

Creating Wikipedia articles from CC-BY content: *how hard can it be?*



Kerry Raymond

kerry.raymond@wikimedia.org.au

Copyright Kerry Raymond 2019. Released under CC-BY-4.0 licence.

File:Mad scientist caricature 2.png. (2014, December 31). *Wikimedia Commons, the free media repository*.
Retrieved 14:35, June 11, 2019
from https://commons.wikimedia.org/w/index.php?title=File:Mad_scientist_caricature_2.png&oldid=144870516.

Small/medium Wikipedias are challenged

- “As readers, many editors see the **XXXX Wikipedia as limited** ...This means that as editors, they are **less likely to contribute** to those Wikipedias because the **content gap that needs to be filled feels too large**. This perception creates a vicious cycle that prevents medium-sized wikis from reaching a critical mass of value.” – WMF Research
- We see the same problem in large language Wikipedias wrt to diversity
- Content gaps reflect contributor gaps, so the usual response is
 - Increase the number of contributors through training and edit-a-thons
 - Hard to scale, and low numbers of people become significant contributors
 - “Wikipedians are born not made”?!

Increasing content

- So let's increase the number of articles and hope that attracts new contributors
- Many contributors never create new articles (anecdotal evidence of bad experience when they first attempted and failed at doing so)
- A look at article histories shows that most edits aren't adding new content but fiddling with existing content
 - Bulk of content often written within the first few edits or by a small number of edits (and small number of contributors) over time
 - Try the "random article" experiment and see for yourself!

"If you want to end war and stuff, you gotta sing loud!"

– Arlo Guthrie in "Alice's Restaurant"


- The infinite **monkey** theorem states that a **monkey** hitting keys at random on a **typewriter** keyboard for an infinite amount of time will almost surely type any given text, such as the complete works of William Shakespeare.
- Let's stop depending on the cottage industry approach to Wikipedia and let's start a Wikipedia industrial revolution
- Let's get contributors to create more articles and add more content using pre-existing text where suitably licenced (e.g. CC-BY) or out of copyright
- *How hard can it be? – Top Gear (many episodes)*

Singing loud! What's involved?

- Manual additions of biographies based on obituaries published in the Trove digitised newspapers (pre-1955 are out of copyright)
 - Copy and paste, rewrite to sound less “Victorian era” (and more MoS)
 - Each paragraph cites the newspaper article
- Random content from Qld Govt webpages under CC-BY licence, mostly histories of Qld Indigenous communities (content gap in en.WP)
 - Copy and paste, less rewrite involved, but original text included its own citations had to be rewritten as Wikipedia citations
 - Each para cites the QG webpage
 - Add an attribution to the References section to the webpage (required by BY)
- Faster but still a lot of manual work involved in writing lede paras, adding wikilinks, infoboxes, categories, adding photos, rewriting citations and other article infrastructure

Singing with an amplifier!

- Queensland Heritage Register was released under a CC-BY license
 - Potential for 1600 new articles, additions to 200 existing articles (daunting!)
 - Had to speed up the process with machine-generation to do as much of the task of writing whole articles as possible
 - Natural language is hard for a computer to read/write
 - Embedded in HTML, XML, and other formats and must first be “scraped” out
 - Always intended for a human to upload the article to Wikipedia and copyedit but desirous to get it right as much as possible to reduce human effort
 - Extensive use of heuristics: “any approach to problem solving or self-discovery that employs a practical method, not guaranteed to be optimal, perfect, logical, or rational, but instead sufficient for reaching an immediate goal” – Wikipedia
 - Risk management approach: probability of getting it wrong * cost of fixing it
 - Need project management to track everything (lots of spreadsheets!)


Queensland Government
Contact us

For Queenslanders
Business and industry

Queensland Government home > For Queenslanders > Environment, land and water > Land, housing and property > Heritage places > Queensland Heritage Register > Search the register > Men's Toilet, Russell Street, Toowoomba


Search the register

- > List all places
- > Show all places on a map
- > Heritage explorer

Men's Toilet, Russell Street, Toowoomba

601381 Russell Street, Toowoomba


General
Significance
History
Description
Gallery



[More images...](#)

| | |
|-----------------|---|
| Classification | State Heritage |
| Register status | Entered |
| Date entered | 6 June 1994 |
| Type | Health and care services: Public toilet |
| Theme | 6.3 Building settlements, towns, cities and dwellings: Developing urban services and amenities |

Location



Men's Toilet, Russell Street, Toowoomba

From Wikipedia, the free encyclopedia

Coordinates: 27°55′S 151°95′E﻿ / ﻿27.559°S 151.9517°E﻿ / -27.559; 151.9517

Men's Toilet is a heritage-listed **public toilet** at Russell Street, **Toowoomba**, **Queensland**, Australia. It was built in 1919. It was added to the **Queensland Heritage Register** on 6 June 1994.^[1]

Contents [hide]

- 1 [History](#)
- 2 [Description](#)
- 3 [Heritage listing](#)
- 4 [References](#)
 - 4.1 [Attribution](#)
- 5 [External links](#)

History [\[edit\]](#)

The urinal was built in 1919 and is situated in Russell Street opposite the **Toowoomba railway station**. The immediate social context for the urinal is to be found in the relevant history of Toowoomba. By the late nineteenth century, the lack of sanitation in Toowoomba became a pressing local issue. Earth closets were in a disgraceful condition and **typhoid fever** was prevalent. It was in these circumstances that Toowoomba's health program, and that of Queensland generally, began to take shape. Boards of Health were established but no real success was achieved until the first years of the twentieth century. Despite improved sanitary services, Toowoomba had 98 cases of enteric fever (typhoid) in 1916 although by 1918 this had dropped to 17. The foremost mover in these plans was Dr Thomas Arthur Price, public health advocate and mayor of Toowoomba in 1919. The lack of a **sewerage system** had a major effect upon the provision and type of sanitation,

Men's Toilet, Toowoomba



Men's Toilet, 2012, prior to its demolition in 2013 and reconstruction in 2015

| | |
|----------------------|---|
| Location | Russell Street, Toowoomba, Queensland, Australia |
| Coordinates | 27°55′S 151°95′E﻿ / ﻿27.559°S 151.9517°E﻿ / -27.559; 151.9517 |
| Design period | 1914 - 1919 (World War I) |
| Built | 1919 |






Queensland Heritage Register

Official name: Men's Toilet, Russell Street, Toowoomba

```

171 <h2>General</h2>
172 <div class="figure"><div class="caption"> <a href="#tab-images" class="opentab">More images&hellip;</a></div></div>
<dl class="gridlist">
173 <dt>Classification</dt>
174 <dd>State Heritage</dd>
175 <dt>Register status</dt>
176 <dd>Entered</dd>
177 <dt>Date entered</dt>
178 <dd>6 June 1994</dd>
179 <dt>Type</dt>
180 <dd>Health and care services: Public toilet</dd>
181 <dt>Theme</dt>
182 <dd>6.3 Building settlements, towns, cities and dwellings: Developing urban services and amenities</dd>
183 <dt>Construction period</dt>
184 <dd>1919, Men's Toilet, Russell Street, Toowoomba (1919 - 1919)</dd>
185 <dt>Historical period</dt>
186 <dd>1914-1919 World War I</dd>
187 </dl>
188 <h3>Location</h3><dl class="gridlist">
189 <dt>Address</dt>
190 <dd>Russell Street, Toowoomba</dd>
191 <dt><acronym title="Local government area">LGA</acronym></dt>
192 <dd>Toowoomba Regional Council</dd>

```

B I      **Advanced** **Special characters** **Help** **Cite**

```

<!-- Article title: '''Men's Toilet, Toowoomba''' siteId: 16144 placeRef:601381 -->
{{Use Australian English|date=October 2014}}
{{Use dmy dates|date=October 2014}}
{{Infobox historic site
| name = Men's Toilet, Toowoomba
| image =Men's Toilet, Russell Street, Toowoomba (2012).jpg
| caption =Men's Toilet, 2012, prior to its demolition in 2013 and reconstruction in 2015
| locmapin = Queensland#Australia
| map_caption =
| coordinates = {{coord|-27.559|151.9517|region:AU-QLD_type:landmark|display=inline,title}}
| location = Russell Street, [[Toowoomba, Queensland|Toowoomba]], [[Queensland]], Australia
| beginning_label = Design period
| beginning_date = 1914 - 1919 (World War I)
| built = 1919
| built_for =
| demolished =
| architect =
| architecture =
| owner =
| designation1 = Queensland Heritage Register
| designation1_offname = Men's Toilet, Russell Street, Toowoomba
| designation1_type = state heritage (built)
| designation1_date = 6 June 1994

```

Getting the information in

- Writing the scraping software:
 - tedious, messy, imperfect process
 - heuristic: better too much extracted than too little
 - extremely dependent on the input format, not re-usable on other projects
 - E.g. sub-headings were text no different to other content (heuristic: short lines at the start of paragraph are sub-headings)
 - Different methods for citations and references in the heritage register entries
 - Sometimes you can't scrape everything – instead warn the human to do a check and scrape manually if needed
 - Used Python's Beautiful Soup module (speaks HTML and XML) and Python modules for working with Shape files (for geo-coordinate calculations from the heritage boundaries expressed as polygons)

Making sense of the information

- The main sections about history, description, and satisfaction of heritage criteria were “plain text”, but they need:
 - Subheadings
 - Wikilinks to other articles
 - Conversion and re-use of citations
 - Cross-references to other heritage properties (unique format QHR1234)

What's in a name? – William Shakespeare

- Wikipedia article titles must be manually chosen in many cases
 - Heritage articles are unique by ID not by title (as Wikipedia articles are) and many heritage titles are unsuitable for Wikipedia titles, e.g. “House”, “Anglican Church”, or name of short-term occupant at time of listing “Ray White Real Estate”
 - But the presence of cross-references between heritage entries means article titles need to be all chosen up-front so cross-references can be implemented as wikilinks in the Wikipedia article
 - The roll-out person can rename the article at roll-out (creates a redirect so the cross-reference wikilinks will still work)

That which we call a rose by any other name would smell as sweet -William Shakespeare

- A rose or a toilet? Does it matter?
- Need the type to write the lede para, fill in an infobox, and create categories
 - A number of “facts” were available, some values came from controlled vocabularies, some were free text
 - Create spreadsheets to map controlled vocabularies and words/phrases of free text to Wikipedia type articles and categories and types within the Wikipedia universe – over 500 types
 - Some needed categories won't exist and Wikipedia doesn't tolerate redlink categories so manual creation of categories at roll-out may be required (may not be a skill of the roll-out person)

Generating articles

- Consulting with the community on the structure and appearance of the generated article
- May take weeks but fight this battle once, not article by article
- Decisions:
 - What information should go in the lede paragraph?
 - What infobox to use? What fields to include as default?
 - What sections and sub-sections should we have as standard?
 - How to cite and attribute the original material?
- People can be incredibly picky on details (no surprise there!)
- Non-IT people have unrealistic expectations of what can be generated
 - Expect absolute perfection, difficult to explain why not

[[Wikilinking]] – making links to other articles

- What text should be wikilinked?
 - In the main sections, in the lede, in the infobox values, in the citations?
- Mapping general text to Wikipedia links has been much studied , but ...
 - Instead take advantage of the limited vocabularies in each heritage section
 - History section talks about Queensland places, places, events
 - Description section talks about architectural features and construction methods
 - Use the category closure of “Queensland” and various architecture/building categories to develop a more limited set of article targets for wikilinks
 - Create spreadsheets with aliases, many of which are stripping disambiguation
 - “Maryborough, Queensland” has the alias of “Maryborough”
 - Manually add/remove articles and their aliases from the spreadsheet (learn as you go with the roll-out when links are missing or inappropriate)

[[Wikilinking]]

- Heuristic: favour longer text matches over shorter ones, e.g. match “Shire of Richmond” & “Richmond Shire” (LGA) in preference to “Richmond” (town)
- Must not “overlink” (link the same thing many times) BUT must not allow an “overlink” to match anything else either
 - “**** [[Richmond Shire]] **** [[Richmond]] Shire ****”
- The slowest part of the article generation
 - Around 18,000 Queensland articles to be potentially link to
 - Around 800 architectural/constructural concepts
 - But faster than trying to link to 5.5M all-of-Wikipedia articles

Photos

- Original CC-BY release of Qld Heritage Register did not include the photos, they were manually uploaded later when released
- NSW Heritage Register released text and photos under CC-BY except when a third party copyright owner was identified for a photo

Bethanga Bridge

Image by: Heritage Victoria

Image copyright owner: Heritage Victoria



Bethanga Bridge

Image by: Bill Nethery

Image copyright owner: Heritage Office



Photos

With the NSW State Heritage Register:

- a webscraping tool downloads the images covered by the CC-BY licence and puts the filename and meta-data into a spreadsheet
 - about 6K out of 9K photos were CC-BY licenced
- The Pabbypan tool was used to bulk upload the photos to Wikimedia Commons (requires a set of photos and a spreadsheet with the meta-data)
- Photos in heritage entries are “galleries” not in-line, so choice of photos and positioning in the Wikipedia article must be done manually.

Photo problems

- Put photos for each heritage entry into a single Commons category called the same name as the Wikipedia article title, so need to know the article title at time of photo upload
 - Harder to rename Commons categories than renaming articles
 - but easier for roll-out if the photos are already available for inclusion
 - Visit each article only once, instead of twice
- Need to generate the Commons category description and super-categories and en.WP and Commons have different category systems
 - Another type mapping problem
 - Cheated and reused the en.WP category mappings
 - Mostly works, apart from “Railway stations” (en.WP) vs “Train stations” (Commons)

No CC-BY sources? Generate from facts!

- Copyright relates to expression not facts
- Many copyright government databases are full of facts
- Qld town/suburb/locality article generator creates stubs using only facts from government databases and SHAPE files (polygons showing boundaries)
- Good for creating ledes, infoboxes, and schools!

Springsure is a town and a locality in the Central Highlands Region, Queensland, Australia.^{[2][3]} In the 2016 census, Springsure had a population of 1103 people.^[1]

Education [edit]

Springsure State School is a government primary and secondary (Prep-10) school for boys and girls at 55 Eclipse Street (24.1157°S 148.0885°E).^{[4][5]} In 2017, the school had an enrolment of 158 students with 21 teachers (16 full-time equivalent) and 14 non-teaching staff (8 full-time equivalent).^[6]

Our Lady of the Sacred Heart Catholic Primary School is a Catholic primary (Prep-6) school for boys and girls at Gap Street (24.1170°S 148.0926°E).^{[4][7]} In 2017, the school had an enrolment of 50 students with 7 teachers (6 full-time equivalent) and 6 non-teaching staff (2 full-time equivalent).^[8]


References [edit]

- ↑ Australian Bureau of Statistics (27 June 2017). "Springsure (SSC)". *2016 Census QuickStats*. Retrieved 20 October 2018.
- ↑ "Springsure - town in Central Highlands Region (entry 31998)". *Queensland Place Names*. Queensland Government. Retrieved 10 June 2019.
- ↑ "Springsure - locality in Central Highlands Region (entry 46976)". *Queensland Place Names*. Queensland Government. Retrieved 10 June 2019.
- ↑ "State and non-state school details". Queensland Government. 9 July 2018. Archived from the original on 21 November 2018. Retrieved 21 November 2018.
- ↑ "Springsure State School". Retrieved 21 November 2018.
- ↑ "ACARA School Profile 2017". Archived from the original on 22 November 2018. Retrieved 22 November 2018.
- ↑ "Our Lady of the Sacred Heart Catholic Primary School". Retrieved 21 November 2018.

External links [edit]

- "Springsure". *Queensland Places*. Centre for the Government of Queensland, University of Queensland.
- Town map of Springsure, 1989.

Springsure
Queensland



Coordinates 24°11′51″S 148°08′86″E﻿ / ﻿24.19750°S 148.14611°E﻿ / -24.19750; 148.14611

Population 1,103 (2016 census)^[1]

Density 7,790/km² (20,175/sq mi)

Postcode(s) 4722

Area 141.6 km² (54.7 sq mi)

Time zone AEST (UTC+10:00)

LGA(s) Central Highlands Region

State Gregory

electorate(s)

Federal Flynn

Division(s)

Localities around Springsure:

| | | |
|------------|-------------------|------------|
| Minerva | Minerva | Arcturus |
| Minerva | Springsure | Orion |
| Cona Creek | Cona Creek | Orion |
| | | Cairdbeign |

Towns and localities in the Central Highlands Region, Queensland [hide]

Albinia · Alsace · Anakie · Arcadia Valley · Arcturus · Argyll · Balcomba · Barnard · Bauhinia · Belcong · Binegang · Blackdown · Blackwater · Bluff · Bogantungan · Boolburra · Buckland · Cairdbeign · Capella · Carbine Creek · Carnarvon Park · Cheeseborough · Chirmside · Comet · Cona Creek · Consuelo · Coomoo · Coorumbene · Cotherstone · Crinum · Dingo · Dromedary · Duaringa · Emerald · Fernlees · Fork Lagoons · Gainsford · Gindie · Goomally · Goowarra · Gordonstone · Hibernia · Humboldt · Jellinbah · Khosh Bulduk · Lilyvale · Lochington · Lowesby · Lowestoff · Mackenzie · Mantuan Downs · Mimosa · Minerva · Mount Macarthur · Mungabunda · Nandowne · Oombabeer · Orion · Retro · Rewan · Rhydding · Rolleston · Rubyvale · Sapphire · Springsure · The Gemfields · Theresa Creek · Tien · Togara · Wallaroo · Wealwandangie · Willows · Willows Gemfields · Witherfield · Wooroona · Wyuna · Yamala

Main Article: Local government areas of Queensland

Categories: Towns in Queensland | Central Highlands Region

Never send a man to do a machine's job!

- The Matrix

- Developing tools to generate Wikipedia articles (or article content):
 - Benefits:
 - Creates more content! Over 4K new articles to date.
 - Makes Wikipedians much more productive
 - Achieves greater / complete coverage of topics in a space, not just the popular topics
 - Results in a more consistent style (and conforms with community consensus)
 - Creates new Wikipedians to some extent (heritage property owners, heritage consultants)
 - Downsides:
 - Does not appeal to all Wikipedians
 - Needs people who can write the tools
 - Needs a roll-out team willing to go the distance

Any questions?