


Práce s jazykem  v rámci
T_EXových dokumentů
<https://github.com/witiko/markdown>

Vít Novotný

Fakulta informatiky Masarykovy Univerzity, Brno

17. prosince 2016

1 Úvod

- Proč používat odlehčené značkování
- Existující softwarová řešení

2 Balíček `markdown.tex`

- Návrh parseru jazyka Markdown
- Použití balíku `markdown.tex` v \LaTeX u

3 Reference

Section 1

Úvod

Proč používat odlehčené značkování

Cožpak mi T_EX nepřijde dost dobrý?

1. Vysoký poměr značkování vůči textu

- Knuth (1986) je z 22 % značkování.
- Downey et al. (2016) je z 21 % značkování.

2. Nulové odstínění uživatele od vnitřností T_EXu

- Sázený dokument se nemusí zkompilovat.

... v souboru `{\tt zpropadená_podtržítka.tex}` ...

- Sázený dokument může skončit v nekonečné smyčce.

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

- Sázený dokument může přistupovat k příkazové řádce systému.

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

3. Dlouhá křivka učení

Proč používat odlehčené značkování

Srovnání \LaTeX s jazykem Markdown

```
\section{Nadpis první úrovně}
\subsection{Nadpis druhé úrovně}
Odstavec se \emph{zvýrazněným textem}
\begin{quotation}
  Citát
\end{quotation}
\begin{verbatim}
  Ukázka zdrojového kódu
\end{verbatim}
\begin{itemize}
  \item Bod seznamu se \alert{silně zvýrazněným textem}
  \item Bod seznamu s odkazem%
    \footnote{Vizte \url{https://odkaz.cz} (Titulek)}
\end{itemize}
\begin{enumerate}
  \item Bod seznamu s \verb`ukázkou zdrojového kódu`
  \item Bod seznamu s obrázkem \includegraphics{obrázek.svg}
\end{enumerate}
```

Proč používat odlehčené značkování

Srovnání \LaTeX s jazykem Markdown

Nadpis první úrovně

Nadpis druhé úrovně

Odstavec se *zvýrazněným textem*

> Citát

~~~~~Ukázka zdrojového kódu.

\* Bod seznamu se **silně zvýrazněným textem**

\* Bod seznamu s [odkazem](https://odkaz.cz/ "Titulek")

1. Bod seznamu s `ukázkou zdrojového kódu`

2. Bod seznamu s ![obrázkem](obrázek.svg "Titulek")

# Proč používat odlehčené značkování

## Jazyk Markdown

*The overriding design goal for Markdown's formatting syntax is to make it **as readable as possible**. The idea is that a Markdown-formatted document should be **publishable as-is, as plain text**, without looking like it's been marked up with tags or formatting instructions. While Markdown's syntax has been influenced by several existing text-to-HTML filters, the single biggest source of inspiration for Markdown's syntax is **the format of plain text email**.*

— Gruber (2004)

- Původní jazyk byl navržen pro přípravu HTML dokumentů.
- Dnes existuje množství nástrojů (Pandoc, MultiMarkdown), které umožňují použití jazyka s jinými jazyky ( $\text{\LaTeX}$ ).

# Proč používat odlehčené značkování

## Výhody jazyka Markdown

### 1. Minimální poměr značkování vůči textu

- Knuth (1986) a Downey et al. (2016) jsou z ~22 % značkování.
- Gillespie et al. (2016) je z 5.5 % značkování.
- Grolemond et al. (2016) je z 3.8 % značkování.

### 2. Úplné odstínění od vnitřností T<sub>E</sub>Xu ...

- Markdownový dokument se po převodu do T<sub>E</sub>Xu zkompiluje.
- Dokument nemůže cyklit ani přistupovat k příkazové řádce.

### 3. ... nebo hybridní značkování

- Markdown byl navržen jako doplněk HTML, ne jako náhrada.
- Strukturně jednoduché sekce mohou být značkovány pouze Markdownem, složité pak pomocí T<sub>E</sub>Xových maker.

### 4. Krátká křivka učení



# Existující softwarová řešení

Švýcarský armádní nůž Pandoc

*If you need to convert files from one markup format into another, Pandoc is your swiss-army knife.*

— MacFarlane (2016b)

- Nástroj pro konverze a víceformátové publikování.
- Podporuje převod mezi desítkami značkovacích jazyků (Markdown,  $\text{\LaTeX}$ , HTML, XML Docbook) a výstupních formátů (ODF, OOXML, PDF).
- Přípravu  $\text{\LaTeX}$ ových dokumentů pomocí Pandocu popisuje Dominici (2014).

# Existující softwarová řešení

Cožpak mi ani Pandoc nepřijde dost dobrý?

1. Nelze ovlivnit výstupní značkování.

```
# Nadpis {#nadpis}
```

```
[Jeden]({#nadpis}) odkaz a [druhý](https://odkaz.cz/).
```

```
↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓↓
```

```
\hypertarget{nadpis}{\section{Nadpis}\label{nadpis}}
```

```
\protect\hyperlink{nadpis}{Jeden} odkaz a
```

```
\href{https://odkaz.cz/}{druhý}.
```

2. Nejedná se o součást T<sub>E</sub>Xových distribucí.

- Markdownové dokumenty nelze přímo editovat na kolaborativních T<sub>E</sub>Xových platformách (ShareL<sub>A</sub>T<sub>E</sub>X, Overleaf).
- Konstatní verze T<sub>E</sub>Xové distribuce negarantuje stabilní výstup.

## Existující softwarová řešení

## Cožpak mi ani Pandoc nepřijde dost dobrý?

### 3. Částečné odstínění od T<sub>F</sub>Xu, částečně hybridní značkování

- Vstupní dokument je po načtení heuristicky očištěn od speciálních znaků plain T<sub>E</sub>Xu:

Toto {vše}  $2^n$  \begin{bude} očištěno, ale \toto{už}  
nikoliv \begin{equation} 2^n \end{equation}  $2^n$ .

Toto  $2^n$  \textbackslash{}begin\{bude\}  
o\textasciitilde{}čištěno, ale \toto{už} nikoliv  
\begin{equation}2^n\end{equation}  $(2^n)$ .

- Nebezpečné vstupy jako

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

jsou Pandocem ponechány beze změny.

## Section 2

Balíček markdown.tex

# Návrh parseru jazyka Markdown

Je T<sub>E</sub>X vhodným jazykem?

Existují parsery formálních jazyků napsané v T<sub>E</sub>Xu. Tyto parsery rozpoznávají regulární (L<sup>A</sup>T<sub>E</sub>X3 Project, 2016) a bezkontextové LL(1) jazyky (Carlisle, 2000). Markdown však není bezkontextový

• Obrácené apostrofy (``) lze psát i v ukázkách kódu. •  
a parser se v degenerovaných případech musí vracet přes celý vstup:

```
[toto není dýmka] (https://dymka.cz/ "Dýmka"
```

Implementace v T<sub>E</sub>Xu tedy je možná, ale díky absenci vhodných datových struktur pro posun po vstupním řetězci bez kopírování také vysoce neefektivní.

# Návrh parseru jazyka Markdown

Co tedy použít namísto T<sub>E</sub>Xu?

*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)

*LuaT<sub>E</sub>X is an extended version of pdfT<sub>E</sub>X using Lua as an embedded scripting language.*

— LuaT<sub>E</sub>X Team (2016)

- Interpretér Lua je dostupný všude tam, kde se nachází LuaT<sub>E</sub>X.
- Pomocí malého triku jej lze spustit i z pdfT<sub>E</sub>Xu a X<sub>Y</sub>T<sub>E</sub>Xu.

# Návrh parseru jazyka Markdown

Co tedy použít namísto T<sub>E</sub>Xu?

- V LuaT<sub>E</sub>Xu můžeme přímo spouštět kód v jazyce Lua:

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

- V pdfT<sub>E</sub>Xu a X<sub>Ǝ</sub>T<sub>E</sub>Xu interpreter spustíme z příkazové řádky:

```
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
```

# Návrh parseru jazyka Markdown

Knihovna Lunamark

- Lunamark (MacFarlane, 2016a) je parser jazyka Markdown napsaný v jazyce Lua.
- Jazyk je specifikován pomocí formalismu Parsing Expression Grammar (PEG) (a s trochou podvádění) s využitím knihovny LPeg psané v jazyce C.
- Veškeré závislosti knihovny byly buďto zakompilované do Lua<sub>T</sub><sub>E</sub>Xu (LPeg, SInunicode), nebo nepotřebné (Cosmo, Alt-getopt).
- Knihovna byla vypuštěna pod licencí Expat (MIT).





# Použití balíku markdown.tex v L<sup>A</sup>T<sub>E</sub>Xu

## Odstíněný a hybridní režim

```
\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  $\backslash\mathrm{TeX}\{\}$   $2^n$ . Foo bar  $\mathrm{T}_{\mathrm{F}}\mathrm{X}$   $2^n$ .

# Použití balíku markdown.tex v L<sup>A</sup>T<sub>E</sub>Xu

# Mapování tokenů jazyka Markdown na T<sub>E</sub>Xová makra

```
\documentclass{article}
\usepackage{markdown}
\markdownSetup{renderers = {
  link = {#1\footnote{Vizte \url{#3} (#4)}}},
}}
\begin{document}
\begin{markdown}
  Foo [bar](https://odkaz.cz/ "Titulek").
\end{markdown}
\end{document}
```

Foo bar<sup>1</sup>.

<sup>1</sup>Vizte <https://odkaz.cz/> (Titulek)

# Použití balíku `markdown.tex` v $\text{\LaTeX}$ u

## Rozšíření syntaxe

- Jazyk Markdown má poměrně chudou syntax. Existuje proto množství syntaktických rozšíření, z nichž některá byla podporována již knihovnou Lunamark:
  - poznámky pod čarou,
  - definiční seznamy,
- V rámci projektu byla přidána podpora pro následující syntaktická rozšíření:
  - citace,
  - alternativní syntax pro poznámky pod čarou,
  - oplocené ukázky zdrojového kódu.




# Použití balíku markdown.tex v L<sup>A</sup>T<sub>E</sub>Xu

## Syntaktická rozšíření – \markdownSetup{inlineFootnotes}

Toto je alternativní syntax poznámek pod čarou.<sup>~</sup>[Tyto poznámky pod čarou se snáze píší, protože není třeba vybírat identifikátor a přesouvat se s kurzorem pod odstavec.]

Toto je alternativní syntax poznámek pod čarou.<sup>4</sup>

<sup>4</sup>Tyto poznámky pod čarou se snáze píší, protože není třeba vybírat identifikátor a přesouvat se s kurzorem pod odstavce. 



## Syntaktická rozšíření – \markdownSetup{citations}

Zde je textová citace @knuth86 a zde je řetězec  
několika @knuth86 [s. 33--35; @gruber04, kap. 1].

Zde je citace s použitím závorek (Knuth, 1986) a zde je řetězec několika (viz Knuth, 1986, s. 33–35; a také Gruber, 2004, kap. 1). Zde je textová citace Knuth (1986) a zde je řetězec několika Knuth (1986, s. 33–35) a Gruber (2004, kap. 1).



# Použití balíku markdown.tex v L<sup>A</sup>T<sub>E</sub>Xu

## Syntaktická rozšíření – \markdownSetup{fencedCode}

```

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

```

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

## Section 3

### Reference



L<sup>A</sup>T<sub>E</sub>X3 PROJECT, 2016. *The l3regex package: regular expressions in T<sub>E</sub>X* [online] [cit. 2016-11-08]. Dostupné z: <http://mirrors.ctan.org/macros/latex/contrib/l3experimental/l3regex.pdf>.



CARLISLE, David, 2000. XMLT<sub>E</sub>X: A non-validating (and not 100% conforming) namespace-aware XML parser implemented in T<sub>E</sub>X. *TUGboat* [online]. Roč. 21, č. 3, s. 193–199 [cit. 2016-11-08]. ISSN 0896-3207. Dostupné z: <https://www.tug.org/TUGboat/tb21-3/tb68carl.pdf>.



DOMINICI, Massimiliano, 2014. An overview of Pandoc. *TUGboat* [online]. Roč. 35, č. 1, s. 44–50 [cit. 2016-08-15]. ISSN 0896-3207. Dostupné z: <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. Verze 6.1.0 [cit. 2016-11-08]. Dostupné z: <http://thinkjava.org/>.



FENN, Jürgen, 2016. Neue Pakete auf CTAN. *Die T<sub>E</sub>Xnische Komödie*. Č. 3/2016. ISSN 1434-5897.



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]. Sv. 37, s. 36–47 [cit. 2016-11-08]. Č. 9. Dostupné z DOI: 10.1145/581478.581483.

# Reference III



FORD, Bryan, 2004. Parsing expression grammars: A recognition-based syntactic foundation. In: *Parsing expression grammars: A recognition-based syntactic foundation*. ACM SIGPLAN Notices [online]. Sv. 39, s. 111–122 [cit. 2016-08-16]. Č. 1. Dostupné z DOI: 10.1145/964001.964011.



GILLESPIE, Colin; LOVELACE, Robin, 2016. *Efficient R programming* [online]. O'Reilly Media [cit. 2016-11-08]. ISBN 978-1-4919-5078-4. Dostupné z: <https://github.com/hadley/r4ds/>.



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

# Reference IV

-  GRUBER, John, 2004. *Markdown* [online] [cit. 2016-08-15].  
Dostupné z:  
<https://daringfireball.net/projects/markdown/>.
-  KNUTH, Donald Ervin, 1986. *The T<sub>E</sub>Xbook* [online]. 3. vyd.  
Addison-Westley [cit. 2016-11-08]. ISBN 0-201-13447-0.  
Dostupné z: <https://mirrors.ctan.org/systems/knuth/dist/tex/texbook.tex>.
-  LUA TEAM, 2016. *Lua: About* [online] [cit. 2016-08-15].  
Dostupné z: <https://www.lua.org/about.html>.
-  LUAT<sub>E</sub>X TEAM, 2016. *LuaT<sub>E</sub>X: Welcome* [online] [cit. 2016-08-15]. Dostupné z: <http://luatex.org/>.
-  MACFARLANE, John, 2016a. *Lunamark: Lua library for conversion between markup formats* [online] [cit. 2016-08-15]. Dostupné z: <https://github.com/jgm/lunamark>.

# Reference V



MACFARLANE, John, 2016b. *Pandoc: a universal document converter* [online] [cit. 2016-08-15]. Dostupné z: <http://pandoc.org/>.



NOVOTNÝ, Vít, 2016a. *A Markdown Interpreter for T<sub>E</sub>X* [online] [cit. 2016-08-17]. Dostupné z: <http://mirrors.ctan.org/macros/generic/markdown/markdown.pdf>.



NOVOTNÝ, Vít, 2016b. *Added support for Pandoc-style citations* [online] [cit. 2016-08-15]. Dostupné z: <https://github.com/jgm/lunamark/pull/20>.



NOVOTNÝ, Vít, 2016c. *Markdown: A package for converting and rendering markdown documents inside T<sub>E</sub>X* [online] [cit. 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] [cit. 2016-08-15].

Dostupné z: [https:](https://twitter.com/overleaf/status/763395560682364928)

[//twitter.com/overleaf/status/763395560682364928](https://twitter.com/overleaf/status/763395560682364928).