

Structures_BibTeX

Elmar Gregor Pitschke

July 28, 2006

Abstract

This PHP class is intended to enable an easy access and creation of BibTeX files. It is intended to be used in projects which want to enable BibTeX import or export.

1 Features

- Parsing of BibTeX Data in a PHP Array
- Adding An Entry
- Exporting to BibTeX
- Optional Validation and Creation of Warnings
- Optional Cutting of Delimiters of the values

2 What is BibTeX?

BibTeX itself is a tool to store information about references. It is mainly used in conjunction with LaTeX and introduces its own File specification. For more information see the corresponding Wikipedia page.

3 Definition of BibTeX and File Specification

For a detailed description of the BibTeX file specification take a look at this summary of BibTeX.

3.1 Special Entries

3.1.1 String

The @string entry is special because it defines some kind of template. This example is taken from the BibTeXing document by Oren Patashnik and it describes how to produce nearly identical title fields for different entries:

```
1 @STRING( WGA = " World Gnus Almanac" )
2 @BOOK( almanac-66,
3       title = 1966 # WGA,
```

```

4         . . .
5
6 @BOOK(almanac-67,
7       title = 1967 # WGA,

```

The two string in both title values are concatenation by the '#' character. This character has to be surrounded by whitespaces.

It also has to be mentioned here that it is only possible to define one parameter using the string entry. So the following entry is invalid:

```

8 @String(mar = "march",
9        apr = "april")

```

This has to be written like this:

```

10 @String(mar = "march")
11 @String(apr = "april")

```

3.1.2 Preamble

The preamble entry is used to include some LaTeX commands in the .bbl file. This files is generated by bibtex on base of the .bib file. Inside the preamble entry string definitions are also allowed. An example looks like this:

```

12 @preamble {"This bibliography was generated on \today"}

```

The preamble entry is only of interest if using in conjunction with bibtex and tex. this is not in the scope of this class and therefore the preamble entry is not supported!

4 Requisites

There are some things which are Requisites in order that the BibTex data may be parsed correctly.

- No '@' is allowed in Comments. This has been validated using 'bibtool'. Therefore if an '@' is detected an entry is about to begin.
- Beginning from the first opening brace after the '@' the entry ends as soon as the amount of opened and closed braces is equal. An escaped brace is not taken into account.

5 Parsing the BibTex File

The Parsing is done by going through every character in the text.

```

13 for($i = 0 ; $i < strlen($this->content) ; $i++)

```

where \$this->content is the BibTex data. Inside this loop the actual character is stored in \$char and the previous character is stored in \$lastchar. The \$lastchar is needed to detect if a brace is escaped. The amount of opened braces is stored

in \$open. The boolean \$entry is used to specify if the actual character is inside an entry or not. If not the character belongs to a comment. Every character of an entry is appended to \$buffer. Finally when an error, in this case unbalanced parenthesis, is detected \$valid is set to false. At first the the beginning of an entry is detected:

```

14 if( ( 0 == $open ) && ( '@' == $char ) ) {
15     $entry = true;
16 }

```

Inside an entry the \$open should be greater zero. Then every '@' is ignored. Outside an entry the \$open has to be Zero (see second item in the Requisites). Then an '@' marks the beginning of an entry and therefore the \$entry is set to true. Then an opening brace is detected:

```

17 elseif ( $entry && ( '{' == $char ) && ( '\\\ ' !=
18     $lastchar ) ) {
19     $open++;

```

Opening braces are only counted if they are inside entries and if they are not escaped. If they are escaped the previous character would be '\'. If an opening brace is detected the value stored in \$open is incremented. Finally an closing brace is detected:

```

20 elseif ( $entry && ( '}' == $char ) && ( '\\\ ' !=
21     $lastchar ) ) {
22     $open--;
23     if( $open < 0 ) {
24         $valid = false;
25     }
26     if( 0 == $open ) {
27         $entry = false;
28         $entrydata = $this->_parseEntry($buffer);
29         if(!$entrydata) {
30             $valid = false;
31         } else {
32             $this->data[] = $entrydata;
33         }
34         $buffer = '';
35     }

```

Closing braces are only counted if they are inside entries and if they are not escaped. If an closing brace is detected the value stored in \$open is decreased. Next the \$open is checked. If this value should be less than zero than there are more closing than opening braces which is not possible in a valid BibTex File. If at this point the value stored in \$open is equal zero the end of an entry is reached. Then \$entry is set to false - we are not inside an entry anymore. The entry stored in \$buffer is parsed separately in the private function _parseEntry. If this functions returns false, something went wrong and valid is set to false. It is important to mention that _parseEntry returns false if the opening delimiter

of a value does not match the closing. This usually happens if the follow up of braces is not correct. If the entry is parsed correctly it is appended to the array `$this->data`. And finally the buffer is cleared. There are two final things to do in the loop:

```

36 if( $entry ) {
37     $buffer .= $char;
38 }
39 $lastchar = $char;

```

If we are inside an entry the actual character is appended to the buffer and the actual character is stored as the previous character in the next cycle. After finishing the loop an error is risen when valid is false. The message than is 'Unbalanced parenthesis'. It is important to mention that Comments are simply dropped.

6 Warnings

Even if BibTeX Data is parsed the content may be not correct - speaking from a format point of view. You may refer to this as non strict. To emphasise this again a warning is detected when something is wrong with the data but the parsing will not break! The detected Warnings include:

- `WARNING_AT_IN_BRACES`: A Value is delimited by Braces. Then inside a @ is not allowed.
- `WARNING_ESCAPED_DOUBLE_QUOTE_INSIDE_DOUBLE_QUOTES`: A Value is delimited by Braces. Then inside no escaped double quote is allowed. The double quotes should be written as:
" "
- `WARNING_UNBALANCED_AMOUNT_OF_BRACES`: The amount of braces inside a value is not equal (opening and closing). The parser fails if in the complete entry this amount is not correct. But if only on an entry it is not correct then this is only a warning. As a matter of fact of the parser does not fail but there is an unbalanced amount of braces in an entry this warning has to be generated something times two.
- `WARNING_MULTIPLE_ENTRIES`: Every entry is identified by a unique string. This warning is created if there are at least two entries with the same identification.

7 Using Structures_BibTeX

Lets suppose you got ten references from PubMed and saved them under "texmed.bib". The content of this BibTeX File looks like this:

```

40 % 9606928 (JID)
41 @Article{pmid16594765,
42     Author="Maccallum, Robert C and Browne, Michael W
         and Cai, Li",

```

```

43     Title="{Testing differences between nested
         covariance structure models: Power analysis and
         null hypotheses}",
44     Journal="Psychol Methods",
45     Year="2006",
46     Volume="11",
47     Number="1",
48     Pages="19--35",
49     Month="Mar"
50     }
51
52
53 % 0375356 (JID)
54 @Article{pmid16606695,
55     Author="Waterhouse, Nigel J and Sutton, Vivien R
         and Sedelies, Karin A and Ciccone, Annette and
         Jenkins, Misty and Turner, Stephen J and Bird,
         Phillip I and Trapani, Joseph A",
56     Title="{Cytotoxic T lymphocyte-induced killing in
         the absence of granzymes A and B is unique and
         distinct from both apoptosis and perforin-
         dependent lysis}",
57     Journal="J Cell Biol",
58     Year="2006",
59     Volume="173",
60     Number="1",
61     Pages="133--144",
62     Month="Apr"
63     }
64
65
66 % 101232529 (JID)
67 @Article{pmid16594196,
68     Author="Wein, Simon",
69     Title="{The firmament of consciousness}",
70     Journal="Palliat Support Care",
71     Year="2005",
72     Volume="3",
73     Number="1",
74     Pages="55",
75     Month="Mar"
76     }
77
78
79 % 101262774 (JID)
80 @Article{pmid16566567,
81     Author="Walker, James and Maccallum, Matt and
         Fischer, Carl and Kopcha, Robert and Saylor,
         Roy and McCall, John",
82     Title="{Sedation using dexmedetomidine in pediatric

```

```

83         burn patients}”,
84     Journal=“J Burn Care Res”,
85     Year=“2006”,
86     Volume=“27”,
87     Number=“2”,
88     Pages=“206--210”,
89     Month=“Mar”
90     }
91
92 % 7603616 (JID)
93 @Article{pmid16580511,
94     Author=“Henderson, Michael A”,
95     Title=“{In reply to drs. Godinez and gombos}”,
96     Journal=“Int J Radiat Oncol Biol Phys”,
97     Year=“2006”,
98     Volume=“64”,
99     Number=“5”,
100    Pages=“1611”,
101    Month=“Apr”,
102    Note=“Letter”
103    }
104
105
106 % 7603509 (JID)
107 @Article{pmid16597590,
108     Author=“Kinross KM and Clark AJ and Iazzolino RM
109     and Humbert PO”,
110     Title=“{E2f4 regulates fetal erythropoiesis through
111     the promotion of cellular proliferation}”,
112     Journal=“Blood”,
113     Year=“2006”,
114     Month=“Apr”,
115     Note=“JOURNAL ARTICLE”
116     }
117
118 % 8205768 (JID)
119 @Article{pmid16598744,
120     Author=“Thomas D and Kansara M”,
121     Title=“{Epigenetic modifications in osteogenic
122     differentiation and transformation}”,
123     Journal=“J Cell Biochem”,
124     Year=“2006”,
125     Month=“Apr”,
126     Note=“JOURNAL ARTICLE”
127     }
128 % 0374236 (JID)

```

```

129 @Article{pmid16565958,
130   Author="Jefford M",
131   Title="{Factors associated with interval adherence
        to mammography screening in a population-based
        sample of New Hampshire women}",
132   Journal="Cancer",
133   Year="2006",
134   Month="Mar",
135   Note="JOURNAL ARTICLE"
136   }
137
138
139 % 101232529 (JID)
140 @Article{pmid16594228,
141   Author="Wein, Simon",
142   Title="{Death of a generation}",
143   Journal="Palliat Support Care",
144   Year="2003",
145   Volume="1",
146   Number="4",
147   Pages="381--383",
148   Month="Dec"
149   }
150
151
152 % 100927353 (JID)
153 @Article{pmid16608535,
154   Author="Sloan EK and Pouliot N and Stanley KL and
        Chia J and Moseley JM and Hards DK and Anderson
        RL",
155   Title="{Tumor-specific expression of alphavbeta3
        integrin promotes spontaneous metastasis of
        breast cancer to bone}",
156   Journal="Breast Cancer Res",
157   Year="2006",
158   Volume="8",
159   Number="2",
160   Pages="R20",
161   Month="Apr",
162   Note="JOURNAL ARTICLE"
163   }

```

So you got different Types with different entries and some comments between the entries. To get this BibTex data in a PHP Array you would do something like this:

```

163 <?php
164 require_once 'Structures_BibTex.php';
165
166 $foo=new Structures_BibTex();

```

```

167 $ret=$foo->loadFile('texmed.bib');
168 if(PEAR::isError($ret)) {
169     print $ret->getMessage();
170     die();
171 }
172 if(PEAR::isError($ret=$foo->parse())) {
173     print $ret->getMessage();
174 } else {
175     print_r($foo->data);
176     print "_____\\n";
177     print "Warnings:\\n";
178     print_r($foo->warnings);
179 }
180 ?>

```

The result would be:

```

180 Array
181 (
182     [0] => Array
183     (
184         [month] => Mar
185         [pages] => 19--35
186         [number] => 1
187         [volume] => 11
188         [year] => 2006
189         [journal] => Psychol Methods
190         [title] => Testing differences between nested
                covariance structure models: Power
                analysis and null hypotheses
191         [author] => Array
192         (
193             [0] => Maccallum, Robert C
194             [1] => Browne, Michael W
195             [2] => Cai, Li
196         )
197         [cite] => pmid16594765
198         [type] => article
199     )
200
201     [1] => Array
202     (
203         [month] => Apr
204         [pages] => 133--144
205         [number] => 1
206         [volume] => 173
207         [year] => 2006
208         [journal] => J Cell Biol
209     )

```

```

210     [title] => Cytotoxic T lymphocyte-induced
211           killing in the absence of granzymes A and
212           B is unique and distinct from both
213           apoptosis and perforin-dependent lysis
214     [author] => Array
215     (
216       [0] => Waterhouse, Nigel J
217       [1] => Sutton, Vivien R
218       [2] => Sedelies, Karin A
219       [3] => Ciccone, Annette
220       [4] => Jenkins, Misty
221       [5] => Turner, Stephen J
222       [6] => Bird, Phillip I
223       [7] => Trapani, Joseph A
224     )
225     [cite] => pmid16606695
226     [type] => article
227   )
228 [2] => Array
229 (
230   [month] => Mar
231   [pages] => 55
232   [number] => 1
233   [volume] => 3
234   [year] => 2005
235   [journal] => Palliat Support Care
236   [title] => The firmament of consciousness
237   [author] => Array
238   (
239     [0] => Wein, Simon
240   )
241   [cite] => pmid16594196
242   [type] => article
243 )
244 [3] => Array
245 (
246   [month] => Mar
247   [pages] => 206--210
248   [number] => 2
249   [volume] => 27
250   [year] => 2006
251   [journal] => J Burn Care Res
252   [title] => Sedation using dexmedetomidine in
253           pediatric burn patients
254   [author] => Array
255   (

```

```

256         [0] => Walker , James
257         [1] => Maccallum , Matt
258         [2] => Fischer , Carl
259         [3] => Kopcha , Robert
260         [4] => Saylor , Roy
261         [5] => McCall , John
262     )
263
264     [cite] => pmid16566567
265     [type] => article
266 )
267
268 [4] => Array
269 (
270     [note] => Letter
271     [month] => Apr
272     [pages] => 1611
273     [number] => 5
274     [volume] => 64
275     [year] => 2006
276     [journal] => Int J Radiat Oncol Biol Phys
277     [title] => In reply to drs. Godinez and
                gombos
278     [author] => Array
279         (
280             [0] => Henderson , Michael A
281         )
282
283     [cite] => pmid16580511
284     [type] => article
285 )
286
287 [5] => Array
288 (
289     [note] => JOURNAL ARTICLE
290     [month] => Apr
291     [year] => 2006
292     [journal] => Blood
293     [title] => E2f4 regulates fetal
                erythropoiesis through the promotion of
                cellular proliferation
294     [author] => Array
295         (
296             [0] => Kinross KM
297             [1] => Clark AJ
298             [2] => Iazzolino RM
299             [3] => Humbert PO
300         )
301
302     [cite] => pmid16597590

```

```

303         [type] => article
304     )
305
306 [6] => Array
307 (
308     [note] => JOURNAL ARTICLE
309     [month] => Apr
310     [year] => 2006
311     [journal] => J Cell Biochem
312     [title] => Epigenetic modifications in
                 osteogenic differentiation and
                 transformation
313     [author] => Array
314         (
315             [0] => Thomas D
316             [1] => Kansara M
317         )
318
319     [cite] => pmid16598744
320     [type] => article
321 )
322
323 [7] => Array
324 (
325     [note] => JOURNAL ARTICLE
326     [month] => Mar
327     [year] => 2006
328     [journal] => Cancer
329     [title] => Factors associated with interval
                 adherence to mammography screening in a
                 population-based sample of New Hampshire
                 women
330     [author] => Array
331         (
332             [0] => Jefford M
333         )
334
335     [cite] => pmid16565958
336     [type] => article
337 )
338
339 [8] => Array
340 (
341     [month] => Dec
342     [pages] => 381--383
343     [number] => 4
344     [volume] => 1
345     [year] => 2003
346     [journal] => Palliat Support Care
347     [title] => Death of a generation

```

```

348         [author] => Array
349             (
350                 [0] => Wein, Simon
351             )
352
353         [cite] => pmid16594228
354         [type] => article
355     )
356
357 [9] => Array
358 (
359     [note] => JOURNAL ARTICLE
360     [month] => Apr
361     [pages] => R20
362     [number] => 2
363     [volume] => 8
364     [year] => 2006
365     [journal] => Breast Cancer Res
366     [title] => Tumor-specific expression of
367                 alphavbeta3 integrin promotes spontaneous
368                 metastasis of breast cancer to bone
369     [author] => Array
370         (
371             [0] => Sloan EK
372             [1] => Pouliot N
373             [2] => Stanley KL
374             [3] => Chia J
375             [4] => Moseley JM
376             [5] => Hards DK
377             [6] => Anderson RL
378         )
379     [cite] => pmid16608535
380     [type] => article
381 )
382 )
383
384 Warnings:
385 Array
386 (
387 )

```

As you can see, every entry has been parsed correctly and no warning has been fired. The first thing in the PHP source we create an instance of `Structures.BibTex`. Then we try to read the content of a file. In the class the BibTex data is stored in `content`. This may be set externally, so there is no need to get that from a file, this can also be for example from an input field. Then the content is parsed. This may break and therefore it is checked for errors. Of course this will no happen here. The parsed data is stored in `data` and this is

simply printed. After that, just in case, the warnings are printed.

Now an example which will fire every known Warning. The BibTeX Data is stored in ill.bib and looks like this:

```
387 The first example has an At inside braces
388 @article{ill1,
389     author="Elmar Pitschke",
390     foo={Where does the @ belong?}
391 }
392
393 The next example illustrates the escaped double quotes
      warning
394 @article{ill2,
395     author="Elmar Pitschke",
396     foo="Here it \" is",
397 }
398
399 The next one is tricky - unbalanced amount of braces in
      entry
400 @article{ill3,
401     author="Elmar Pitschke",
402     foo="Here it { is",
403     bar="And here the } second"
404 }
405
406 And finally we define the second identification again
407 @article{ill2,
408     author="Elmar Pitschke",
409     foo="This one is correct",
410 }
```

Using the same PHP script as above the output would look something like this:

```
410 Array
411 (
412     [0] => Array
413         (
414             [foo] => Where does the @ belong?
415             [author] => Array
416                 (
417                     [0] => Elmar Pitschke
418                 )
419             [cite] => ill1
420             [type] => article
421         )
422     [1] => Array
423         (
424             [foo] => Here it \"is
425         )
426 )
```

```

427 .....[ author ] => Array
428 .....(
429 .....[0] => Elmar Pitschke
430 .....)
431
432 .....[ cite ] => ill2
433 .....[ type ] => article
434 .....)
435
436 .....[2] => Array
437 .....(
438 .....[ bar ] => And here the } second
439 .....[ foo ] => Here it { is
440 .....[ author ] => Array
441 .....(
442 .....[0] => Elmar Pitschke
443 .....)
444
445 .....[ cite ] => ill3
446 .....[ type ] => article
447 .....)
448
449 .....[3] => Array
450 .....(
451 .....[ foo ] => This one is correct
452 .....[ author ] => Array
453 .....(
454 .....[0] => Elmar Pitschke
455 .....)
456
457 .....[ cite ] => ill2
458 .....[ type ] => article
459 .....)
460
461 )
462 -----
463 Warnings:
464 Array
465 (
466 .....[0] => Array
467 .....(
468 .....[ warning ] => WARNING_AT_IN_BRACES
469 .....[ entry ] => {Where does the @ belong?}
470 .....[ wholeentry ] => @article{ill1 ,
471 .....author="Elmar Pitschke" ,
472 .....foo={Where does the @ belong?}
473 .....
474 .....)
475
476 .....[1] => Array

```

```

477 (
478 [warning] =>
    WARNING_ESCAPED_DOUBLE_QUOTE_INSIDE_DOUBLE_QUOTES
479 [entry] => "Here it \" is"
480 [wholeentry] => @article{ill2 ,
481   author="Elmar Pitschke" ,
482   foo="Here it \" is" ,
483
484   )
485
486 [2] => Array
487 (
488   [warning] =>
    WARNING_UNBALANCED_AMOUNT_OF_BRACES
489   [entry] => "And here the } second"
490   [wholeentry] => @article{ill3 ,
491   author="Elmar Pitschke" ,
492   foo="Here it { is" ,
493   bar="And here the } second"
494
495   )
496
497 [3] => Array
498 (
499   [warning] =>
    WARNING_UNBALANCED_AMOUNT_OF_BRACES
500   [entry] => "Here it { is"
501   [wholeentry] => @article{ill3 ,
502   author="Elmar Pitschke" ,
503   foo="Here it { is" ,
504   bar="And here the } second"
505
506   )
507
508 [4] => Array
509 (
510   [warning] => WARNING_MULTIPLE_ENTRIES
511   [entry] => ill2
512   [wholeentry] =>
513
514   )
515 )

```

As you can see every entry is parsed correctly but still there are some warnings that you should usually look after. Another possibility to use Structures.BibTex is to add another entry and export it to BibTex. We suppose you have a BibTex File called yourfile.bib and want to add some content to it. Then you would do something like this:

```
515 <?php
```

```

516 require_once 'Structures_BibTex.php';
517
518 $foo=new Structures_BibTex();
519 $ret=$foo->loadFile('yourfile.bib');
520 if(PEAR::isError($ret)) {
521     print $ret->getMessage();
522     die();
523 }
524 if(PEAR::isError($ret=$foo->parse())) {
525     print $ret->getMessage();
526 } else {
527     $add=array();
528     $add['type']='article';
529     $add['author'][]='John.Doe';
530     $add['author'][]='Jane.Doe';
531     $add['title']='Foo.Bar';
532     $add['cite']='fool';
533     $foo->addEntry($add);
534     print $foo->bibTex();
535 }
536 ?>

```

The constructor has two possible parameter. Both are boolean values to set two modes. The first one whether the delimiters should be stripped from the values and the second whether warnings should be fired. Both are set to true per default. And here is another example if you want to see this class live in action.

8 History

After some struggling about the correct way to parse the data the first beta release has been released.

- 29.05.06: Squashed a bug when an equal sign has been used inside an abstract entry.
- 30.05.06: Another little Bug squashed. Some (PHP) Warnings were generated when running in E_ALL Mode.
- 03.06.06: Mayor Bug: Any entry like this: "{foo} bar" is perfectly legal but has been ignored because when the first delimiters are cut the opening { do not match any closing. This problem is now solved. Another bug has been that the script accidentally raised an Pear Error (Unbalanced parenthesis) when an unsupported entry has been found. This is of course solved, too. Also added the function getStatistic to compare the parsed data with output from bibtool (bibtool -@ foo.bib) to make sure no entry is lost!
- 06.06.06: Started the Pear Voting on the Package

- 10.06.06: Changed the way options are set. Now they are set using a hash table. How this is done has been copied from the DB Pear package. The advantage is that it is simple to introduce new parameters and old code still works. And most of the parameters are optional, but when initialising the class and you want to set the third optional parameter you also have to set the first two. Therefore I think it's better to set the values using a hash table. Also introduced the method `setOption` to have a generic way for setting options. Now all specialised methods to set an option are no longer needed.
- 16.06.06: Beginning to source out the options from the parsing. Added `Option` to unwrap and wrap the text.
- 22.06.06: Added the parsing of authors, which was quite a work. Also added a numerous amount of tests to make sure this works.
- 15.07.06: Adding parsing of authors including necessary tests. Adding RTF export. Minor changes.

9 ToDo

There are still some things to do and will ever be. Speaking in other words the request on features and the debugging will never end. Here is a list of things still to come:

- Outside of an BibTeX entry an `@` is not allowed. This has to be checked (see `plbib.bib`)
- Rewrite the check for multiple entries `-i`. Quite not optimal!
- Automatic Correcting of entries when a warning occurs
- Better adjusting of the BibTeX export
- Rewrite the parsing to do it more OOP Style `-i`. Better unit testing
- Text export
- SQL export (this idea is still in my head, no special idea here)
- Implementing more Warnings like checking if the type is known
- automatic creation of cite when adding an entry without one
- Parsing and interpreting of `@comment` and `@string`

It is easier to change the specification to fit the program than vice versa. Alan Perlis