FOSSCOMM 2017

# Mediawiki and AutoWikiBrowser

Marios Magioladitis

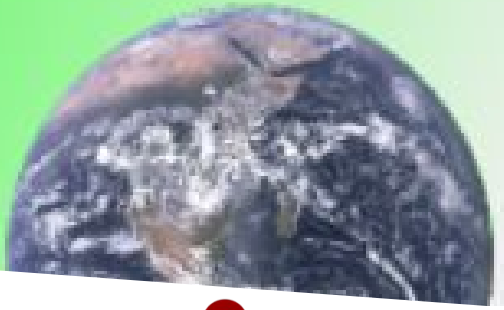
WIKIMEDIA
Community User Group
GREECE

05.11.2017

# AutoWikiBrowser

The semi-automated wiki editor

Version 5.9.0.1



## Mediawiki and AutoWikiBrowser

Marios Magioladitis

WIKIMEDIA
Community User Group
GREECE

05.11.2017

# Wikipedia:

- Created in 2001 by Larry Sanger and Jimmy Wales
- Owned by non-profit Wikimedia Foundation
- Anyone can edit
- 40 million articles in 299 languages
- 25k-60k pages requests per second
- It is accessed via website, mobile apps, search engines, etc.

# Mediawiki

- Free and open-source wiki software.
- Originally developed by Magnus Manske and improved by Lee Daniel Crocker.
- Initial release: 2002
- 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

MediaWiki has had 79.849 commits made by 776 contributors, representing 540.379 lines of code

```php
/**
 * Make URL indexes, appropriate for the el_index field of externallinks.
 *
 * @param string $url
 * @return array
 */
function wfMakeUrlIndexes( $url ) {
  $bits = wfParseUrl( $url );

  // Reverse the labels in the hostname, convert to lower case
  // For emails reverse domainpart only
  if ( $bits['scheme'] == 'mailto' ) {
    $mailparts = explode( '@', $bits['host'], 2 );
    if ( count( $mailparts ) === 2 ) {
      $domainpart = strtolower( implode( '.', array_reverse( explode( '.', $mailparts[1] ) ) ) );
    } else {
      // No domain specified, don't mangle it
      $domainpart = '';
    }
    $reversedHost = $domainpart . '@' . $mailparts[0];
  } else {
    $reversedHost = strtolower( implode( '.', array_reverse( explode( '.', $bits['host'] ) ) ) );
  }
  // Add an extra dot to the end
  // Why? Is it in wrong place in mailto links?
  if ( substr( $reversedHost, -1, 1 ) !== '.' ) {
    $reversedHost .= '.';
  }
  // Reconstruct the pseudo-URL
  $prot = $bits['scheme'];
  $index = $prot . $bits['delimiter'] . $reversedHost;
```

# Mediawiki

- https://www.mediawiki.org
- Extensions and themes

# 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.

# Mediawiki

## Editing Pride and Prejudice

B I {} Advanced Special characters Help

'''''Pride and Prejudice''''' is a romance novel by [[Jane Austen]], first published in 1813. The story charts the emotional development of the protagonist, [[Elizabeth Bennet]], who learns the error of making hasty judgements and comes to appreciate the difference between the superficial and the essential. The comedy of the writing lies in the depiction of [[manners]], [[education]], and [[marriage]] and [[money]] in the [[Regency era|British Regency]].

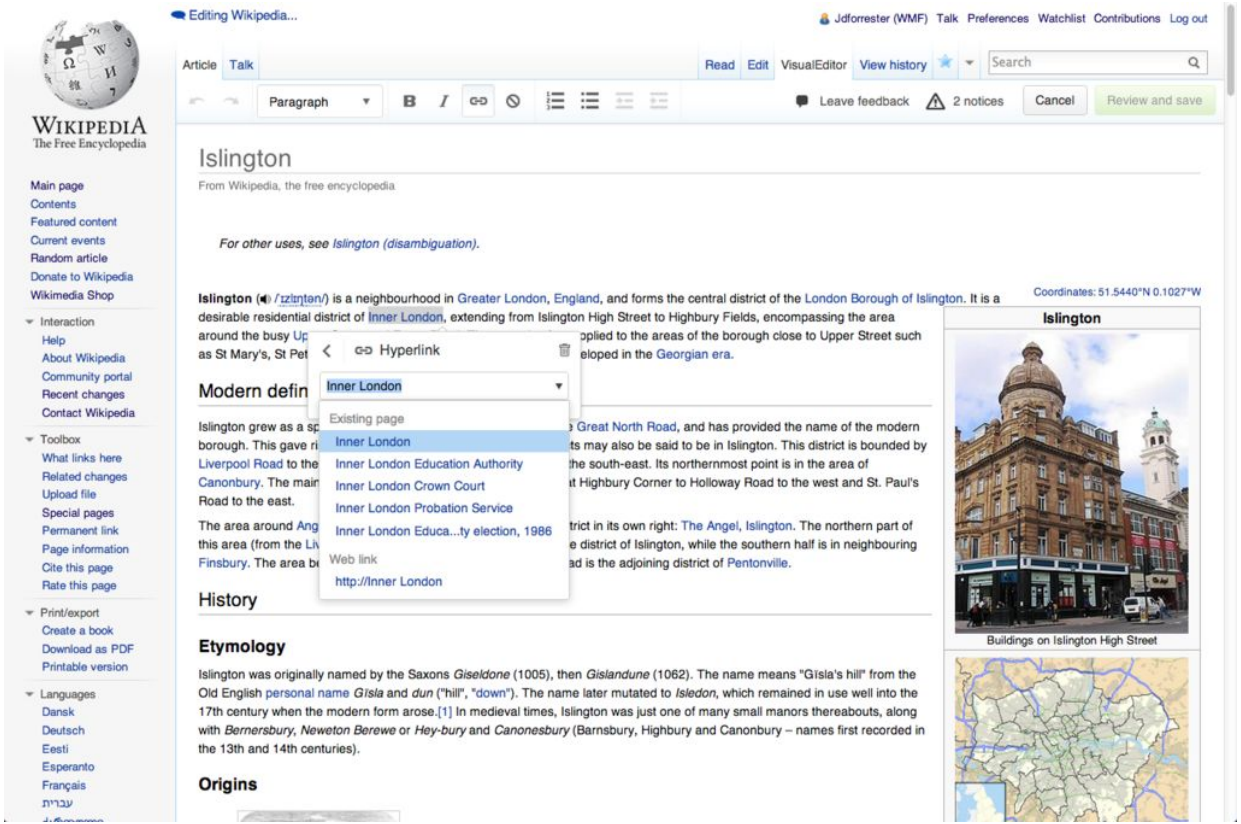
Mr Bennet of the Longbourn estate has five daughters, but his property is entailed, meaning that none of the girls can inherit it. Since his wife had no fortune, it is imperative that one of the girls marries well in order to support the others on his death. However, Jane Austen's opening line, "It is a truth universally acknowledged that a single man in possession of a good fortune must be in want of a wife" is a sentence filled with irony and playfulness. The novel revolves around the necessity of marrying for love, not simply for monetary reasons, despite the social pressures to make a good (i.e. wealthy) match.

''Pride and Prejudice'' retains the fascination of modern readers, consistently appearing near the top of lists of "most-loved books" among both [[Literary critics|literary scholars]] and the general public. It has become one of the most popular novels in [[English literature]], with over 20 million copies sold, and paved the way{{specify|date=December 2016}} for many archetypes that abound in modern literature. Continuing interest in the book has resulted in a number of dramatic adaptations and an abundance of novels and stories imitating Austen's memorable characters or themes.<ref>{{cite web|url=http://www.monstersandcritics.com/dvd/reviews/article_1475660.php/Pride_and_Prejudice_%E2%80%93_Blu-ray_Review|title=Monstersandcritics.com|date=7 May 2009|publisher=Monstersandcritics.com|archive-url=https://web.archive.org/web/20091026154330/http://www.monstersandcritics.com:80/dvd/reviews/article_1475660.php/Pride_and_Prejudice_%E2%80%93_Blu-ray_Review|archive-date=26 October 2009|dead-url=yes|accessdate=27 January 2012}}</ref> The 2005 film, ''[[Pride & Prejudice (2005 film)|Pride and Prejudice]]'', starring [[Keira Knightley]] and [[Matthew Macfadyen]] is the most recent Hollywood adaption of

# Mediawiki

MediaWiki

visual editor

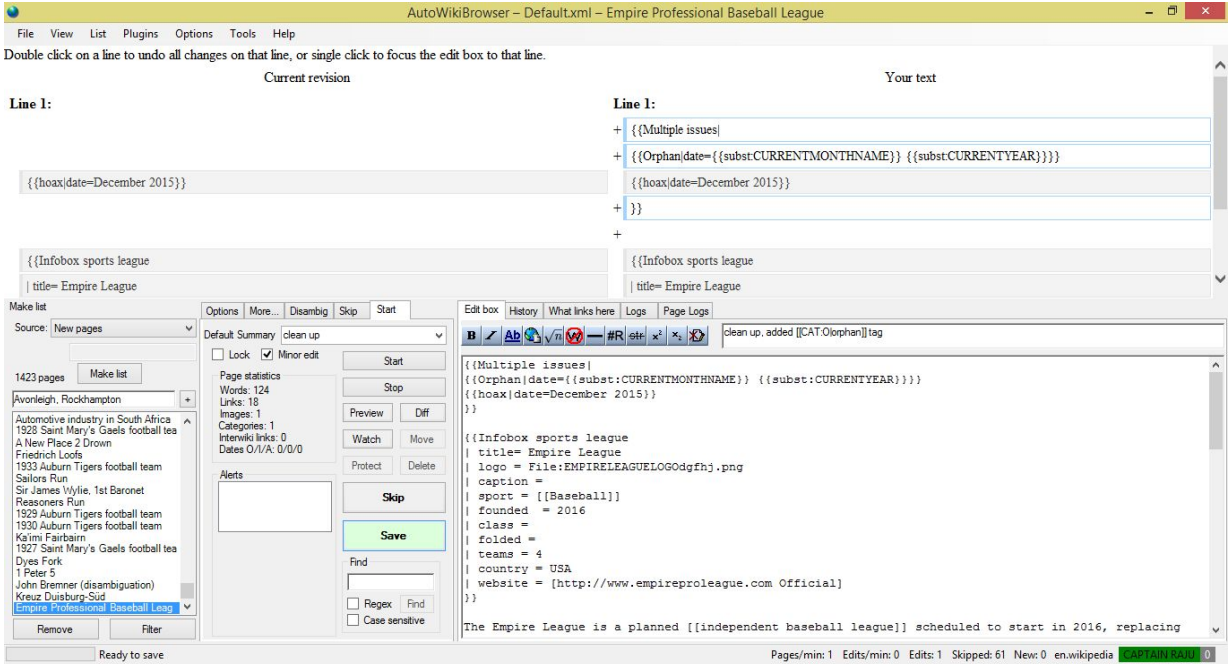


# Wikipedia:

enwiki:
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.

# AutoWikiBrowser



# 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#

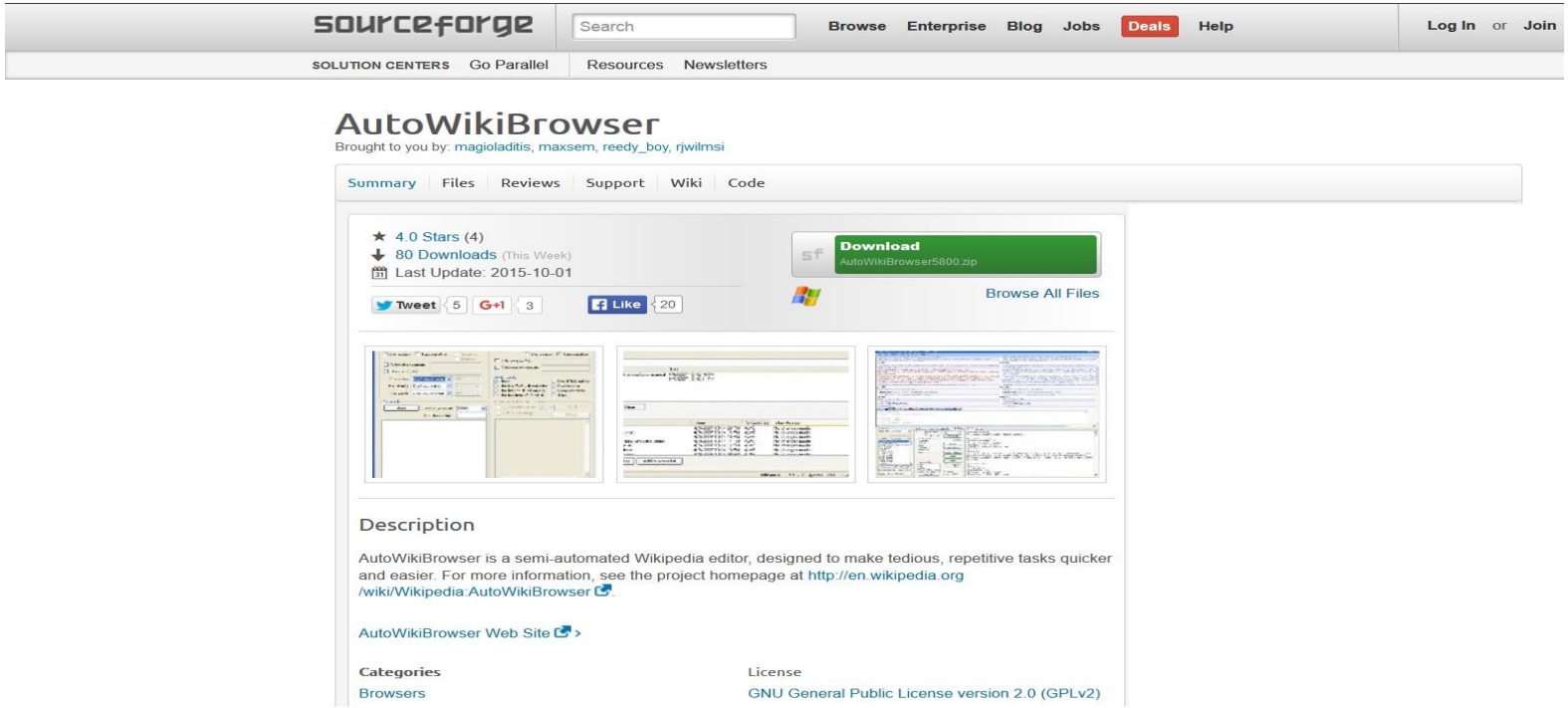
# AWB is powerful

- 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

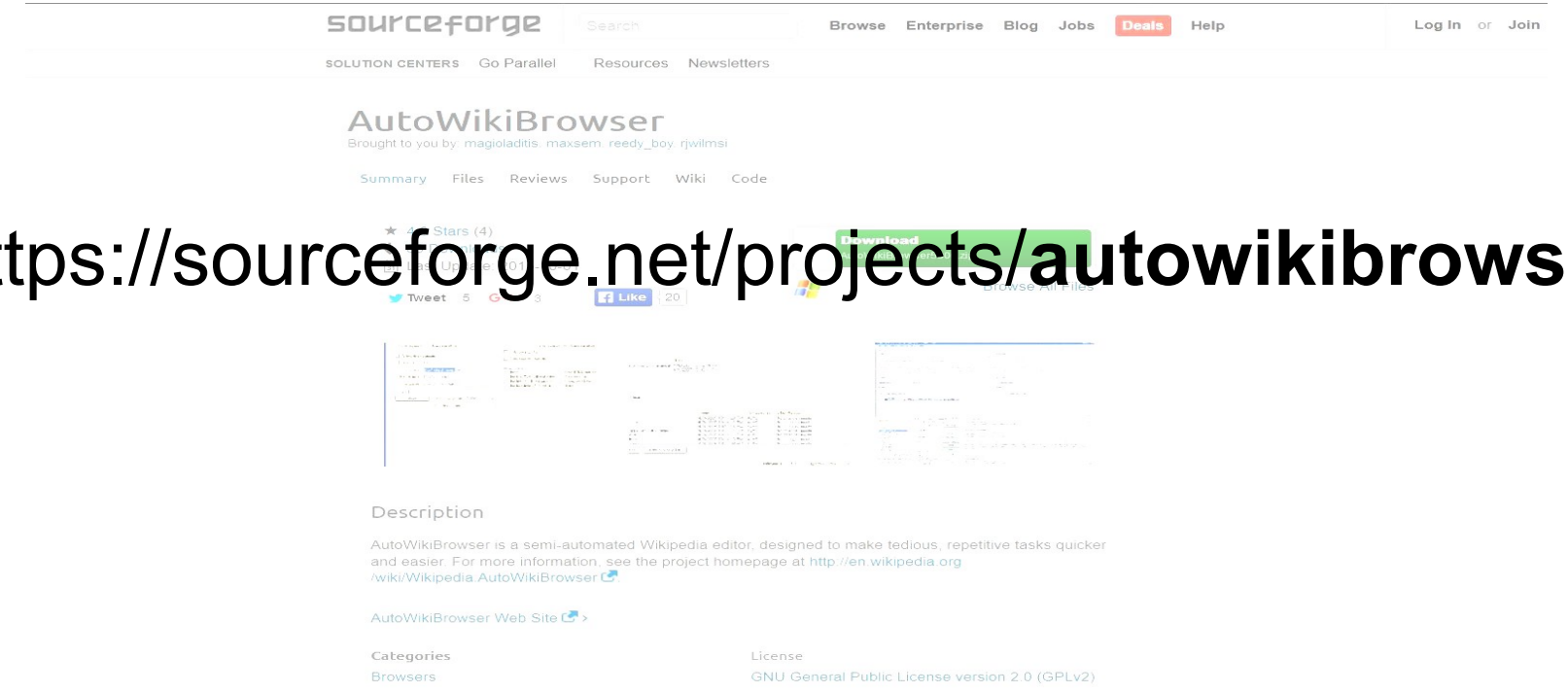
- > 3 millions lines of code [including libraries]
- 633K comments
- 1.6Mb code in C#
- 957 years of effort (COCOMO model)
- Many plugins
- Edit box supports the Microsoft Text Services Framework for use with speech recognition/handwriting applications
- Licence: GPL-2.0+
- https://www.openhub.net/p/AutoWikiBrowser

# Code is available at Sourceforge

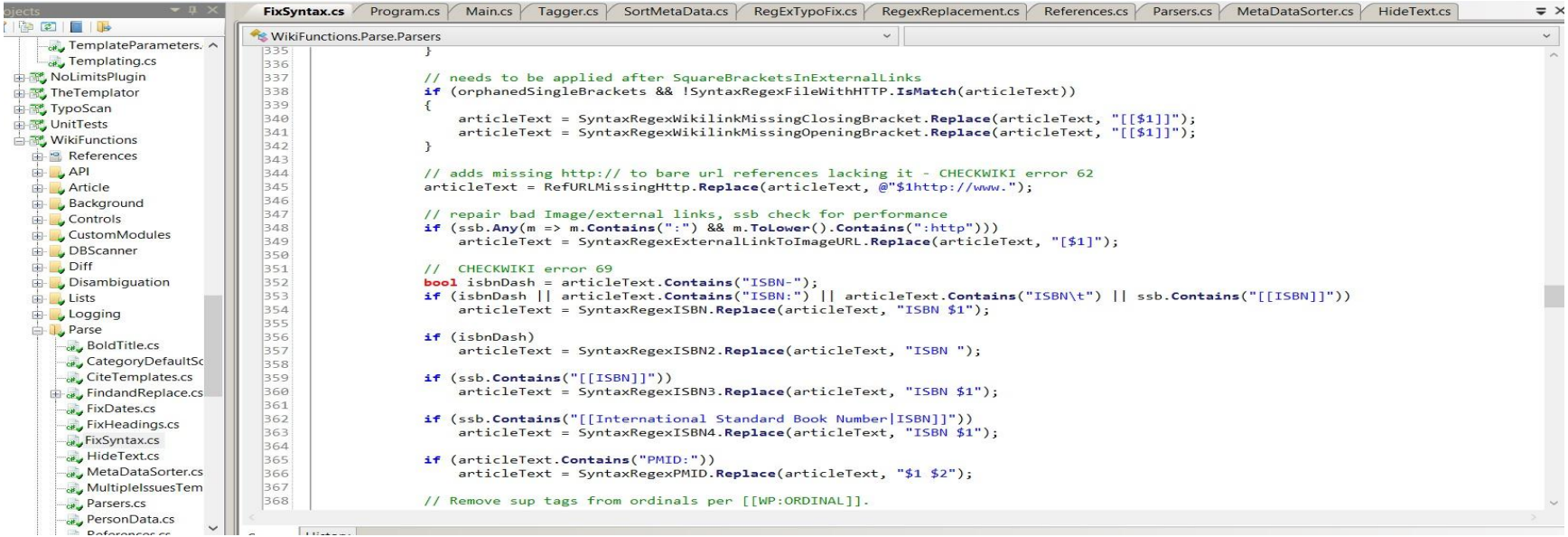


# Code is available at Sourceforge

https://sourceforge.net/projects/**autowikibrowser/**



# The code

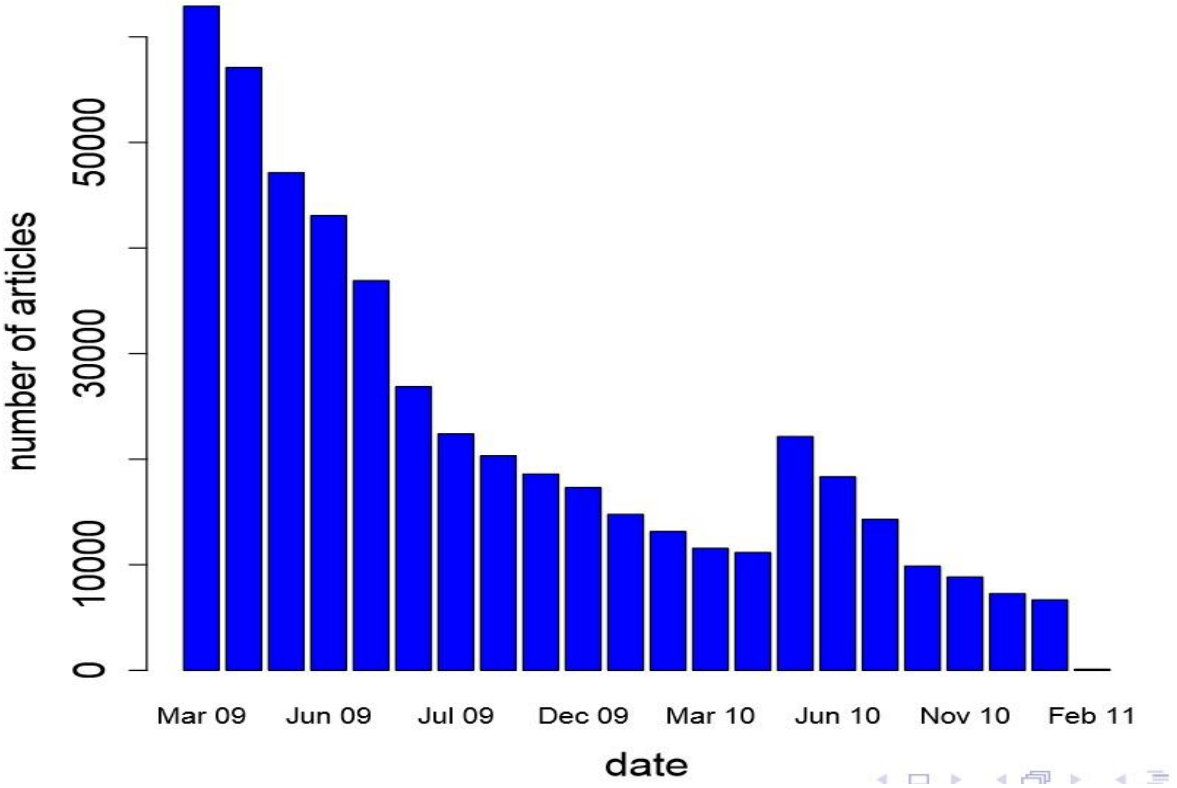

FixSyntax.cs
Program.cs
Main.cs
Tagger.cs
SortMetaData.cs
RegExTypoFix.cs
RegexReplacement.cs
References.cs
Parsers.cs
MetaDataSorter.cs
HideText.cs
WikiFunctions.Parse.Parsers
TemplateParameters.
Templating.cs
NoLimitsPlugin
TheTemplator
TypoScan
UnitTests
WikiFunctions
References
API
Article
Background
Controls
CustomModules
DBScanner
Diff
Disambiguation
Lists
Logging
Parse
BoldTitle.cs
CategoryDefaultSc
CiteTemplates.cs
FindandReplace.cs
FixDates.cs
FixHeadings.cs
FixSyntax.cs
HideText.cs
MetaDataSorter.cs
MultipleIssuesTem
Parsers.cs
PersonData.cs
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]].

# The code

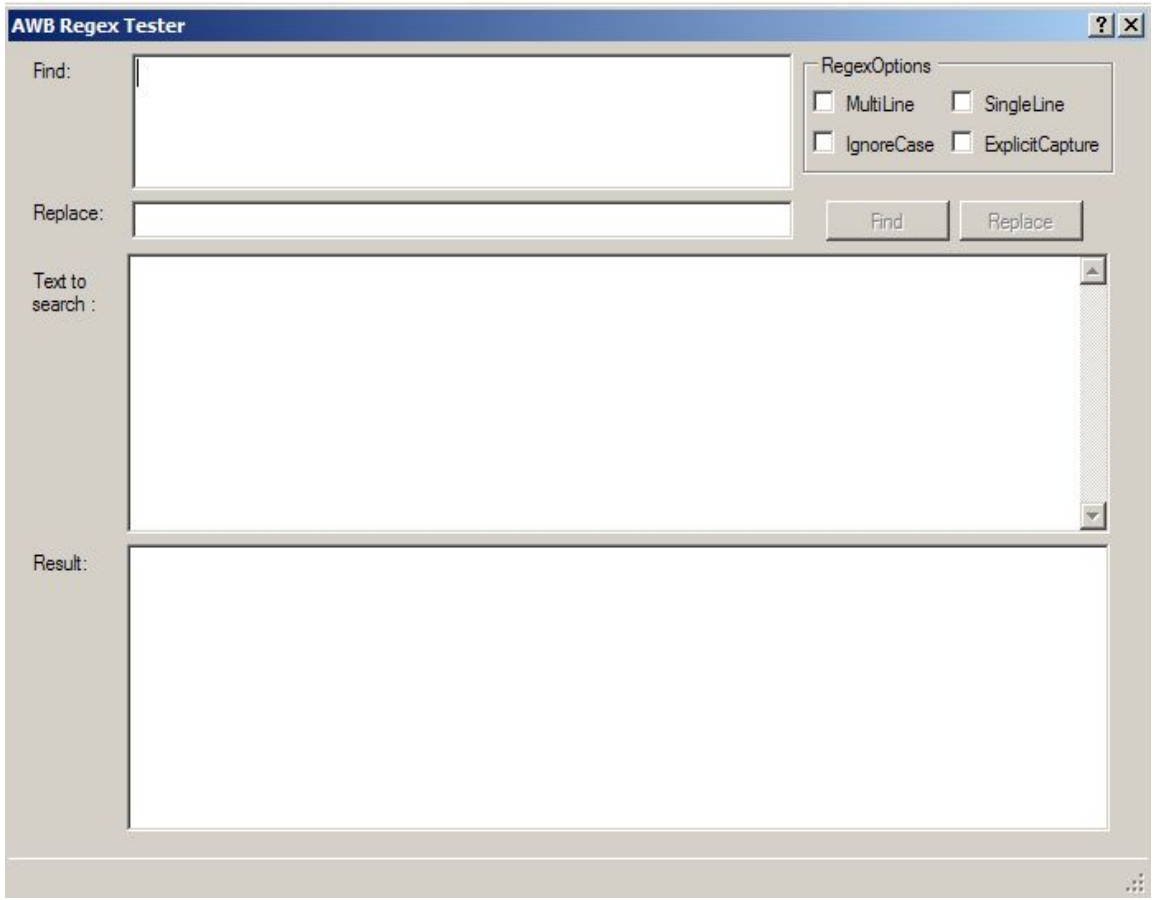
WikiFunctions.Parse.Parsers | FixSyntax(string articleText)

```
250             bool badHttpLinks = Tools.DeduplicateList((from Match m in HttpLinks.Matches(articleText.ToLower()) select m.Value).ToList()).Any(s
251
252             if (badHttpLinks)
253                 articleText = MultipleHttpInLink.Replace(articleText, "$1");
254
255             articleText = MultipleFtpInLink.Replace(articleText, "$1");
256
257             if (badHttpLinks && TemplateExists(alltemplates, WikiRegexes.UrlTemplate))
258                 articleText = WikiRegexes.UrlTemplate.Replace(articleText, m => m.Value.Replace("http://http://", "http://"));
259
260             if (badHttpLinks && !SyntaxRegexHTTPNumber.IsMatch(articleText))
261             {
262                 articleText = MissingColonInHttpLink.Replace(articleText, "$1://$2");
263                 articleText = SingleTripleSlashInHttpLink.Replace(articleText, "$1://$2");
264                 articleText = articleText.Replace("https://http://", "https://");
265             }
266
267             if (CellpaddingTypoQuick.IsMatch(articleText))
268                 articleText = CellpaddingTypo.Replace(articleText, "$1cellpadding");
269
270             if (SimpleTagsList.Any(s => s.Contains("font")))
271                 articleText = RemoveNoPropertyFontTags.Replace(articleText, "$1");
272
```

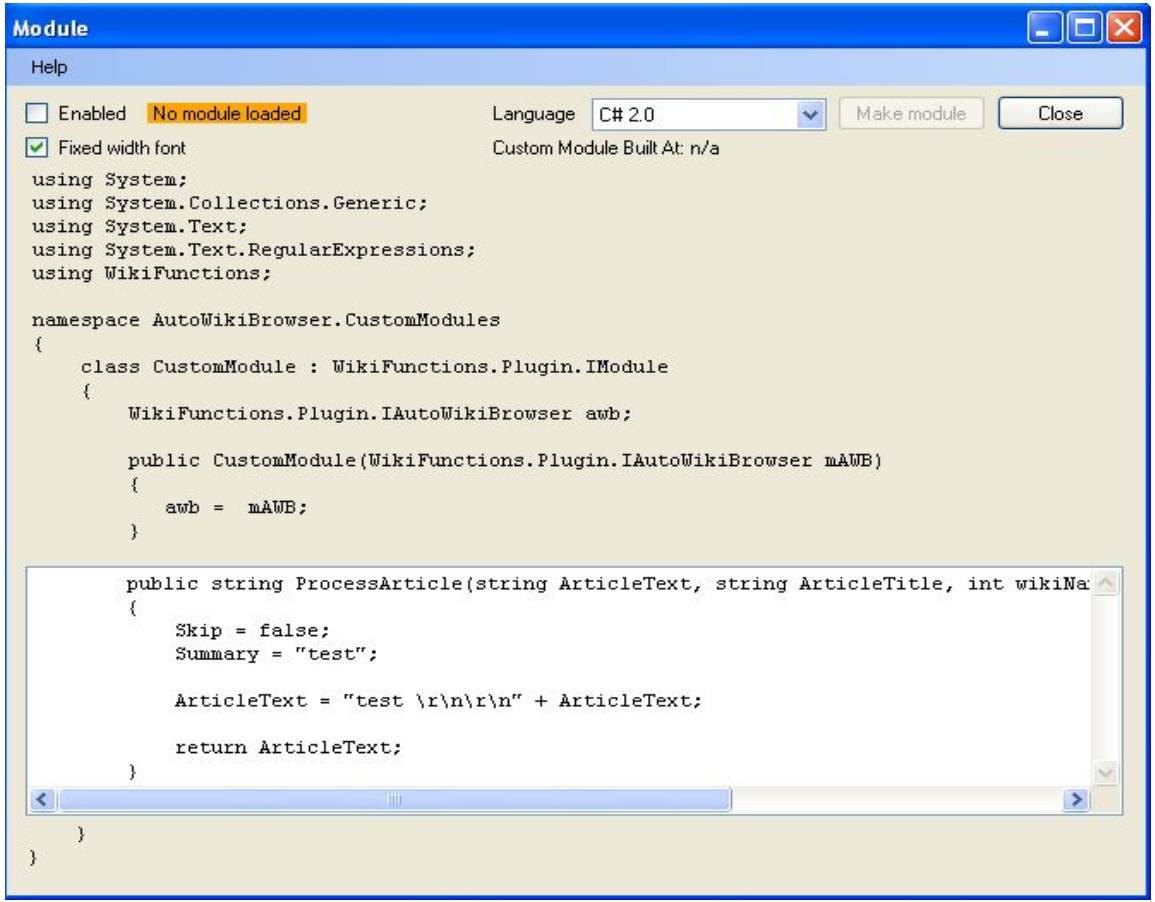
# An example: BLP tags



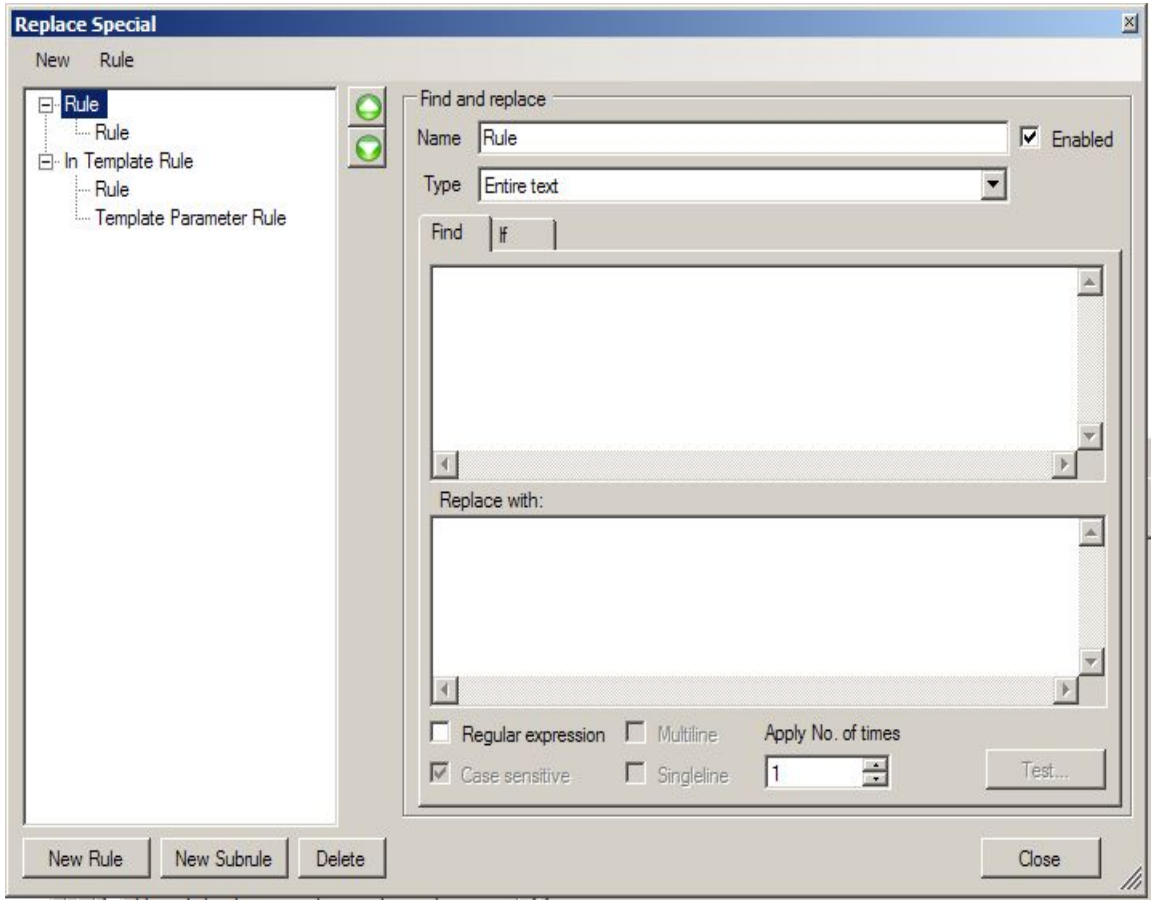
# Advanced tools



# Advanced tools



# Advanced tools



# The team

https://commons.wikimedia.org/wiki/File:AWB team @ Wikimania 2013 5.JPG (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

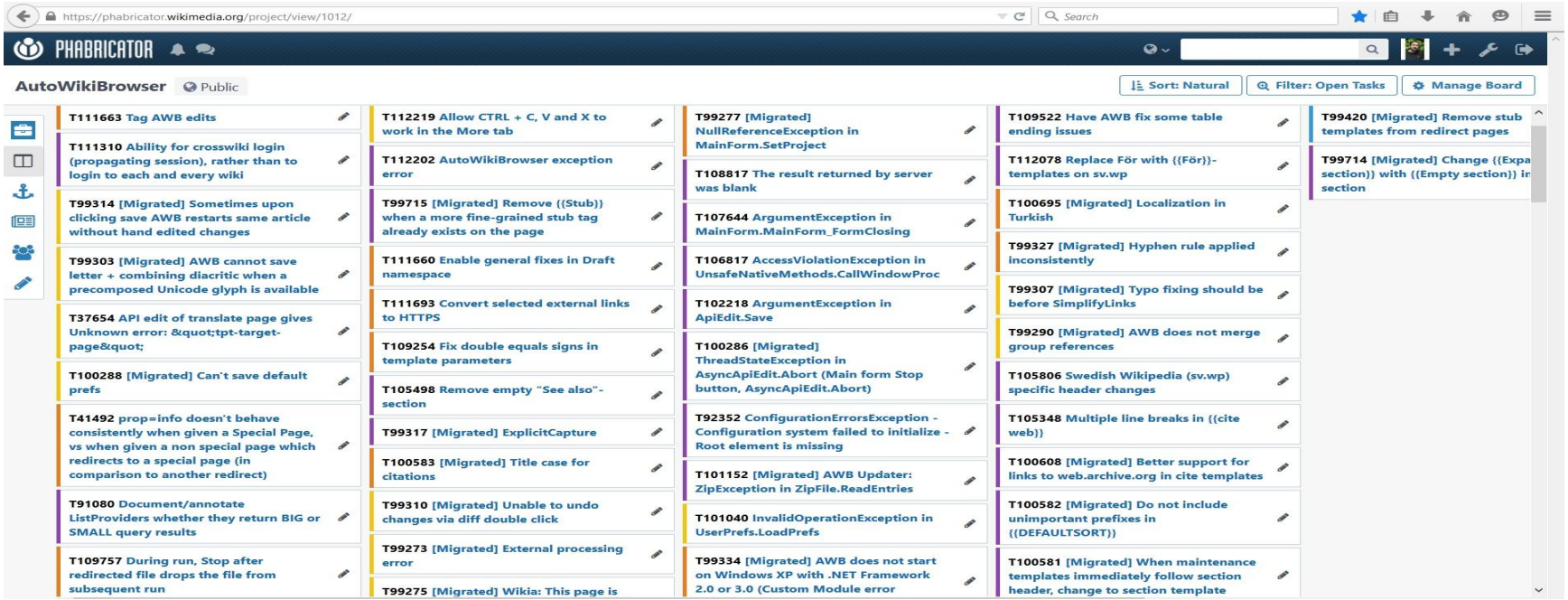
# The team

https://commons.wikimedia.org/wiki/File:AWB_team_@_Wikimania_2014_1.jpg (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

# 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!



# Acknowledgments


- The speaker is member of Wikimedia Community User Group Greece and received a WMF grant to travel and participate in this Conference
- Slides are available on Wikimedia Commons