MlBibTeX now handles Unicode*

Jean-Michel HUFFLEN

Abstract

A new version of MlBibTeX can deal with the full range of Unicode and can process .bib files written using most byte-based encodings. We describe the new organisation of this version and show how to use the executable files built by the installation procedure. We also summarize the syntactic extensions implemented within .bib files, some originating from new fields introduced by the biblatex package.

Keywords MlBibTeX, kernel and derived programs, interface with Scheme, recognised formats and encodings, output routines, biblatex package, ConTeXt.

Streszczenie

Nowa wersja MlBibTeX-a radzi już sobie z unikodem w pełnym zakresie i potrafi przetwarzać pliki .bib zapisane z użyciem większości kodowań jednobajtowych. Zostanie opisana nowa organizacja tej wersji oraz sposób używania plików wykonywalnych, jakie buduje procedura instalacyjna. Zostaną zwięźle omówione rozszerzenia syntaktyczne zaimplementowane w plikach .bib, z których niektóre mają źródło w nowych polach pakietu biblatex.

Słowa kluczowe MlBibTeX, jądro i programy pochodne, interfejs do Scheme, rozpoznawane formaty i kodowania, procedury wyjściowe, biblatex paket, ConTeXt.

Introduction

Let us recall that the MlBibTeX[1] program aims to be a 'better' BibTeX, that is, a 'better' bibliography processor for documents written using LaTeX.

Since its beginning, this project has particularly focused on multilingual features. Then it has also provided better functions from a point of view related to programming. For example, the sort function used within BibTeX's bibliography styles [13] can only be customised by redefining *one* sort key, built by concatenating strings.[2] On the contrary, sort functions handled by MlBibTeX can be more easily adapted or redefined. Although MlBibTeX includes a rich collection of 'predefined' order relations, such a *modus operandi* means that users interested in *ad hoc* sort procedures are able to write functions in Scheme [14], the implementation language of MlBibTeX. That may be viewed as restrictive, but much synergy exists among LaTeX users, so we think that the advantages of this approach outweigh the drawbacks: programmers can help non-programmers. On another point, MlBibTeX went beyond exclusively generating LaTeX 'References' sections: it can also generate bibliographies according to other output formats, some examples being ConTeXt [1], XML[3]-like formats, or simple texts.

In [7], we recalled the successive steps of the development of MlBibTeX and announced a new version (1.4), more new features being described in [8]. This new version's main point is the ability to deal with the full range of the Unicode encoding and character standard [15]. So MlBibTeX is now able to process bibliography database (.bib) files encoded with conventions other than ASCII[4] and Latin 1, an extension suitable for western European languages. This new version will be publicly available in Summer 2017. Hereafter, after a short review of MlBibTeX's organisation (§1), we progressively describe this new version's state about the formats recognised (§2), the bibliography styles which may be used (§3), and the output routines for each output format (§4).

1 MlBibTeX's organisation

We detailed MlBibTeX's organisation in [9, Fig. 5]. Let us recall that this program gets information from an .aux file about *citation keys* and .bib files, and also looks into the preamble of a .tex master file for the languages used throughout a LaTeX document if the babel package is loaded. Parsing .bib files results in an (S)XML[5] tree. A *bibliography style* is applied to this tree, and *output routines* allow the result of such a style to conform to an output format's needs. For example, different output routines are called in order to build bibliographies for documents using LaTeX and ConTeXt, as explained in [9].

In [4] we explained that MlBibTeX is composed of a *kernel*, upon which *executable programs* are built.[6] The programs listed here have been updated:

`mlbibtex` aims to replace BibTeX;

`mlbiblatex` builds bibliographies (.bbl source files) suitable for the biblatex package [12]; it can be an

[*] Previously entitled: *MlBibTeX Now Deals with Unicode*. Polish title: *MlBibTeX od teraz rozumie Unicode*.

[1] **M**ulti**L**ingual BibTeX.

[2] BibTeX can only perform *lexicographic sorts*; its sort procedure cannot deal with numbers.

[3] e**X**tensible **M**arkup **L**anguage.

[4] **A**merican **S**tandard **C**haracter **I**nformation **I**nterchange.

[5] **S**cheme implementation of **XML** [11].

[6] We can statically determine all the modules composing such an executable program. Besides, each program has its own arguments, some being irrelevant for other programs. That is why we think that building separate programs is better. But if end-users prefer to have only one program with more options, we can do that with a wrapper program written using a script language.

```
%encoding = latin1

@BOOK{henze1973,
      AUTHOR = {first => Hans Werner,
                last => Henze},
      TITLE = {Zweites Violinkonzert für
               Sologeiger, Tonband,
               Baß-bariton und 33
               Instrumentalisten},
      PUBLISHER = {B. Scott Söhne},
      ADDRESS = {Mainz},
      YEAR = 1973,
      LANGUAGE = german}
```

Figure 1: Example using the Latin 1 encoding.

```
%encoding = latin2

@BOOK{morys-twarowski2016,
      AUTHOR = {first => Michael,
                last => Morys-Twarowski},
      TITLE = {Polskie Imperium. {Wszystkie
               kraje podbite przez
               rzeczpospolitą}},
      PUBLISHER = {Ciekawostki Historyczne.pl},
      ADDRESS = {Kraków},
      DATE = {2016-02-17},
      LANGUAGE = polish}
```

Figure 2: Example using the Latin 2 encoding.

```
%encoding = utf8

@BOOK{lem1964,
      AUTHOR = {Stanisław Lem},
      TITLE = {Bajki robotów},
      PUBLISHER = {Wydawnictwa Literackiego},
      YEAR = 1964,
      LANGUAGE = polish}
```

Figure 3: Example using the UTF-8 encoding.

alternative to the Biber bibliography processor [10];

`mlbibcontext` generates bibliographies suitable for ConTeXt;

`mlbib2xml` converts .bib files according to the XML format internally used by MlBibTeX.

The `hal` program, used to populate the HAL[7] open-archive site [3] has not yet been updated.[8]

2 Formats recognised

The new `%encoding` directive at the beginning of a .bib file, allows the encoding of the file to be specified. Some extensions of ASCII — e.g., Latin 1, Latin 2 — are now recognised. More precisely, most *byte-based* encodings are handled, in particular UTF[9]-8. The UTF-16 encoding, based on 16-bit units, will be added to the allowed encodings later. We recommend end-users specify information about encoding explicitly, even though MlBibTeX tries to guess a .bib file's encoding, because it may be difficult to guess correctly. Let us consider the `file` command, generally used to determine such encodings on operating systems such as Linux and Mac OS X. Applying this command to the files of Figs. 1 and 2 reports that the used encodings belong to ISO-8859, a series of 8-bit character encodings — including Latin 1 (ISO-8859-1) for western European languages and Latin 2 (ISO-8859-2) for eastern European Latin-alphabet languages — but gives no more precise information.[10]

Let us be clear that a text may use citation keys belonging to *several* .bib files with different encodings, for example, the three files given in Figs. 1–3 (notice the German letter 'ß' directly included in Fig. 1 and the Polish diacritical signs in Figs. 2 and 3). All the syntactic extensions for .bib files are still usable, including the new syntax for people's names by means of *keywords* (cf. Figs. 1 and 2). Most of the fields added by the biblatex package are recognised,[11] too; an example is the `DATE` field, used within Fig. 2 instead of the fields `YEAR`, `MONTH` and `DAY`.[12]

By default, MlBibTeX looks for .bib files for bibliographical entries, the default encoding of such files being Latin 1. It can also parse XML files for bibliographical entries, according to the `mlbiblio` format used by MlBibTeX.[13] The bibliographical entries cited throughout a document can be saved as an XML file, too. Hereafter we give two simple examples of using the interface with Scheme. It consists of Scheme definitions put in *initialisation files* located in your home directory. On Unix-based systems, the executable programs derived from MlBibTeX's kernel look for the following initialisation files:

`mlbibtex` ⟸ ~/.mlbibtex
`mlbiblatex` ⟸ ~/.mlbiblatex
`mlbibcontext` ⟸ ~/.mlbibcontext

[7] ***H**yper-**A**rticle en **L**igne*, that is, 'hyper-article on-line'.

[8] Since the format used for metadata by this site has changed, a new version of this program requires major rewriting; this will be done for a future release.

[9] **U**nicode **T**ransformation **F**ormat.

[10] It is unlikely that one end-user uses .bib files with these two encodings, so changing the default input encoding — as shown below — may fix this problem. But relying on this technique is error-prone.

[11] By 'recognised', we mean that a *type* is associated with such a field, and type-checking is performed as soon as corresponding values are parsed.

[12] This last field is recognised by MlBibTeX, but is not used by 'old' BibTeX's standard bibliography styles.

[13] Conventions given in [2] by means of a DTD (**D**ocument **T**ype **D**efinition) are now refined using XML Schema [17].

```
\documentclass{article}

\usepackage[T1]{fontenc}
%%  \usepackage[utf8]{inputenc}

\begin{document}

Did you hear \cite{henze1973}? And did you read
\cite{lem1964,morys-twarowski2016}?

\bibliography{figure-1,figure-2,figure-3}
\bibliographystyle{plain}

\end{document}
```

Figure 4: LaTeX document using Figs. 1–3's entries.

In particular, you can:

- allow MlBibTeX to look for an ⟨*f*⟩-mlbiblio.xml file when an ⟨*f*⟩.bib file is not found:

  ```
  ((bib-files-functions-pv 'set)
   (list s-parse-bib-file
         sxmlh-parse-mlbiblio-xml-file))
  ```

- change the default encoding of .bib files:

  ```
  ((encodings-pv
    'set-default-4-bib-files)
   'utf8)
  ```

You can use *prefixes* for different namespaces as described in [5], and put *inexact* information according to [6]'s syntax, but only with the two programs `mlbibtex` and `mlbibtex2xml`. The programs `mlbiblatex` and `mlbibcontext` have not incorporated these features yet.

3 Bibliography styles

BibTeX's standard bibliography styles written using [13]'s language can be used by the executable program `mlbibtex`, even if some fields introduced by the biblatex package are used instead of standard fields — e.g., the `DATE` field, instead of the standard fields `YEAR` and `MONTH`. Styles written using the nbst[14] language can be used, too. The two executable programs `mlbiblatex` and `mlbibcontext` use *direct styles* — using MlBibTeX's terminology, such styles are wholly written in Scheme [4]; these styles have been updated.

4 Output routines

The encoding of an output file generated by our programs is:

ASCII for a file suitable for LaTeX, unless another encoding is given within the master file's preamble by means of the inputenc or as an option of the `mlbiblatex` program;

UTF-8 for a file suitable for ConTeXt (the option allowing the choice of an encoding has been removed) or an XML file built by the `mlbib2xml` program, unless another encoding is given as an option.

Now we give a simple example by considering the LaTeX document given in Fig. 4. Let us recall that 'old' BibTeX operates on .aux files and never reads .tex files. On the contrary, MlBibTeX reads both an .aux file and the preamble of the corresponding .tex file. If Fig. 4 is processed *as it is*, the first reference built by the executable program `mlbibtex` looks like:

```
\bibitem{henze1973}
Hans Werner Henze.
\newblock {\em Zweites Violinkonzert
f\"{u}r Sologeiger, Tonband,
Ba{\ss}-bariton... } ...
```

that is, all the accented letters are replaced by the TeX commands used to produce them, since the encoding is supposed to be ASCII. If the line concerning the inputenc package in Fig. 4 is uncommented, this first reference becomes:

```
\bibitem{henze1973}
Hans Werner Henze.
\newblock {\em Zweites Violinkonzert für
Sologeiger, Tonband, Baß-bariton... } ...
```

that is, the .bbl file built by MlBibTeX is encoded using UTF-8.

5 Conclusion

We need to revise the installation procedure, some points now being unsatisfactory. The complete documentation also needs to be updated. But now MlBibTeX is ready to deal with Unicode.

6 Acknowledgements

Many thanks to the Polish translators: Ryszard Kubiak for the abstract and Jerzy B. Ludwichowski for the keywords. Thanks to GUTenberg, the French-speaking TeX users group, which offered me a grant for participating in this TUG@BachoTeX 2017 conference. I am also grateful to this definitive version's proofreaders: Karl Berry and Barbara Beeton.

References

[1] ConTeXt Garden: *Bibliographies in MkII.* April 2012. `http://wiki.contextgarden.net/Bibliography`.

[14] **N**ew **B**ibliography **ST**yles. Let us recall that this language is close to the first version of XSLT (e**X**tensible **S**tylesheet **L**anguage **T**ransformations) [16].

[2] Jean-Michel HUFFLEN: "Multilingual Features for Bibliography Programs: From XML to MlBIBTEX". In: *EuroTEX 2002*, pp. 46–59. Bachotek, Poland. April 2002.

[3] Jean-Michel HUFFLEN: "From Bibliography Files to Open Archives: The Sequel". In: Karl BERRY, Jerzy B. LUDWICHOWSKI and Tomasz PRZECHLEWSKI, eds., *Proc. EuroBachoTEX 2011 Conference*, pp. 61–66. Bachotek, Poland. April 2011.

[4] Jean-Michel HUFFLEN: "MlBIBTEX and Its New Extensions". In: *Proc. 6th ConTEXt Meeting & EuroTEX 2012*, pp. 82–91. Breskens, The Netherlands. October 2012.

[5] Jean-Michel HUFFLEN: "Managing Name Conflicts and Aliasing with MlBIBTEX". In: Tomasz PRZECHLEWSKI, Karl BERRY, Bogusław JACKOWSKI and Jerzy B. LUDWICHOWSKI, eds., *What Can Typography Gain from Electronic Media? Proc. BachoTEX 2014 conference*, pp. 13–16. Bachotek, Poland. April 2014.

[6] Jean-Michel HUFFLEN: "Dealing with Ancient Works in Bibliographies". *ArsTEXnica*, Vol. 18, pp. 81–86. In Proc. GUIT meeting 2014. October 2014. `http://www.guitex.org/home/images/ArsTeXnica/AT018/hufflen-verona.pdf`.

[7] Jean-Michel HUFFLEN: "From MlBIBTEX 1.3 to 1.4". In: Tomasz PRZECHLEWSKI, Karl BERRY, Bogusław JACKOWSKI and Jerzy B. LUDWICHOWSKI, eds., *Various Faces of Typography. Proc. BachoTEX 2015 conference*, pp. 13–17. Bachotek, Poland. April 2015.

[8] Jean-Michel HUFFLEN: "MlBIBTEX 1.4: The New Version". *ArsTEXnica*, Vol. 20, pp. 35–39. In Proc. GUIT meeting 2015. October 2015.

[9] Jean-Michel HUFFLEN: "MlBIBTEX & ConTEXt: Face-to-Face". In: *Proc. 9th ConTEXt Meeting*, pp. 27–48. Nasbinals, France. Abridged version. September 2016.

[10] Philip KIME and François CHARETTE: *biber. A Backend Bibliography Processor for biblatex. Version biber 2.7 (biblatex 3.7)*. 5 December 2016. `https://ctan.org/pkg/biber`.

[11] Oleg E. KISELYOV: *XML and Scheme*. September 2005. `http://okmij.org/ftp/Scheme/xml.html`.

[12] Philipp LEHMAN, with Philip KIME, Audrey BORUVKA and Joseph WRIGHT: *The biblatex Package. Programmable Bibliographies and Citations. Version 3.7*. 16 November 2016. `https://ctan.org/pkg/biblatex`.

[13] Oren PATASHNIK: *Designing BIBTEX Styles*. February 1988. Part of the BIBTEX distribution.

[14] Alex SHINN, John COWAN, and Arthur A. GLECKLER, with Steven GANZ, Aaron W. HSU, Bradley LUCIER, Emmanuel MEDERNACH, Alexey RADUL, Jeffrey T. READ, David RUSH, Benjamin L. RUSSEL, Olin SHIVERS, Alaric SNELL-PYM and Gerald Jay SUSSMAN: *Revised^7 Report on the Algorithmic Language Scheme*. 6 July 2013. `http://trac.sacrideo.us/wg/raw-attachment/wiki/WikiStart/r7rs.pdf`.

[15] THE UNICODE CONSORTIUM: *Unicode 9.0.0*. June 2016. `http://www.unicode.org/versions/Unicode9.0.0/`.

[16] W3C: *XSL Transformations (XSLT). Version 1.0*. W3C Recommendation. Edited by James Clark. November 1999. `http://www.w3.org/TR/1999/REC-xslt-19991116`.

[17] W3C: *XML Schema*. December 2008. `http://www.w3.org/XML/Schema`.

◇ Jean-Michel HUFFLEN
FEMTO-ST (UMR CNRS 6174)
& University of Bourgogne
Franche-Comté
16, route de Gray
25030 Besançon Cedex
France
`jmhuffle (at) femto-st dot fr`
`http://members.femto-st.fr/jean-michel-hufflen`