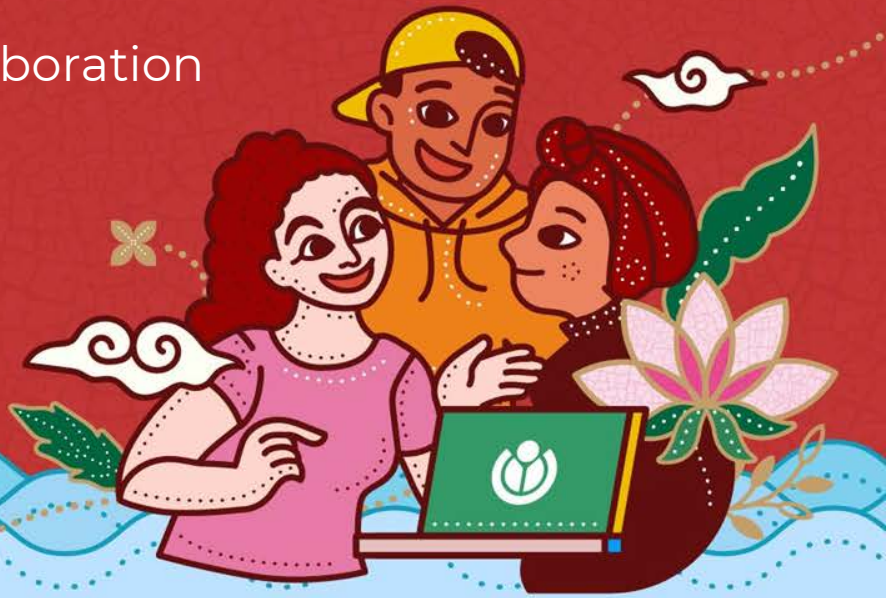
The New Zealand Thesis Project: A Nation's Dissertations

A Wikimedia and academic library collaboration

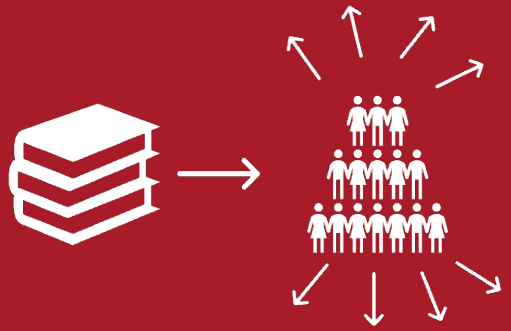
Tamsin Braisher
Deborah Fitchett



Aim: to upload bibliographic metadata from our public thesis collections to Wikidata to make them more accessible

WIKIDATA

66k theses from 13 institutions uploaded



People and main subjects connected

We can ask interesting questions about our data

What collections?

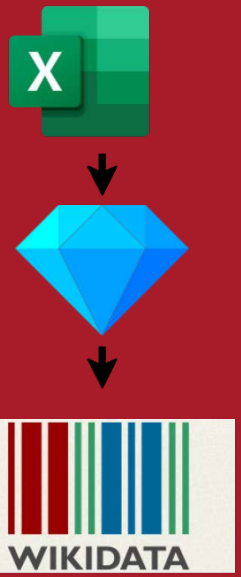
- Bibliographic data from thesis collections in institutional repositories
- Some items available for download, some as metadata only
- Data provided by all 8 universities, 5 polytechnics

What is Wikidata?

- Massive linked open database
- > 1 billion items
- > 300 languages
- Machine and human readable
- Contributes to Google searches, voice assistants, infoboxes & more
- Aggregator for identifiers

Timeline and process

- Institutional buy-in early 2022
- Data aggregation, cleaning, reconciliation March–July
- Upload July–August and QIDs returned to libraries
- Main subjects, matching of authors continue

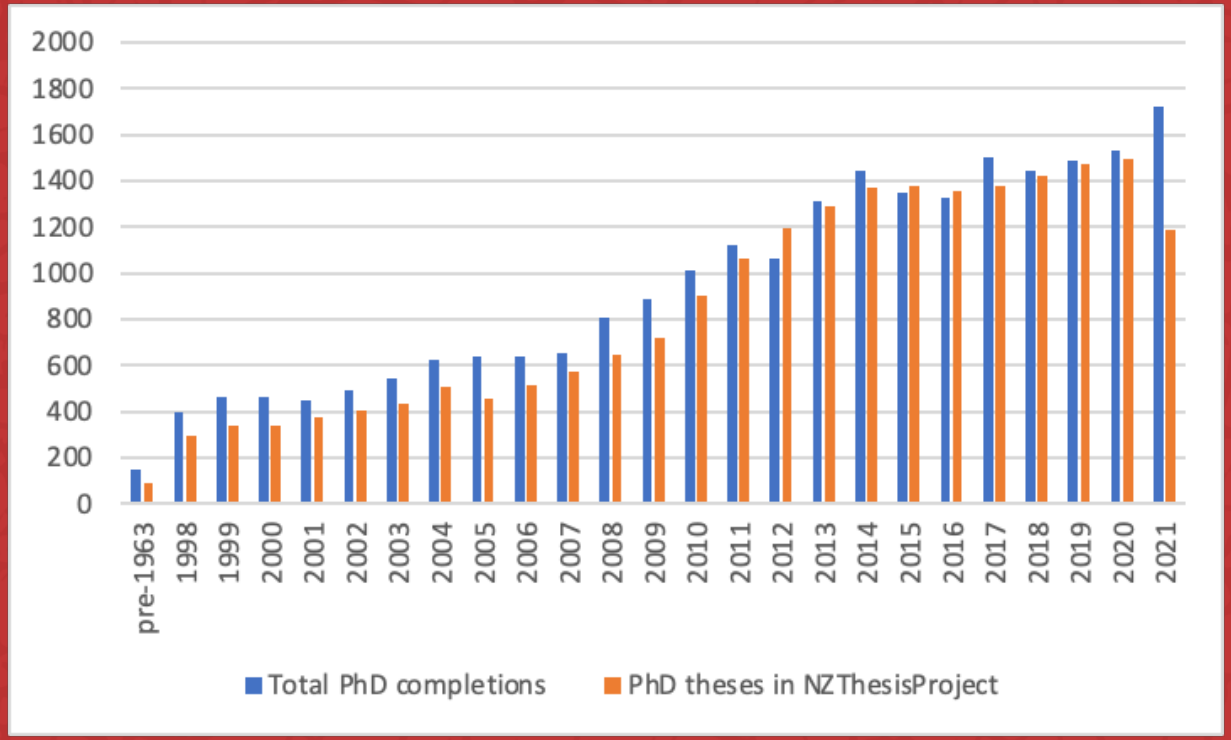


- Deborah Fitchett
  Academic librarian, Lincoln University
- Mike Dickison
  Wikimedian and librarian
- Tamsin Braisher
  Wikimedian, data wrangler
- Siobhan Leachman
  Wikimedian

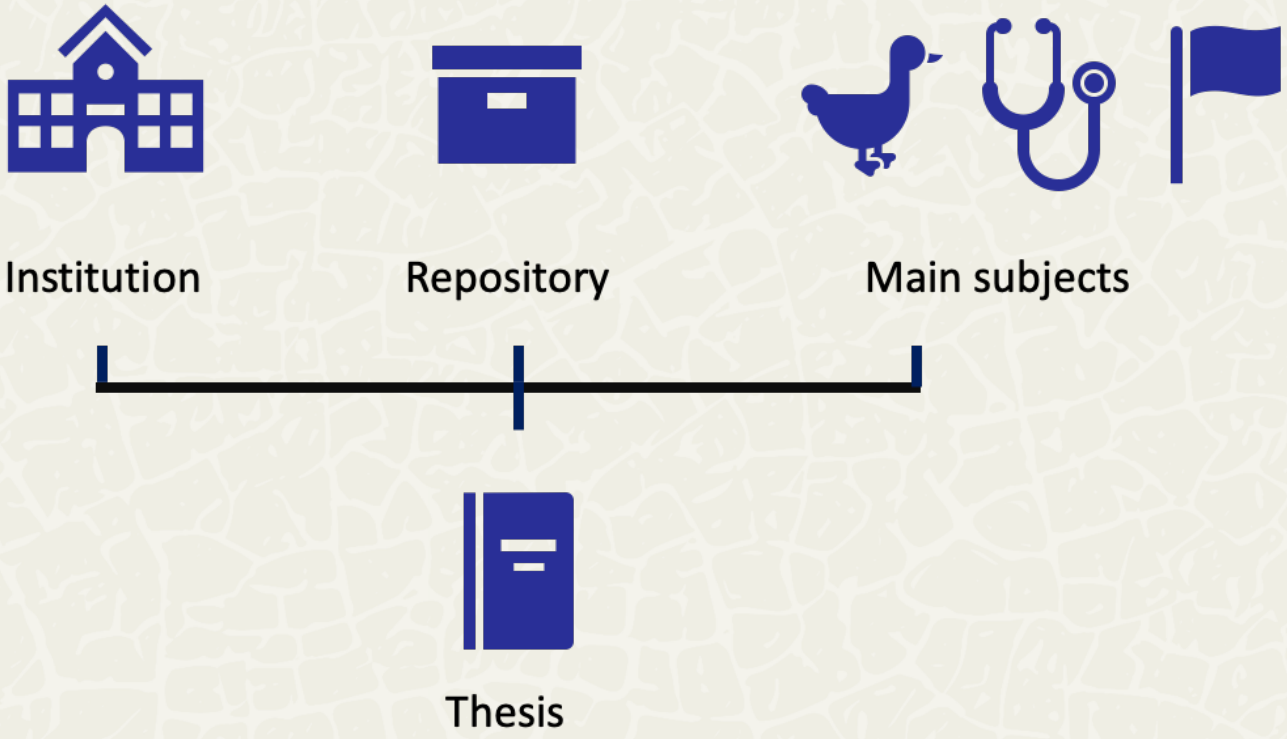
Dataset description

- >66,000 items
- 13 types of thesis
- Several languages
- Time frame 1907–2022
- Handle ID / DOI (permalinks)
- Four NZ-specific controlled vocabularies plus 280,000 rows of uncontrolled keywords
- Missing data, variations in data entry practice
- Qualifications not in Wikidata

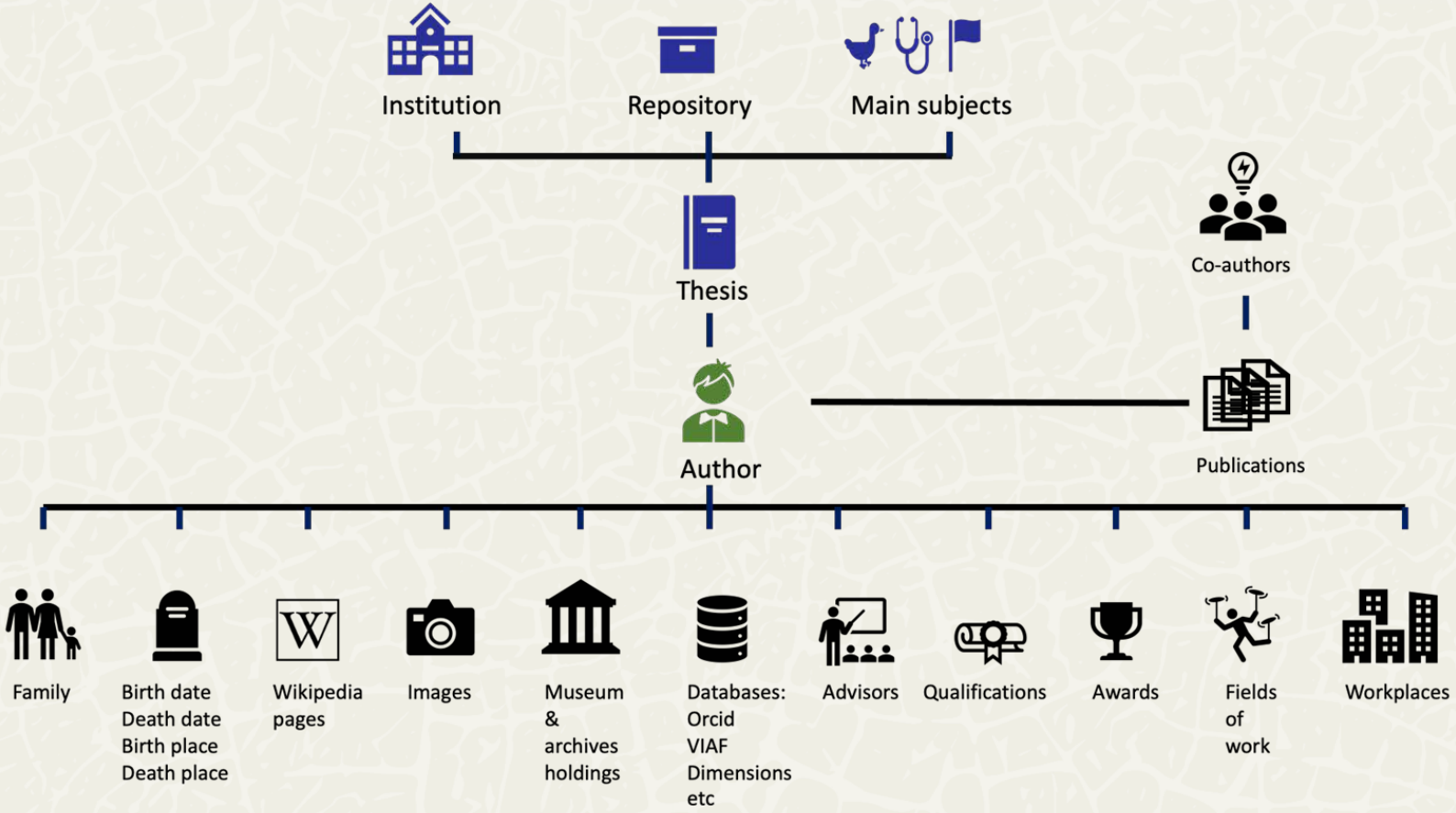
PhD thesis completeness



Thesis links in Wikidata, unlinked author



Thesis links in Wikidata, linked author

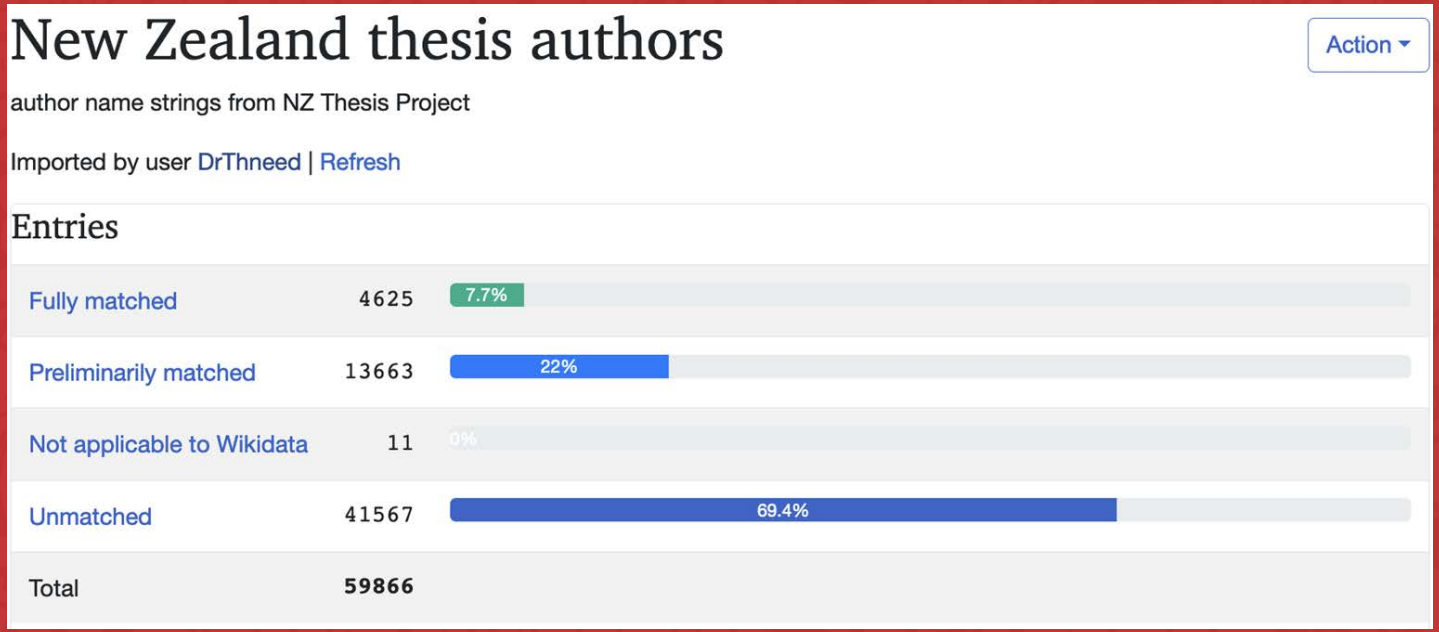


Uploading to Wikidata

- Full schema on project page on Wikidata
- Matched likely subsets of people before upload
- New people items created only when new information
- Unique Wikidata identifiers (QIDs) roundtripped - returned to institutions for inclusion in repository metadata
    - Linked author for >11,000 theses (16%)
    - Advisors on >39,000 theses (60%)
    - Main subjects for >80% of theses

Matching authors

10% of authors matched before upload of thesis
Mix'n'match catalogue for matching remaining authors



New Zealand thesis authors: C. R. Mason	author of 1987 masters thesis at University of Canterbury titled Rhizomatous legumes for hawkweed dominated grasslands	*Preliminarily matched*
Cher [Q12003]	Actress, singer, singer-songwriter, composer, film director, character actor, record producer, television actor, model, film actor, manufacturer, and recording artist (*1946) ♀; Primetime Emmy Award, Academy Award for Best Actress, Lucy Award, Grammy Awards, Kennedy Center Honors, and Palme d'Or; member of Sonny & Cher and Allman and Woman; child of John Sarkisian and Georgia Holt; spouse of Gregg Allman and Sonny Bono ↻	Confirm | Remove [all]

Matching subjects

280k rows of freetext keywords, covering 65% of the theses
Approximately 40% complete in mapping keywords to Wikidata

Four controlled vocabularies covering 13% of the theses

- ANZSRC 2008
- ANZSRC 2020
- Ngā Upoko Tukutuku
- Marsden subjects

ANZSRC codes partially mapped to Wikidata, Mixnmatch tool

All Upoko Tukutuku keywords used on theses mapped

Citing theses on Wikipedia

1536 people in the project are on English Wikipedia
Cite thesis on author page, linking back to repository
Link notable authors and advisors

Alexander Gerst

Pages in 35 languages

Beatrice Tinsley

Pages in 33 languages

What have we connected to?

TROVE

BRITISH LIBRARY

- 44 National Library records
- 364 Turnbull library records
- 454 in Czech National Library records
- 25 Online Cenotaph records
- 30 “archives held at”
- 162 in Trove
- 843 date of death
- 3000 awards
- 27 places named after

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Papers Past

Natural History Museum

Identifying public domain theses

Wikidata author records sometimes hold date of death
So we can build a query to list theses that are now out of copyright:
https://w.wiki/7YT2
And repositories can update their records accordingly.

Rights

https://researcharchive.lincoln.ac.nz/pages/rights

Access Rights

Digital thesis can be viewed by current staff and students of Lincoln University only. If you are the author of this item, please contact us if you wish to discuss making the full text publicly available.

File restricted to university members

File open access, marked as public domain

Files

Name: campbell_magsc.pdf
Size: 56.17 MB
Kind: Adobe PDF

Permalink

https://hdl.handle.net/10182/14297

Rights

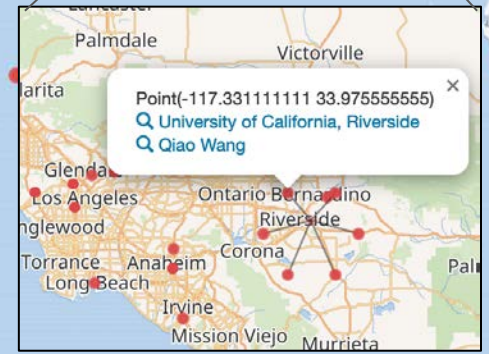
Public Domain: This work is free of known copyright restrictions in New Zealand.

Wikidata visualisations

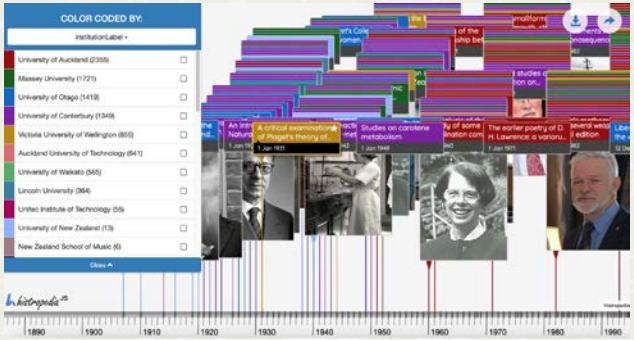
Run the example queries live at the Wikidata project page
https://www.wikidata.org/wiki/Wikidata:WikiProject_NZThesisProject

Queries can be run on the entire dataset or for a specific institution

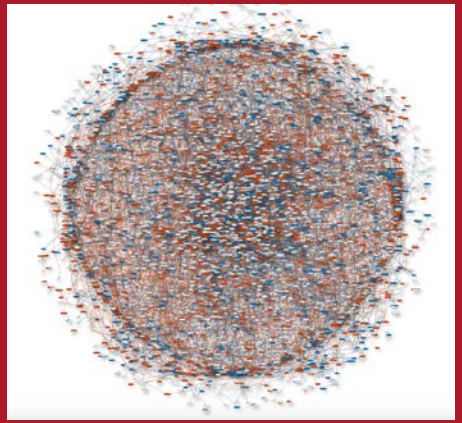
Where have people been employed?



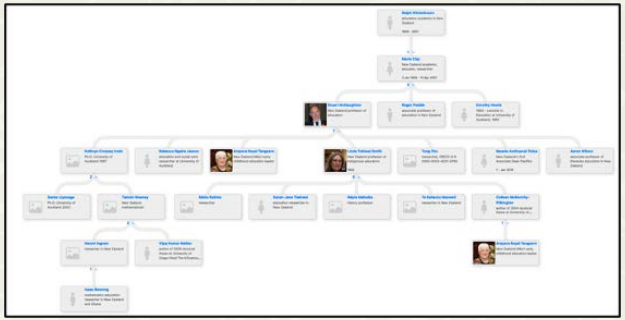
Sample visualisations of thesis data



Histropedia timelines



Doctoral links



Academic family trees

Scholia

organization / Q45135879 Improve data

Department of Zoology, University of Otago (Q45135879)

Table of Contents

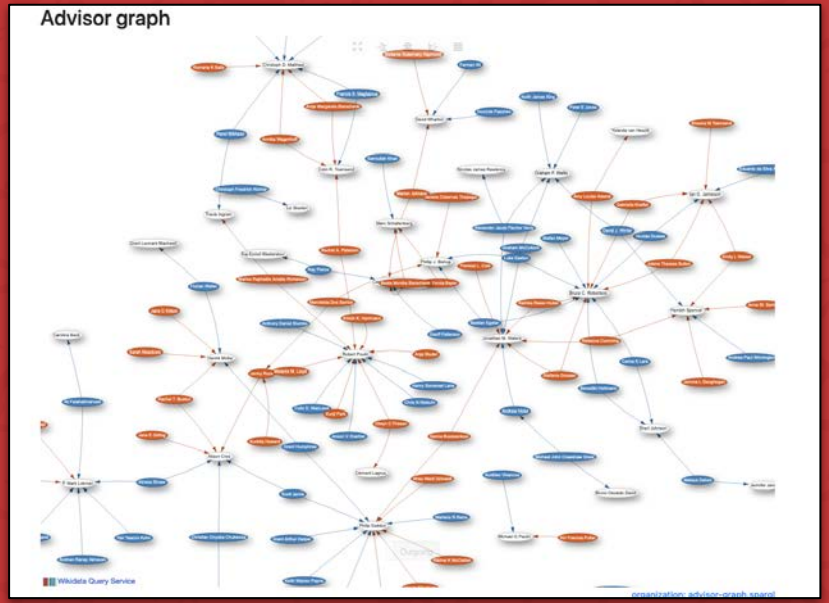
Employees and affiliated

Past and present employees, affiliated, and members

Reload

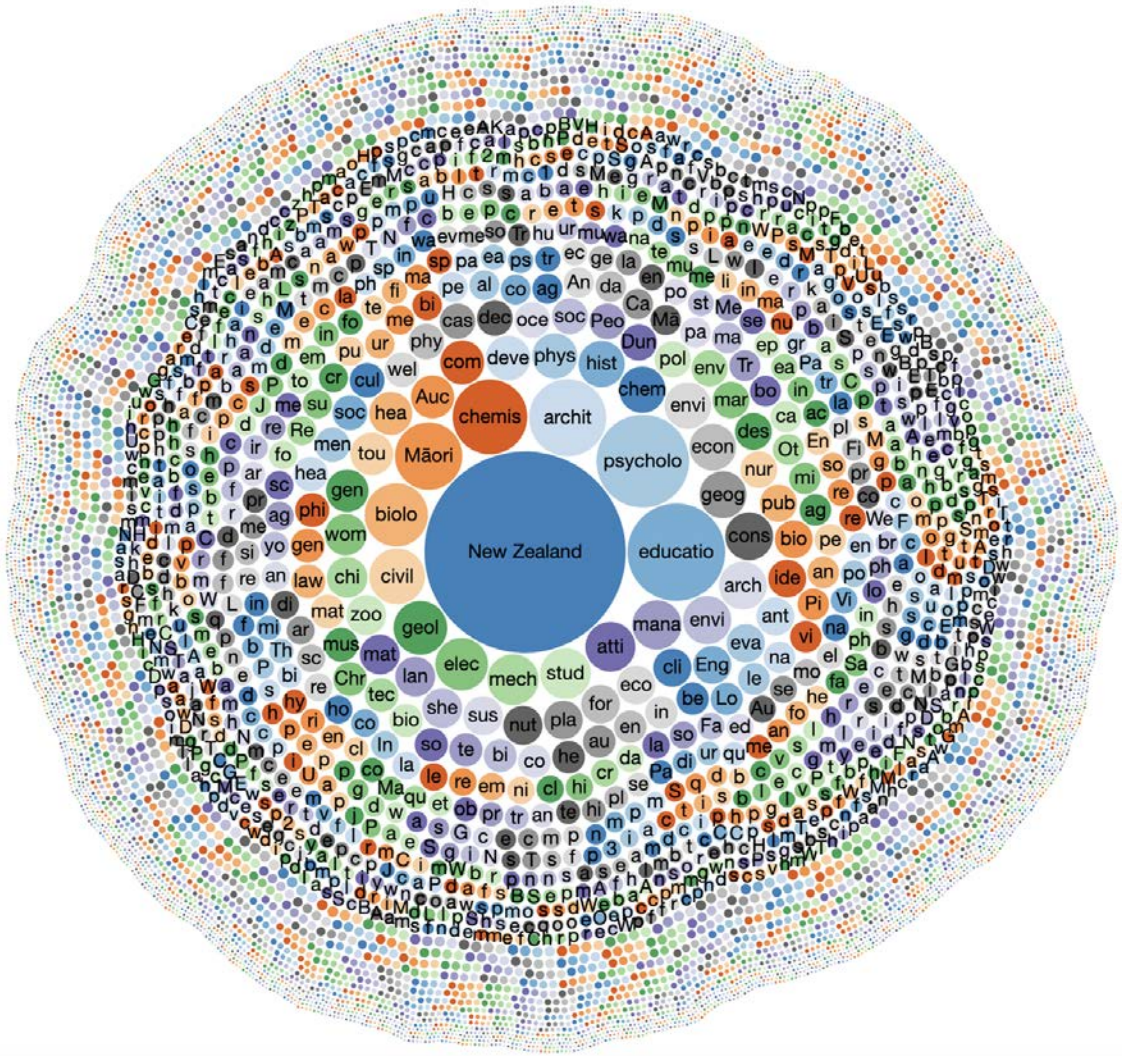
Show 10 entries Search:

Works	Wikis	Researcher	Researcher description	Orcid
518	2	Robert Poulin	evolutionary ecologist and parasitologist	0000-0003-1390-1206
216	1	Corey J. A. Bradshaw	Canadian zoologist	0000-0002-5328-7741
158	2	Hamish Spencer	New Zealand evolutionary biologist	0000-0001-7531-597X
139	1	Jonathan M. Waters	zoologist in New Zealand	0000-0002-1514-7916

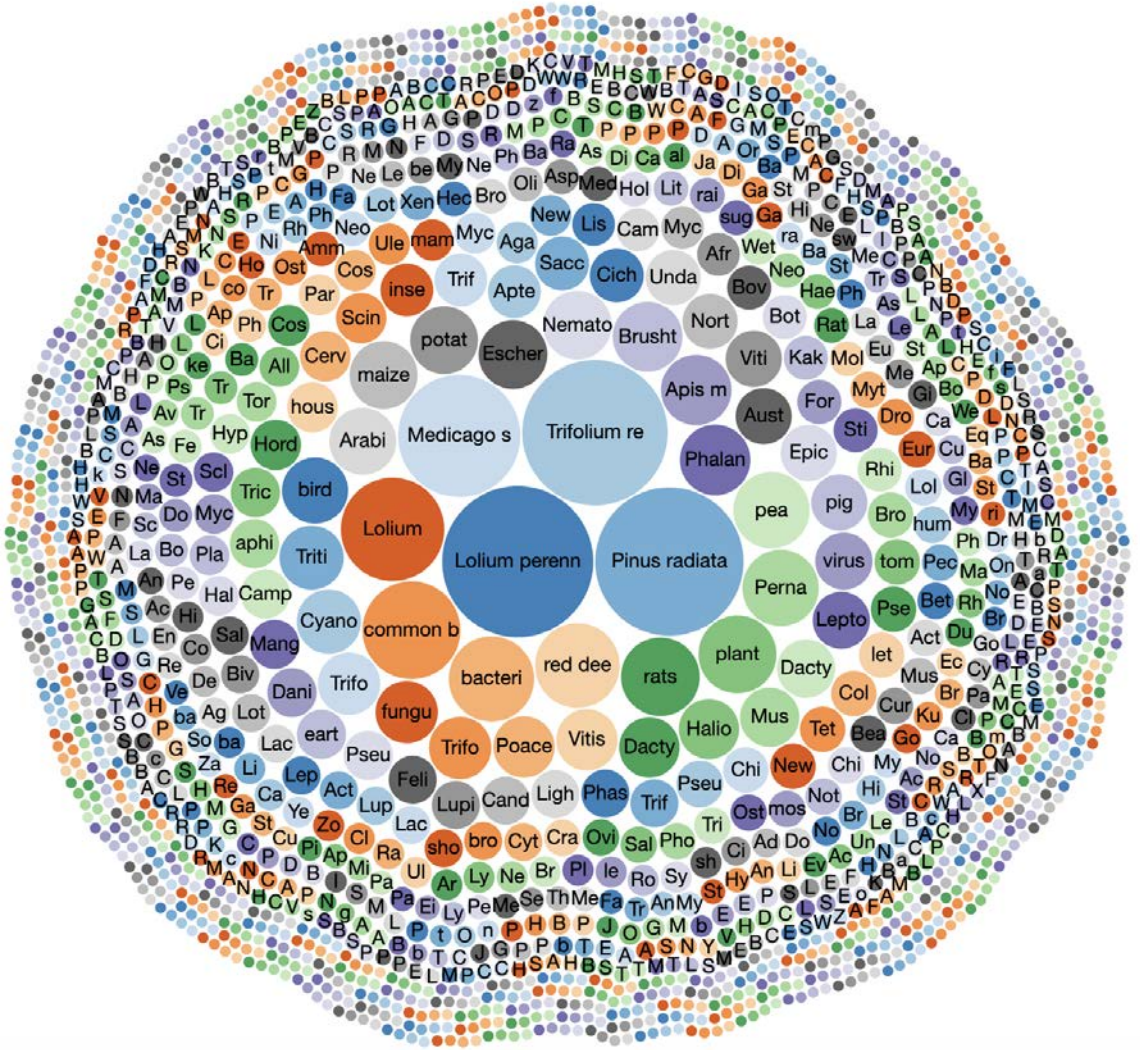


Main subjects

- You say tomato...I say *Solanum lycopersicum*... Wikidata says Q23501
- Keywords 40% mapped so far
- Ambiguous terms, very specific terms, missing Wikidata items, redundancy

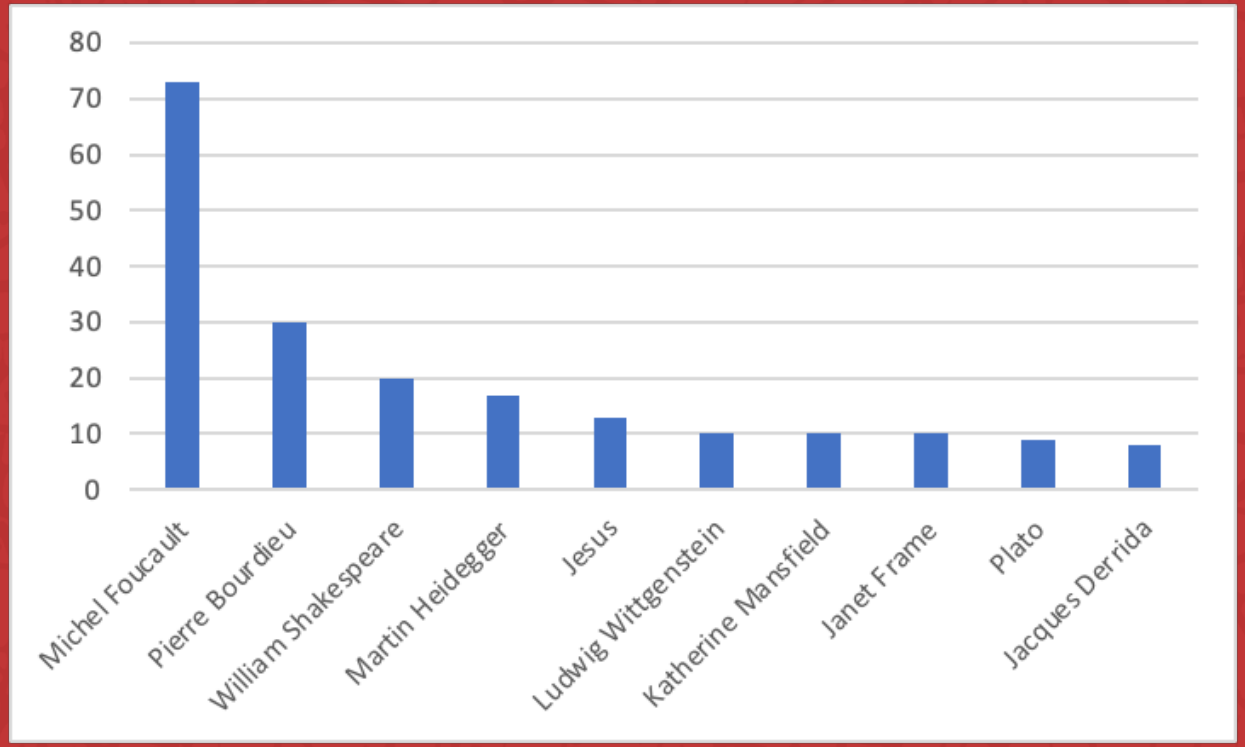


Taxa we study

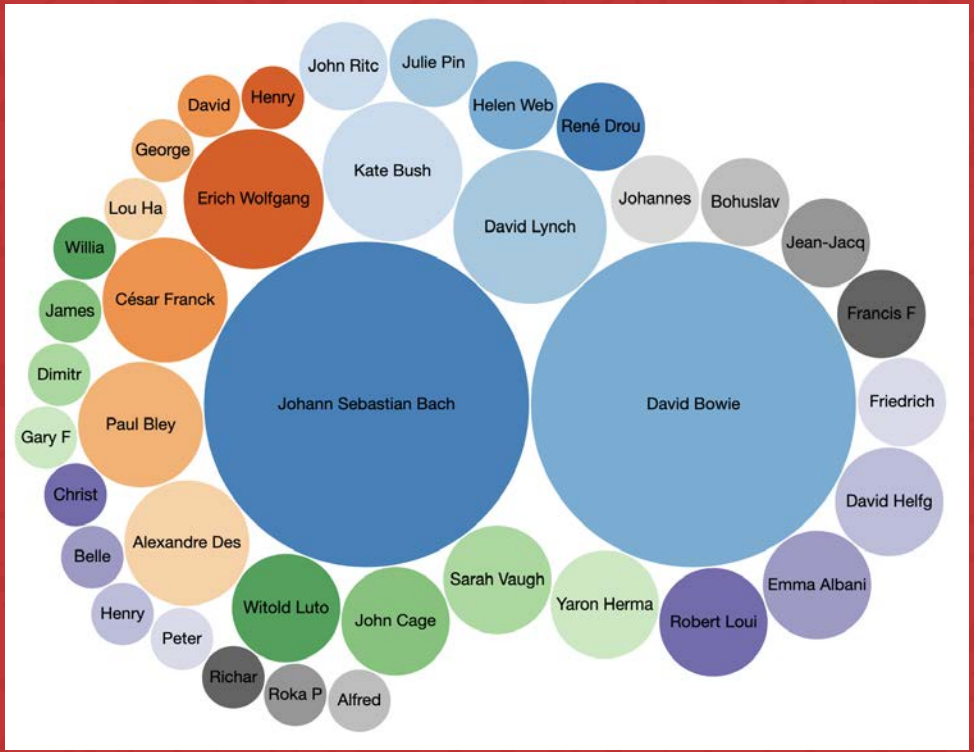


Places we study

Top ten people as main subjects



Bach or Bowie?



Future work

- Project & Wikidata community continue to match authors and add main subjects
- Machine learning for matching remaining keywords
- Process for libraries to keep adding theses themselves
- Measure impact in a year or two
- OCLC and ProQuest identifiers for theses
- Connect NZ authors to their theses overseas

Thanks

- All libraries who donated data
- Wikidata and Wikipedia community
- WANZ for $5k data cleaning grant and grant to attend LIANZA

Got a repository or other kind of collection?
Let’s talk!

Deborah Fitchett

deborah.fitchett@lincoln.ac.nz
ORCID: 0000-0002-7927-3321

Tamsin Braisher

DrThneed on Wiki
tamsin.braisher@otago.ac.nz
@thneed.bsky.social
@CabbageTree@mastodon.nz

Photo credits

Slide 2 Wikidata logo public domain

Slide 3 University of Auckland logo, Yvette, CC BY-SA 4.0. Massey University arms, Stanley Bannerman CC BY-SA 4.0. UC logo, University of Canterbury CC0. VUW logo, Victoria University of Wellington, CC0. University of Waikato image, New Zealand Tertiary Union, CC SA 2.0. University of Otago logo, Ulrich Lange, CC0. Otago Polytech logo, Otago Polytechnic, CC BY 3.0. AUT logo, Auckland University of Technology, CC0. Lincoln University logo, Lincoln University , CC BY 4.0

Slide 4 Excel logo public domain, OpenRefine logo public domain, Wikidata logo public domain

Slide 5 OpenRefine/Wikidata logo by Nikki, public domain

Slide 6,7,8,23 Tamsin Braisher CC BY 4.0

Slide 10 Mixnmatch screenshot CC0

Slide 11 Mixnmatch screenshot CC0. Cher Amy Stricula CC BY-SA

Slide 13 Alexander Gerst - Neutral Buoyancy Laboratory 1.jpg, by NASA, public domain, Beatrice Tinsley by Pelopanton, CC-BY-SA 4.0

Slide 14 Trove, Auckland Museum, Antilived CC BY-SA 4.0. Turnbull Library, Genet CC BY SA 4.0. British Library, public domain. NLNZ fair use. Papers Past unofficial mockup of logo. Natural History Museum logo, public domain. Smithsonian logo, public domain.

Slide 16 Image gallery from Wikimedia Commons, various licences. From Sparql query https://w.wiki/5WDK

Slide 17 Screenshot Sparql query, CC0 https://w.wiki/5Cfa

Slide 18 Entitree https://www.entitree.com/ CC BY-SA 4.0, Histropedia timeline CC BY-SA http://histropedia.com/faq.html,
screenshot advisor links Sparql query https://w.wiki/5Vye

Slide 19 Scholia CC BY-SA https://scholia.toolforge.org/

Slide 20/21/22/24 Sparql queries CC0 (https://w.wiki/5Vad, https://w.wiki/6rpa, https://w.wiki/6rpZ, https://w.wiki/6wbF)

Slide 27 Deborah's photo: Anon, used with permission
Tamsin's photo: Stephen A'Court, under contract by WANZ, CC-BY SA 4.0

Impact

- Does describing theses well on Wikidata and citing them on Wikipedia increase traffic to repositories?