

OSCAL

AutoWikiBrowser: Dealing with data sets

Marios Magioladitis

14.05.2017

AutoWikiBrowser

The semi-automated wiki editor

Version 5.9.0.1



AutoWikiBrowser: Dealing with data sets

Marios Magioladitis

14.05.2017

Wikipedia: 5,500,000 articles

- **Big Data:** Manual data processing is inadequate to deal with that amount of data
- Great need to apply same rules in all pages
- Many editors, not all familiar with Mediawiki and wikisyntax
- Vandalism
- Typos
- Mass re-categorisations etc.

Mediawiki

- Free and open-source wiki software.
- Originally developed by Magnus Manske and improved by Lee Daniel Crocker.
- It runs on many websites, including Wikipedia, Wiktionary and Wikimedia Commons.
- It is written in the PHP programming language and stores the contents into a database.
- The software is optimized to efficiently handle large projects, which can have terabytes of content and hundreds of thousands of hits per second.

Mediawiki

- Achieving scalability through multiple layers of caching and database replication has been a major concern for developers.
- On Wikipedia more than 1000 automated and semi-automated bots and other tools have been developed to assist in editing.

AutoWikiBrowser

- Created in 2006
- Open-source
- Semi-automated MediaWiki editor for **Windows**
- > 100 code syntax fixes with the use of regular expressions
- Fully compatible with all Wikipedia and their sister projects
- Allows plugins
- Written in C#

AWB is powerful

- Page edits since English Wikipedia was set up (2001): 852,605,210 (Special:Statistics)
- AWB edits in en.wiki since 2009: 108,633,087 (<https://tools.wmflabs.org/awb/stats/>)
- AWB edits in en.wiki 2008-2009: 4,904,034 (<https://tools.wmflabs.org/awb/stats/>)
- Vi.wiki edits with AWB: 193,854,162
- Lietuval.it edits with AWB: 875,640,318

The code at a glance

- > 3 millions lines of code [including libraries]
- 633K comments
- 1.6Mb code in C#
- 957 years of effort (COCOMO model)
- Many plugins
- Edit box supports the Microsoft Text Services Framework for use with speech recognition/handwriting applications
- Licence: GPL-2.0+
- <https://www.openhub.net/p/AutoWikiBrowser>

Code is available at Sourceforge

sourceforge [Browse](#) [Enterprise](#) [Blog](#) [Jobs](#) [Deals](#) [Help](#) [Log In](#) or [Join](#)

[SOLUTION CENTERS](#) [Go Parallel](#) [Resources](#) [Newsletters](#)

AutoWikiBrowser

Brought to you by: [magioladitis](#), [maxsem](#), [reedy_boy](#), [rjwilmsi](#)

[Summary](#) [Files](#) [Reviews](#) [Support](#) [Wiki](#) [Code](#)

★ 4.0 Stars (4)

↓ 80 Downloads (This Week)

📅 Last Update: 2015-10-01

🐦 Tweet 5

➕ 3

👍 Like 20

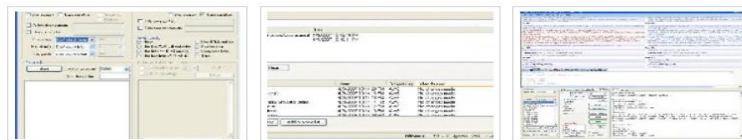


Download

AutoWikiBrowserS600.zip



[Browse All Files](#)



Description

AutoWikiBrowser is a semi-automated Wikipedia editor, designed to make tedious, repetitive tasks quicker and easier. For more information, see the project homepage at <http://en.wikipedia.org/wiki/Wikipedia:AutoWikiBrowser>

[AutoWikiBrowser Web Site](#)

Categories

Browsers

License

GNU General Public License version 2.0 (GPLv2)

Code is available at Sourceforge



AutoWikiBrowser

Brought to you by: magioladitis, maxsem, reedy_boy, rjwilmsi

Summary Files Reviews Support Wiki Code

★ 4 Stars (4)

Download 16,000

Twitter 5

Like +20

Download

Browse All Files

<https://sourceforge.net/projects/autowikibrowser/>



Description

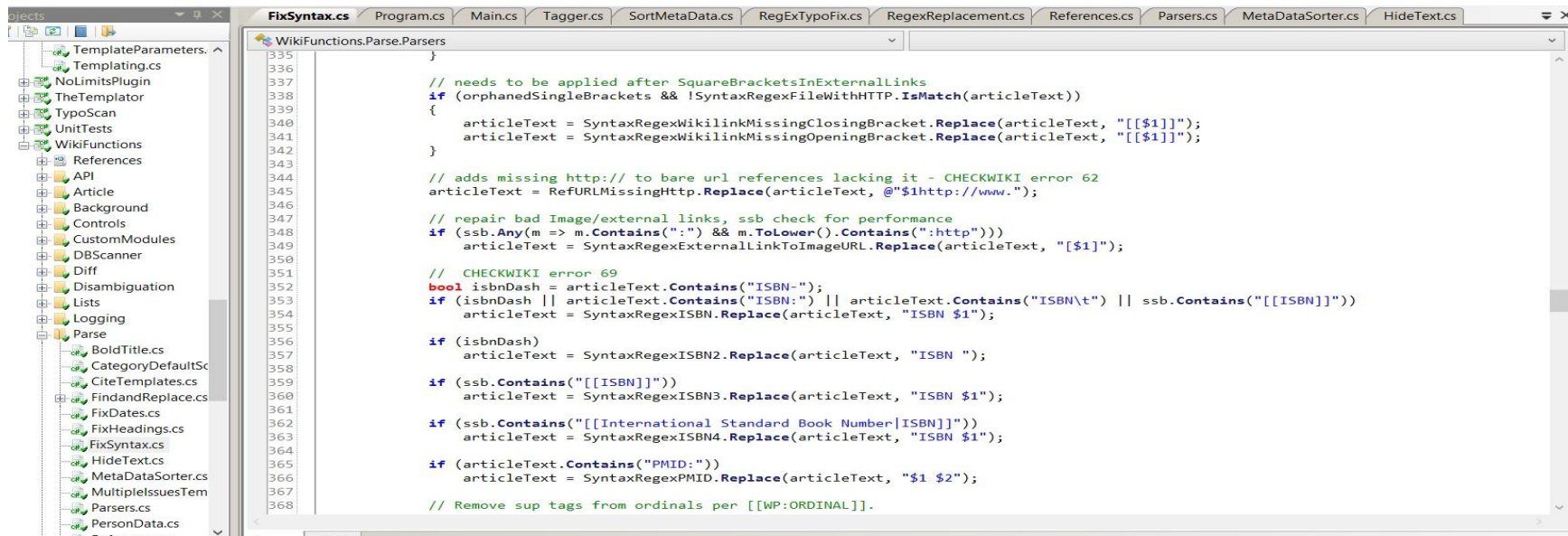
AutoWikiBrowser is a semi-automated Wikipedia editor, designed to make tedious, repetitive tasks quicker and easier. For more information, see the project homepage at <http://en.wikipedia.org/wiki/Wikipedia:AutoWikiBrowser>

[AutoWikiBrowser Web Site](#)

Categories
Browsers

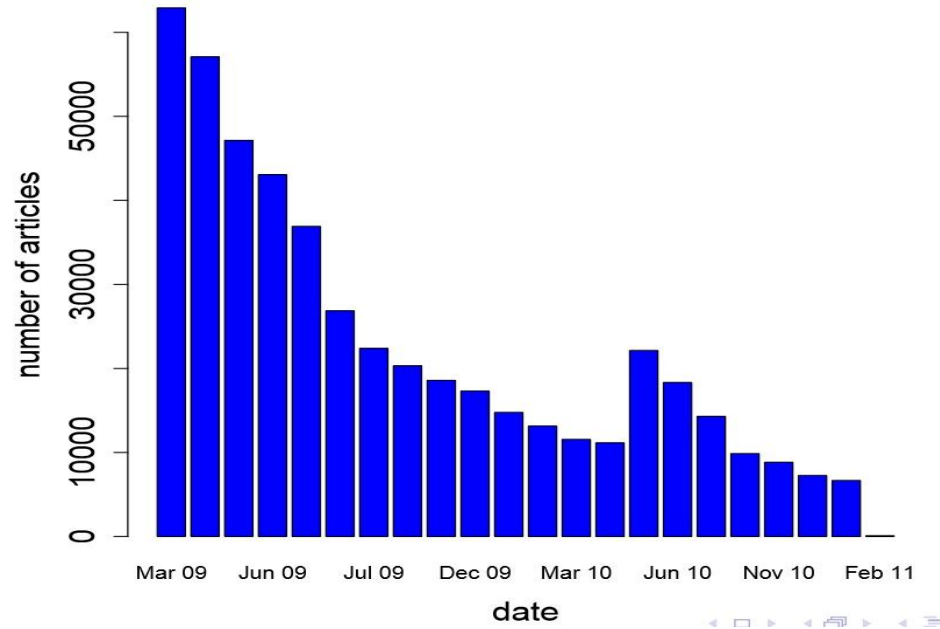
License
GNU General Public License version 2.0 (GPLv2)

The code

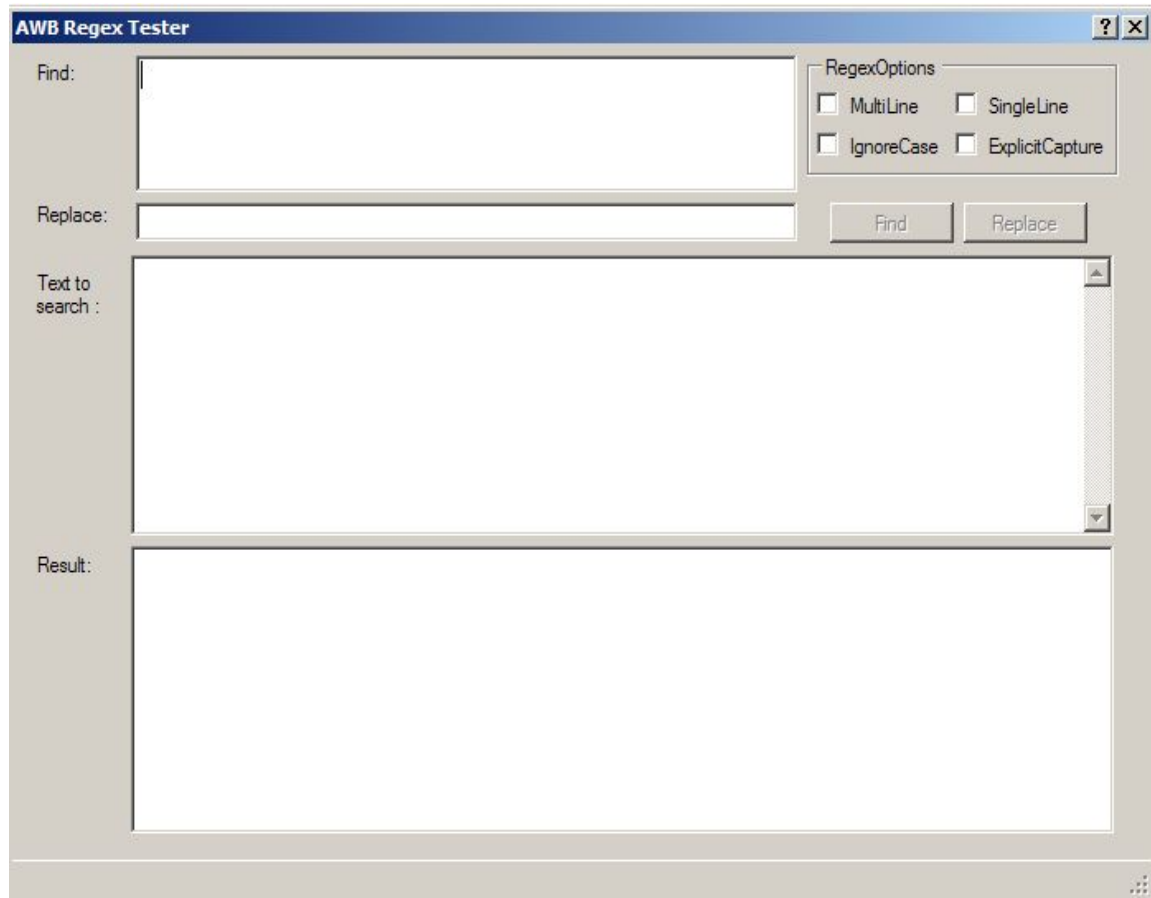


```
335 }
336
337 // needs to be applied after SquareBracketsInExternalLinks
338 if (orphanedSingleBrackets && !SyntaxRegexFileWithHTTP.IsMatch(articleText))
339 {
340     articleText = SyntaxRegexWikilinkMissingClosingBracket.Replace(articleText, "[[{$1}]");
341     articleText = SyntaxRegexWikilinkMissingOpeningBracket.Replace(articleText, "[[{$1}]");
342 }
343
344 // adds missing http:// to bare url references lacking it - CHECKWIKI error 62
345 articleText = RefURLMissingHttp.Replace(articleText, @"{$1http://www.}");
346
347 // repair bad Image/external links, ssb check for performance
348 if (ssb.Any(m => m.Contains(":") && m.ToLower().Contains("http")))
349     articleText = SyntaxRegexExternalLinkToImageUrl.Replace(articleText, "{$1}");
350
351 // CHECKWIKI error 69
352 bool isbnDash = articleText.Contains("ISBN-");
353 if (isbnDash || articleText.Contains("ISBN:") || articleText.Contains("ISBN\t") || ssb.Contains("[[ISBN]]"))
354     articleText = SyntaxRegexISBN.Replace(articleText, "ISBN $1");
355
356 if (isbnDash)
357     articleText = SyntaxRegexISBN2.Replace(articleText, "ISBN ");
358
359 if (ssb.Contains("[[ISBN]]"))
360     articleText = SyntaxRegexISBN3.Replace(articleText, "ISBN $1");
361
362 if (ssb.Contains("[[International Standard Book Number|ISBN]]"))
363     articleText = SyntaxRegexISBN4.Replace(articleText, "ISBN $1");
364
365 if (articleText.Contains("PMID:"))
366     articleText = SyntaxRegexPMID.Replace(articleText, "$1 $2");
367
368 // Remove sup tags from ordinals per [[WP:ORDINAL]].
```

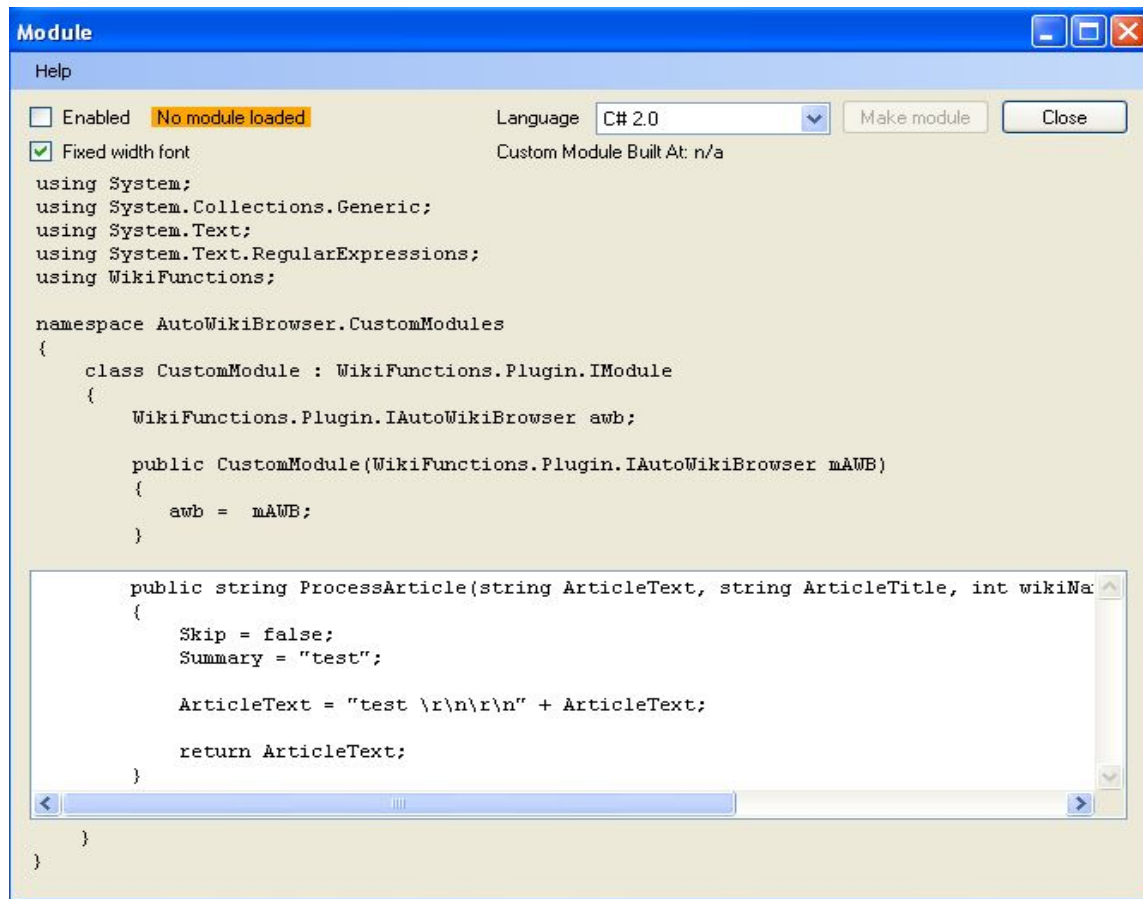

An example: BLP tags



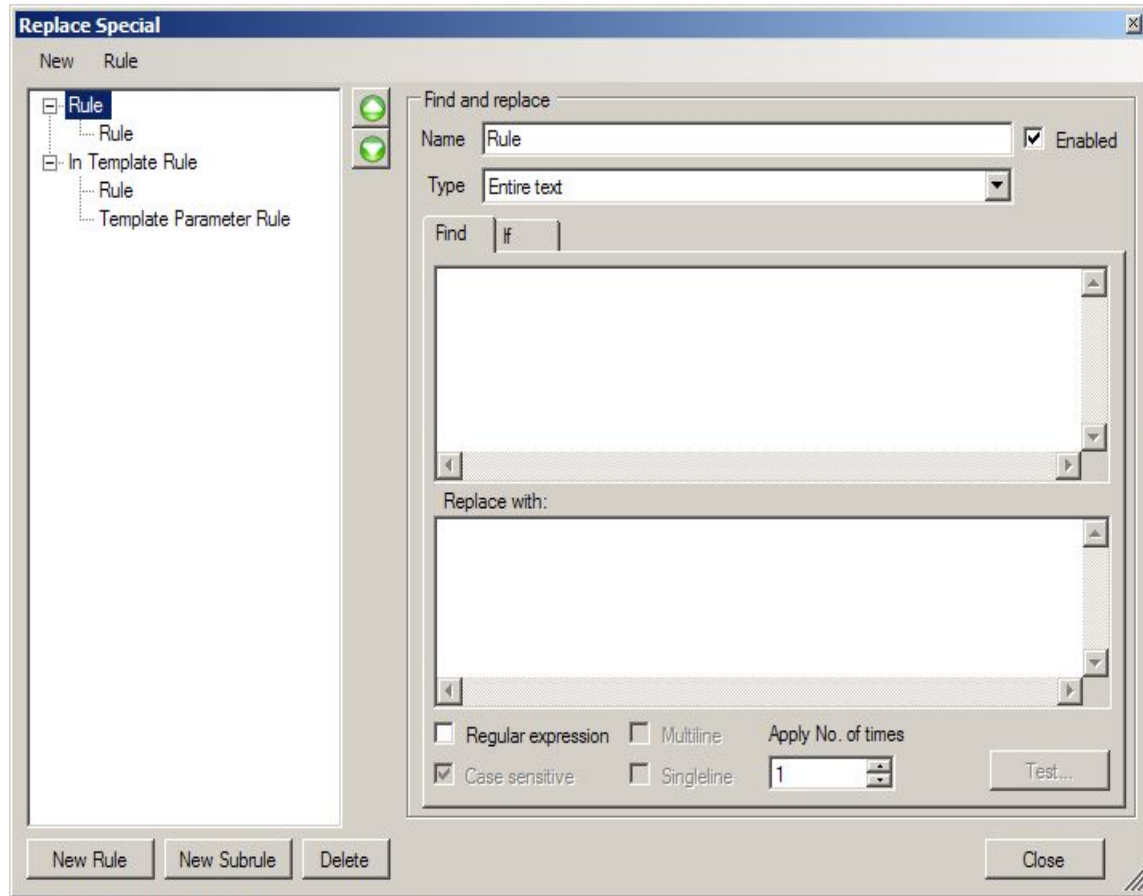
Advanced tools



Advanced tools



Advanced tools



The team



Phabricator

- It is a suite of web-based software development collaboration tools, including the Differential code review tool, the Diffusion repository browser, the Herald change monitoring tool, the Maniphest bug tracker and the Phriction wiki.
- Phabricator integrates with Git, Mercurial, and Subversion. It is available as free software under the Apache License, version 2.
- Phabricator was originally developed as an internal tool at Facebook. Phabricator's principal developer is Evan Priestley. Priestley left Facebook to continue Phabricator's development in a new company called Phacility.

Report bugs and ideas!

https://phabricator.wikimedia.org/project/view/1012/

PHABRICATOR

AutoWikiBrowser Public

Sort: Natural Filter: Open Tasks Manage Board

T111663 Tag AWB edits	T112219 Allow CTRL + C, V and X to work in the More tab	T99277 [Migrated] NullReferenceException in MainForm.SetProject	T109522 Have AWB fix some table ending issues	T99420 [Migrated] Remove stub templates from redirect pages
T111310 Ability for crosswiki login (propagating session), rather than to login to each and every wiki	T112202 AutoWikiBrowser exception error	T108817 The result returned by server was blank	T112078 Replace För with {{För}}-templates on sv.wp	T99714 [Migrated] Change {{(Expansion section)}} with {{(Empty section)}} in section
T99314 [Migrated] Sometimes upon clicking save AWB restarts same article without hand edited changes	T99715 [Migrated] Remove {{Stub}} when a more fine-grained stub tag already exists on the page	T107644 ArgumentException in MainForm.MainForm_FormClosing	T100695 [Migrated] Localization in Turkish	
T99303 [Migrated] AWB cannot save letter + combining diacritic when a precomposed Unicode glyph is available	T111660 Enable general fixes in Draft namespace	T106817 AccessViolationException in UnsafeNativeMethods.CallWindowProc	T99327 [Migrated] Hyphen rule applied inconsistently	
T37654 API edit of translate page gives Unknown error: "tpt-target-page"	T111693 Convert selected external links to HTTPS	T102218 ArgumentException in ApiEdit.Save	T99307 [Migrated] Typo fixing should be before SimplifyLinks	
T100288 [Migrated] Can't save default prefs	T109254 Fix double equals signs in template parameters	T100286 [Migrated] ThreadStateException in AsyncApiEdit.Abort (Main form Stop button, AsyncApiEdit.Abort)	T99290 [Migrated] AWB does not merge group references	
T41492 prop=info doesn't behave consistently when given a Special Page, vs when given a non special page which redirects to a special page (in comparison to another redirect)	T105498 Remove empty "See also"-section	T92352 ConfigurationErrorsException - Configuration system failed to initialize - Root element is missing	T105806 Swedish Wikipedia (sv.wp) specific header changes	
T91080 Document/annotate ListProviders whether they return BIG or SMALL query results	T99317 [Migrated] ExplicitCapture	T101152 [Migrated] AWB Updater: ZipException in ZipFile.ReadEntries	T105348 Multiple line breaks in {{cite web}}	
T109757 During run, Stop after redirected file drops the file from subsequent run	T100583 [Migrated] Title case for citations	T101040 InvalidOperationException in UserPrefs.LoadPrefs	T100608 [Migrated] Better support for links to web.archive.org in cite templates	
	T99310 [Migrated] Unable to undo changes via diff double click	T99334 [Migrated] AWB does not start on Windows XP with .NET Framework 2.0 or 3.0 (Custom Module error)	T100582 [Migrated] Do not include unimportant prefixes in {{DEFAULTSORT}}	
	T99273 [Migrated] External processing error		T100581 [Migrated] When maintenance templates immediately follow section header, change to section template	
	T99275 [Migrated] Wikia: This page is			

Acknowledgments

- The speaker received a WMF grant to travel and participate in this Conference
- Slides are available on commons