Scraping Wikipedia to make links

Rebecca Barabas

Some disclaimers

This presentation is based on a personal project, and is not related to my position at the Library of Congress.

I am an intermediate coder

Initial Problem

- As a part of my Wikipedia editing, I often consult Women in Red's lists generated by Wikidata scraping
- However, there were several instances in which, as I researched someone to write a page, I found the existing Wikipedia page. This means there is a separate Wikidata page, and that resources could be better used on other pages

Image from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red/Missing_articles_by_occupation/Librarians

Wikipedia:WikiProject Women in Red/Missing articles by occupation/Rugby League

Add languages

Project page Talk | Read Edit source View history

From Wikipedia, the free encyclopedia

< Wikipedia:WikiProject Women in Red | Missing articles by occupation

WiR redlist index: *Rugby League*

Welcome to WikiProject Women in Red (WiR). Our objective is to turn **red links** into **blue ones**. Our scope is women's biographies, women's works, and women's issues, broadly construed.

This list of red links is intended to serve as a basis for creating new articles on the English Wikipedia. All new articles must satisfy Wikipedia's notability criteria; red links on this list may or may not qualify.

This table lists women Rugby League players, referees and coaches for which Wikipedia lacks a biography article. It was generated using Wikidata for Wikipedia:WikiProject Women/Women in Red. See Template:Women in Red for other lists by focus area or by country.

This list is automatically generated from data in Wikidata and is periodically updated by Listeriabot. Update the list now ISPARQL IFind images

Edits made within the list area will be removed on the next update!

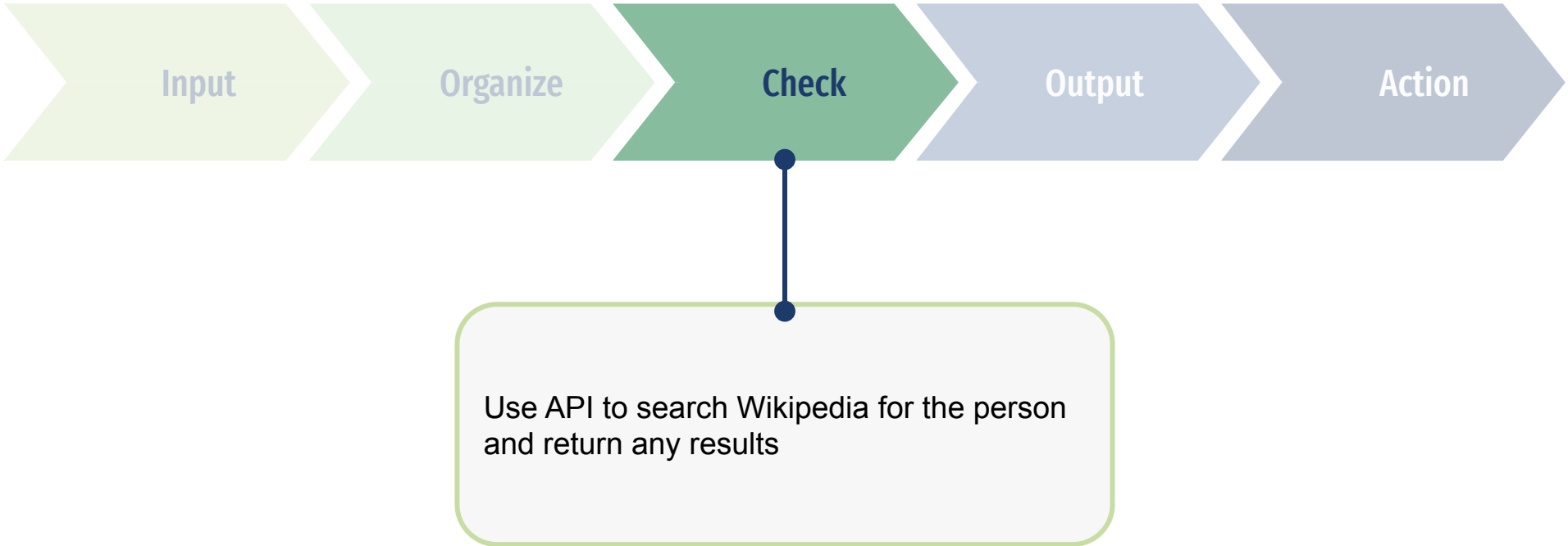
name	image	description	country of citizenship	date of birth	date of death	place of birth	place of death	item	site links
Elisa Ciria		**Elisa Ciria** **Born** -11T00:00: **Occupation(s)** rugby league player, judoka Rugby league player and judoka (*1987) ♀		1987-11-11				Q108038804	1
Cristina Song-Puche		**Cristina Song-Puche** **Born** -25T00:00: Yecla **Nationality** France, Spain **Alma mater** University of Alicante **Occupation(s)** rugby league player, rugby sevens player Rugby league player and rugby sevens player (*1989) ♀	France Spain	1989-10-25		Yecla		Q63145797	1
		Leila Bessahli **Born** -08T00:00:							

Input | **Organize** | Check | Output | Action

Create list with identifying data (Name, Wikidata code, relevant links)

```
        if type(column) == list:
          if column[0].get('_attributes').get('class') == ['new']:
            item = dict_payload[row_count]['td']
            for value in item:
              if value.get('a'):
                if value['a'][0]['_attributes']['title'].startswith('d:Q'):
                  person_list.append({
                      'name':column[0].get('_value'),
                      'wikidata_code':value['a'][0]['_value'],
                      'wikidata_link':'https://wikidata.org/wiki/'+value['a'][0]['_value']
                    })
                  break
    row_count += 1
print(person_list)
```

```
[{'name': 'Elisa Ciria', 'wikidata_code': 'Q108038804', 'wikidata_link': 'https://wikidata.org/wiki/Q108038804'},
```


Input
Organize
Check
Output
Action
Use API to search Wikipedia for the person and return any results

Input → Organize → Check → **Output** → Action

Click here for a better table

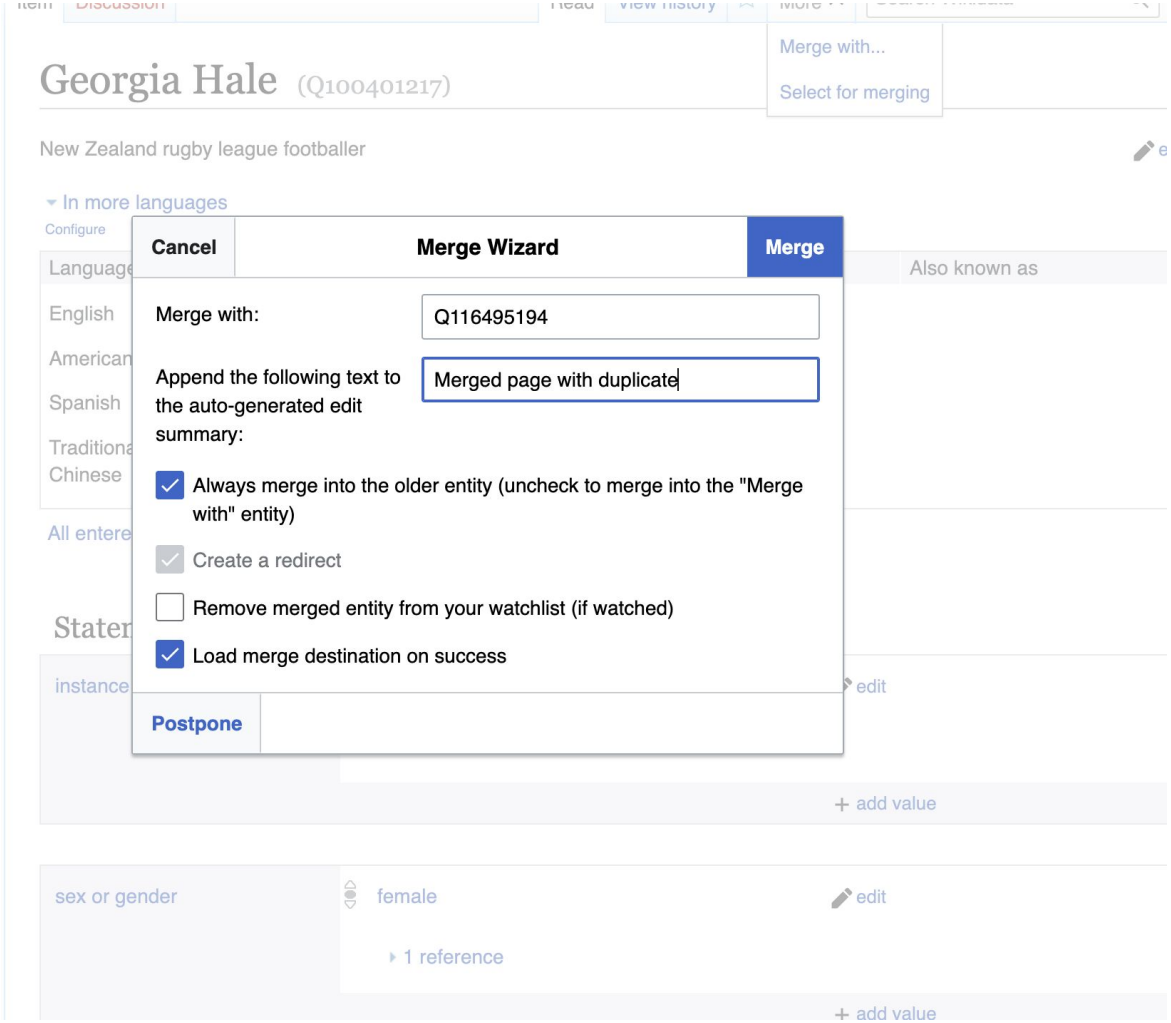
	name	wikidata_code	wikidata_link	possible_matches
0	Jane Banks	Q116475202	https://wikidata.org/wiki/Q116475202	['http://en.wikipedia.org/wiki/June_Banks']
1	Adriana Felix	Q116492978	https://wikidata.org/wiki/Q116492978	['http://en.wikipedia.org/wiki/Adriano_Félix_T...
2	Francesca Goldthorp	Q116494457	https://wikidata.org/wiki/Q116494457	['http://en.wikipedia.org/wiki/Francesca_Goldt...
3	Georgia Hale	Q116495194	https://wikidata.org/wiki/Q116495194	['http://en.wikipedia.org/wiki/Georgia_Hale', ...
4	Karen Shaw	Q116502747	https://wikidata.org/wiki/Q116502747	['http://en.wikipedia.org/wiki/Karen_Shaw_Petr...
5	Emma Slowe	Q116502906	https://wikidata.org/wiki/Q116502906	['http://en.wikipedia.org/wiki/Emma_Flower_Tay...
6	Joanna Will	Q116504483	https://wikidata.org/wiki/Q116504483	['http://en.wikipedia.org/wiki/Joanna_Williams...

Print to table

Review each person in the table. If names look like they might be the same person, perform a manual check. Merge Wikidata items.
To enable merge: Preferences → Gadgets → merge
To merge: More → Merge with → enter Wikidata code for new item

Input | Organize | Check | Output | **Action**

index	name	wikidata_code	wikidata_link	possible_matches
0	Jane Banks	Q116475202	https://wikidata.org/wiki/Q116475202	['http://en.wikipedia.org/wiki/June_Banks']
1	Adriana Felix	Q116492978	https://wikidata.org/wiki/Q116492978	['http://en.wikipedia.org/wiki/Adriano_Félix_Teixeira']
2	Francesca Goldthorp	Q116494457	https://wikidata.org/wiki/Q116494457	['http://en.wikipedia.org/wiki/Francesca_Goldthorp']
3	Georgia Hale	Q116495194	https://wikidata.org/wiki/Q116495194	['http://en.wikipedia.org/wiki/Georgia_Hale', 'http://en.wikipedia.org/wiki/Georgia_Hale_(rugby_league)', 'http://en.wikipedia.org/wiki/Georgia_Hall', 'http://en.wikipedia.org/wiki/Georgina_Hale', 'http://en.wikipedia.org/wiki/Georgia_Hase' 'http://en.wikipedia.org/wiki/Georgia-Palestine_relations']


Merge with...
Select for merging
Georgia Hale (Q100401217)
New Zealand rugby league footballer
In more languages
Configure
Language
Also known as
English
American
Spanish
Traditional Chinese
All entered
Statements
instance
edit
+ add value
sex or gender
female
edit
1 reference
+ add value
Cancel
Merge Wizard
Merge
Merge with:
Q116495194
Append the following text to the auto-generated edit summary:
Merged page with duplicate
Always merge into the older entity (uncheck to merge into the "Merge with" entity)
Create a redirect
Remove merged entity from your watchlist (if watched)
Load merge destination on success
Postpone

Pain points

01

It can be difficult to confirm that two pages are for the same person.

02

Women in Red is a pretty unique project, so there is not a lot of reusability.

03

Longer lists take a long time to review

Next Step – The Dinner Party, Judy Chicago

Image from Wikimedia.
https://en.wikipedia.org/wiki/The_Dinner_Party#/media/File:Judy_Chicago_The_Dinner_Party.JPG

List of women in the *Heritage Floor*

9 languages

Article Talk | Read Edit View history

From Wikipedia, the free encyclopedia

This list documents all 998 mythical, historical and notable women whose names are displayed on the handmade white tiles of the *Heritage Floor* as part of Judy Chicago's *The Dinner Party* art installation (1979); there is also one man listed, Kresilas, who was mistakenly included in the installation as he was thought to have been a woman called Cresilla. The names appear as they are spelled on the floor. Since 2007 the installation has been on permanent exhibition in the Elizabeth A. Sackler Center for Feminist Art at the Brooklyn Museum, New York.

Contents

A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · R · S · T · U · V · W–Z

Notes · References

This is a sortable list. Click on the column headers to reorder.

Name	Birthdate	Location	Group (Related Place Setting)	Notes
Abella of Salerno	Flourished c. 1350	Salerno, Italy	Trotula	Taught general medicine at Schola Medica Salernitana, the earliest medical school in Europe. She specialised in embryology,[1] and published two treatises.[2]
Abigail	Flourished c. 965 BC[3]	Ancient Israel	Judith	Earliest female pacifist in biblical record.[4] Her husband defied the future King David and to avoid war, she went behind her husband's back and prepared food for David, in order to keep the peace. Her husband died from the shock of her actions, and she would go on to become the third wife of David.[3][4]
Abigail Adams	1744	Massachusetts, United States	Anne Hutchinson	Abolitionist, women's education advocate, influential political figure.[5]

Screenshot of
https://en.wikipedia.org/wiki/List_of_women_in_the_Heritage_Floor

Overview

Input

URL

for the table of women on the Heritage Floor

Check

Wikidata pages

to see if The Dinner Party is mentioned

Print

CSV

of which pages have Dinner Party mentioned

Pain points

01

Attributes are not consistent across pages

02

Difficulty encountered when name on the *Dinner Party* page differs from Wikipedia page

03

There can only be one person in a cell

Thank you!

Template from Slidesgo