown from Within LaTeX

Mapping Markdown Tokens to TeX Macros

```
\documentclass{article}
\usepackage{markdown}
\markdownSetup{renderers = {
  link = {#1\footnote{See \url{#3} (#4)}},
}}
\begin{document}
\begin{markdown}
  Foo [bar](http://link.com "Link").
\end{markdown}
\end{document}
```

Foo bar[1].

[1] See http://link.com (Link)

Using Markdown from Within LaTeX

Syntax Extensions

- Some syntax extensions were already supported by Lunamark:
  - footnotes,
  - definition lists,
- New syntax extensions were added as a part of the project:
  - citations,
  - fenced code blocks.

Using Markdown from Within LaTeX

Syntax Extensions – `\markdownSetup{footnotes}`

```
Here is a footnote reference,[^1] and another.[^long]

[^1]: Here is the footnote.

[^long]: Here's one with multiple blocks.

    Subsequent paragraphs are indented to show that
they belong to the footnote.
```

Here is a footnote reference,[2] and another.[3]

[2] Here is the footnote.

[3] Here's one with multiple paragraphs.

Subsequent paragraphs are indented to show that they belong to the footnote.

Using Markdown from Within LaTeX

Syntax Extensions – `\markdownSetup{definitionLists}`

```
Term 1
:   Definition

Term 2
:   Definition with

    multiple paragraphs
```

Term 1 Definition 1

Term 2 Definition
with multiple paragraphs

Using Markdown from Within LaTeX

Syntax Extensions – `\markdownSetup{citations}`

```
Here is a parenthetical citation [@knuth86] and
a string of several [see @knuth86, pp. 33-35;
also @gruber04, chap. 1].

Here is a text citation @knuth86 and a string of
several @knuth86 [pp. 33-35; @gruber04, chap. 1].
```

Here is a parenthetical citation (Knuth, 1986) and a string of several (see Knuth, 1986, pp. 33-35; also Gruber, 2004, chap. 1).
Here is a text citation Knuth (1986) and a string of several Knuth (1986, pp. 33-35) and Gruber (2004, chap. 1).

Using Markdown from Within LaTeX

Syntax Extensions – \markdownSetup{fencedCode}

```
~~~ js
if (a > b)
  return c + 4;
else
  return d + 5;
~~~~~~
```

```js
if (a > b)
  return c + 4;
else
  return d + 5;
```

Section 3

Conclusion

Conclusion

The Missing Pieces of the Puzzle

- Apart from the LaTeX interface, the package also exposes Lua, plain TeX and ConTeXt interfaces.
- The package includes 82 pages of user and technical documentation. (Novotný, 2016a)
- A section on writing LaTeX documents in Markdown was added to the fithesis3 sample documents.
- The package was released under the LaTeX Project Public License (LPPL) 1.3 on the Comprehensive TeX Archive Network (CTAN), GitHub, and the faculty GitLab. (Novotný, 2016c) It is available in updated TeX Live 2016.

Conclusion

Reception by the Community

- The syntax extensions were backported to Lunamark and merged by MacFarlane, resulting in a new minor version release of the library (0.5.0). (Novotný, 2016b)
- The package was featured on the twitter profile of Overleaf – a major online service for preparing LaTeX documents – along with original example documents. (Overleaf, 2016)
- The package was reviewed in the bulletin of the German TeX Users Group (DANTE e.V.). (Fenn, 2016, pp. 43)
- An article about the package has been accepted for publication in the bulletin of the Czechoslovak TeX Users Group (CSTUG).

Section 4

Bibliography

Bibliography I

LaTeX3 PROJECT, 2016. *The l3regex package: regular expressions in TeX* [online] [visited on 2016-11-08]. Available from: http://mirrors.ctan.org/macros/latex/contrib/l3experimental/l3regex.pdf.

CARLISLE, David, 2000. XMLTeX: A non-validating (and not 100% conforming) namespace-aware XML parser implemented in TeX. *TUGboat* [online]. Vol. 21, no. 3, pp. 193–199 [visited on 2016-11-08]. ISSN 0896-3207. Available from: https://www.tug.org/TUGboat/tb21-3/tb68carl.pdf.

Bibliography II

DOMINICI, Massimiliano, 2014. An overview of Pandoc. *TUGboat* [online]. Vol. 35, no. 1, pp. 44–50 [visited on 2016-08-15]. ISSN 0896-3207. Available from: http://tug.org/TUGboat/tb35-1/tb109dominici.pdf.

DOWNEY, Allen B.; MAYFIELD, Chris, 2016. *Think Java: How to Think Like a Computer Scientist* [online]. Green Tea Press. Version 6.1.0 [visited on 2016-11-08]. Available from: http://thinkjava.org/.

FENN, Jürgen, 2016. Neue Pakete auf CTAN. *Die TEXnische Komödie*. No. 3/2016. ISSN 1434-5897.

Bibliography III

FORD, Bryan, 2002. Packrat Parsing: Simple, powerful, lazy, linear time, functional pearl. In: *Packrat Parsing: Simple, powerful, lazy, linear time, functional pearl*. *ACM SIGPLAN Notices* [online]. Vol. 37, pp. 36–47 [visited on 2016-11-08]. No. 9. Available from DOI: 10.1145/581478.581483.

FORD, Bryan, 2004. Parsing expression grammars: A recognition-based syntactic foundation. In: *Parsing expression grammars: A recognition-based syntactic foundation*. *ACM SIGPLAN Notices* [online]. Vol. 39, pp. 111–122 [visited on 2016-08-16]. No. 1. Available from DOI: 10.1145/964001.964011.

Bibliography IV

GILLESPIE, Colin; LOVELACE, Robin, 2016. *Efficient R programming* [online]. O'Reilly Media [visited on 2016-11-08]. ISBN 978-1-4919-5078-4. Available from: https://github.com/hadley/r4ds/.

GROLEMUND, Garrett; WICKHAM, Hadley, 2016. *R for Data Science* [online]. O'Reilly Media [visited on 2016-11-08]. ISBN 978-1-4919-1039-9. Available from: https://github.com/hadley/r4ds/.

GRUBER, John, 2004. *Markdown* [online] [visited on 2016-08-15]. Available from: https://daringfireball.net/projects/markdown/.

Bibliography V

KNUTH, Donald Ervin, 1986. *The TEXbook* [online]. 3rd ed. Addison-Westley [visited on 2016-11-08]. ISBN 0-201-13447-0. Available from: https://mirrors.ctan.org/systems/knuth/dist/tex/texbook.tex.

LUA TEAM, 2016. *Lua: About* [online] [visited on 2016-08-15]. Available from: https://www.lua.org/about.html.

LUATEX TEAM, 2016. *LuaTEX: Welcome* [online] [visited on 2016-08-15]. Available from: http://luatex.org/.

MACFARLANE, John, 2016a. *Lunamark: Lua library for conversion between markup formats* [online] [visited on 2016-08-15]. Available from: https://github.com/jgm/lunamark.

Bibliography VI

MACFARLANE, John, 2016b. *Pandoc: a universal document converter* [online] [visited on 2016-08-15]. Available from: http://pandoc.org/.

NOVOTNÝ, Vít, 2016a. *A Markdown Interpreter for TEX* [online] [visited on 2016-08-17]. Available from: http://mirrors.ctan.org/macros/generic/markdown/markdown.pdf.

NOVOTNÝ, Vít, 2016b. *Added support for Pandoc-style citations* [online] [visited on 2016-08-15]. Available from: https://github.com/jgm/lunamark/pull/20.

Bibliography VII

NOVOTNÝ, Vít, 2016c. *Markdown: A package for converting and rendering markdown documents inside TEX* [online] [visited on 2016-08-15] Available from: http://ctan.org/pkg/markdown, https://github.com/Witiko/markdown, and https://gitlab.fi.muni.cz/xnovot32/markdown.

OVERLEAF, 2016. *Two great examples of how to use #markdown with @Overleaf – thanks @liantze!* [online] [visited on 2016-08-15]. Available from: https://twitter.com/overleaf/status/763395560682364928.