How Australians are represented in Wikipedia

The Annual Report of the Wikihistories Project

Michael Falk Heather Ford Kelly Tall Tamson Pietsch

The Wikihistories project would like to acknowledge the Gadigal people of the Eora Nation, and the Wurundjeri Woi-wurrung and Bunurong peoples, upon whose ancestral lands our University campuses stand. We would also like to pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for this land.

2023-12-01

Table of contents

1 Introduction **3**

2 Analysis **6**

2.1 Australianness 6

2.1.1 In the category system 7

2.1.2 In Wikidata 9

2.2 Gender 11

2.3 Indigeneity 16

2.4 Historical period 20

3 Conclusion **24**

4 Appendix: Data **26**

4.1 Wikipedia's category system 26

4.2 Wikidata's query service 27

4.3 Metadata 28

References **29**

1 Introduction

The team at the wikihistories project is interested in how Wikipedia represents Australian people, places and events over time. National studies of Wikipedia are rare. The encyclopedia is usually treated as a representation of global trends. We now have excellent studies that have demonstrated how Wikipedia represents men at the expense of women (e.g. see Konieczny and Klein 2018), and people and places in North America and Western Europe rather than the Majority World (e.g. see Beytía 2020). But we have very few quantitative national studies of Wikipedia.

National studies of Wikipedia are critical to our understanding of representation. Wikipedia is a powerful producer of notability in Australia as it is in many nations, as we recently argued (Ford, Pietsch, and Tall 2023). Who does and does not have a Wikipedia biography matters. *How* Wikipedia represents people also matters. Being well represented on Wikipedia makes a difference for one's reputation and for the kinds of opportunities available to people in work and life.

In this, our first year of the wikihistories project, our focus is on how Wikipedia represents *people*. We ask how English Wikipedia represents Australians. Do we see the same trends in terms of gender and ethnicity so that globally studies of Wikipedia have identified? How well has English Wikipedia done over time to represent people in Australia? These are important questions but they are by no means easy to answer. In order to map Wikipedia's representation of Australians over time, we first need to understand which biographies on Wikipedia refer to Australians. The question of who is recognized as an Australian is a question beset with complexities and politics before we even get to Wikipedia. On Wikipedia, further complexities arise.

Who counts as Australian in Wikipedia and its sister-projects? How are Wikipedia articles or Wikidata items marked as 'Australian' by the system? How does Wikipedia define or represent 'Australianness'? In this report, we try to address these questions using a dataset of biographical articles culled from English Wikipedia. We have two primary aims: (1) to reveal the definition of 'Australianess' implicit in Wikipedia's systems; (2) to assess how well Wikipedia represents the diversity of Australians. As we will see, Wikipedia works implicitly with numerous inconsistent definitions of 'Australianness', but it is nonetheless possible to make some reasonable observations about how well it represents the diversity of people who make up 'Australia'.

People can be represented in Wikipedia in many ways. They may have an empty article that redirects to another article that concerns them (e.g. Azaria Chamberlain). They may be mentioned in the text of an article (e.g. Sasha Soldatow or the 'manual workers'

in Sydney Push). Or they might have a full-blown article devoted to their biography (e.g. Ada Cambridge). For the sake of our analysis, we focus on the third kind of representation: the biographical article. This allows us to quantify our analysis, because each article in the dataset equals one person.

We explain how we construct the dataset in Appendix: Data. Broadly, we use two methods to identify 'Australian' biographies in Wikipedia. First, we use Wikipedia's category system to find all biographies contained in Category:Australian people. Second, we use Wikidata's query service to find all biographies whose corresponding Wikidata item has a property suggestive of 'Australianness' (e.g. a People Australia ID). Using these two methods, we identify 83,013 biographies on English Wikipedia that are marked as 'Australian' in some way. We enrich these biographies with additional Metadata for further analysis. As Figure 1.1 shows, the number of Australian biographies has consistently grown since the early days of Wikipedia.

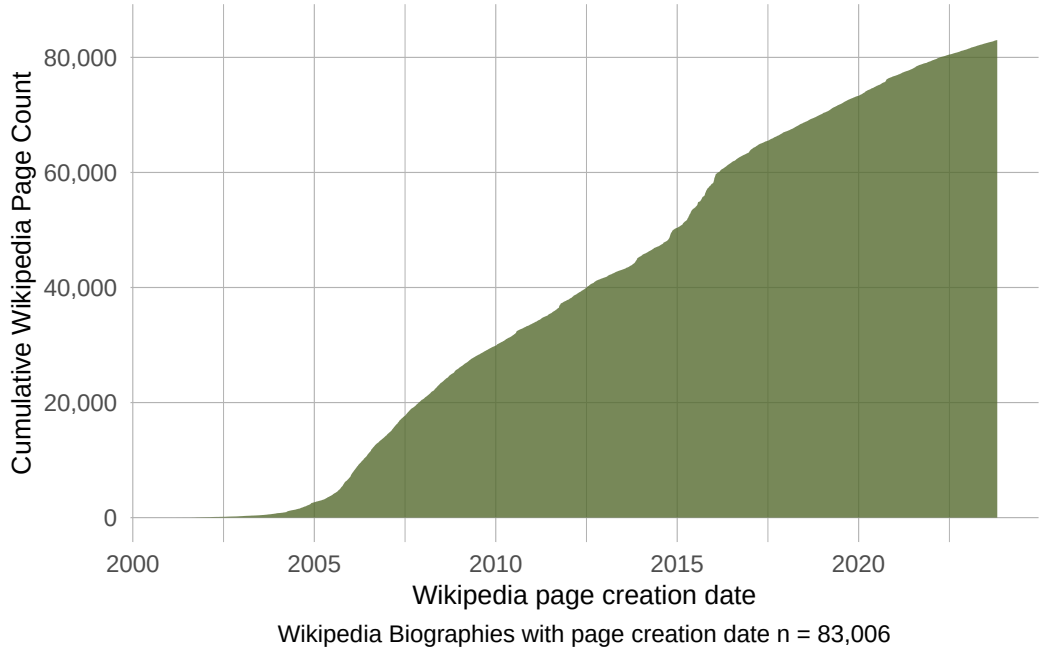

Figure 1.1: Number of articles in the Australian biographies dataset over time

Our Analysis of the data unfolds in four main sections. In Australianness, we consider what it means for a person to be 'Australian' in Wikipedia. Where does the system draw the line(s) between Australians and non-Australians? We show that the system draws many different lines, which do not always conform with one another. In the subsequent sections of analysis, we consider how Wikipedia captures the diversity of Australia's people, along the axes of Gender, Indigeneity and Historical period. We demonstrate that it is possible to make some generalisations about how Australians are represented in Wikipedia. In particular, we observe that the encyclopaedia is highly

conservative: generally, Wikipedia assumes that an 'Australian' is a white cisgender male; only if a person departs from this norm are their racial or sexual characteristics explicitly described. While representation of some minority groups may be improving (e.g. the gap between Indigenous and settler Australians), other representation gaps remain tenacious (e.g. the gender gap).

In Conclusion, we propose some guidelines for future research into Wikipedia's systemic biases. We show that it is possible to judge how well Wikipedia represents a certain group by counting pages, but that this method needs to be carefully planned and contextualised to account for local factors.

Figure 1.2: How do Wikipedians construct Australia online?

2 Analysis

2.1 Australianness

In this report, we do not seek to define what it means to be 'Australian'. Instead, we aim to identify the 'markers of Australianness' that define what it means to be 'Australian' in Wikipedia. In Wikipedia, the primary marker of Australianness is membership of `Category:Australian people`. In Wikidata, no marker can be said to be primary, but the most frequent in our dataset is P27(Q408), a property meaning that a person is a 'citizen' of Australia. In both cases, we will see that there is a sharp divide between the way Wikipedians have defined the marker, and they way they have used it in practice.

There are two reasons we refrain from defining 'Australian' for ourselves. In the first case, our opinion is not at all important—what matters is Wikipedia's opinion. In the second case, it is exceptionally difficult to define 'Australian' in a non-arbitrary way. In two dimensions, **space** and **time**, the concept of 'Australian' poses startling difficulties.

In **space**, it is difficult to judge whether a person is close enough to Australia to count as Australian. At one end of the specturm is someone who is born, lives and dies in Australia, such as Oodgeroo Noonuccal. There can be no dispute that such a person is Australian. At the other end of the spectrum is someone who has only a passing cultural or economic connection to Australia, such as Steve Wozniak, one of the more surprising members of `Category:Australian people`. Then there is the problem of the diaspora. Are the children of Australians Australian, even if born overseas? If, like Rupert Murdoch, you leave Australia and renounce your citizenship, do you remain Australian? There may be easy answers to these questions in specific contexts, for example, when calculating a person's tax liability or entitlement to vote, but in the context of an encyclopaedic project like Wikipedia, they cannot be easily resolved.

In **time**, it is difficult to judge when in history it becomes reasonable to describe someone as 'Australian'. Humans first arrived in Australia some 60,000 years ago, but they did not call the continent 'Australia'. The term 'Terra Australis' ('Southern Land') was coined by 16th-century Europeans. At this time, 'Terra Australis' was merely a proposal, as no European had visited either Australia or Antarctica. In the end, neither Australia nor Antarctica matched European expectations about the 'Southern Land', though Australia, being the first that Europeans visited, acquired the name. New South Wales was invaded by the British in 1788. The term 'Australia' was substituted for 'Terra Australis' in 1813. The colonies of Australia federated into a single Commonwealth in 1901. The resulting Commonwealth of Australia only achieved full independence from Britain in

1988. The problem of time is compounded by geology. In the course of Australia's human occupation, the very landmass has altered considerably. Only 10,000 years ago, the Australian mainland, Papua New Guinea and Tasmania formed a single contiguous landmass that geologers call Sahul. Are all the people who occupied Sahul at any point in time 'Australians'? Or can only residents or citizens of the country 'Australia' be so named? Like space, time wreaks havoc on the category 'Australian'.

The editors of Wikipedia and Wikidata grapple with both these problems. The result is a kind of categorical anarchy, where different parts of the system work with different concepts of 'Australianess' and apply different criteria of connection. It is for this reason that we talk of 'markers of Australianness' in Wikipedia. How do these markers work?

2.1.1 In the category system

The primary marker of Australianness in Wikipedia, as we have seen, is membership of `Category:Australian people`. The category offers simple criteria for membership:

> This category lists notable Australian-born people, or people who identify themselves as Australian.

A user who clicks on the hyperlink 'Australian' will be taken to the Wikipedia page for Australia, which defines Australia as follows:

> **Australia**, officially the **Commonwealth of Australia**, is a sovereign country comprising the mainland of the Australian continent, the island of Tasmania, and numerous smaller islands.

These definitions seem to offer a strong answer to the problems of space and time. 'Australia' is the 'Commonwealth of Australia', the sovereign state that has existed since 1901. In space, a person must be 'born' in Australia, or 'identify themselves' as a member of the country. In time, Australia is the Commonwealth of Australia, and so has only existed since 1901.

The problem is, this explicit definition does not at all describe the contents of `Category:Australian people`. In the spatial dimension, the category includes living people such as Steve Wozniak and Darrell Duffie who were neither born in Australia nor identify themselves with it. In the time dimension, the category includes many historical figures who long preceded the Commonwealth of Australia, such as Bennelong, Mary Reiby, Robert Lowe and Louisa Atkinson. Finally, it lists thousands of entities that aren't people, such as First Ever, the sportswear manufacturer, and 'Hand on Your Heart', the 1989 single by Kylie Minogue. How has `Category:Australian people` expanded to include so many pages that contradict its own explicit definition?

The answer lies in the category's size. When editors include articles in `Category:Australian people`, they do not necessarily do so intentionally. Category:Australian people is the root node of a vast category graph. The editor who categorised Steve Wozniak as a

staff member of the University of Technology Sydney was probably unaware that they were thereby categorising Wozniak as an 'Australian person'. The category for academic staff at UTS lies several layers deep in the category graph, 4 steps away from the root (Figure 2.1). Perhaps an editor may quibble that a person who works at UTS is not necessarily 'from Sydney', but in the present state of the category graph, there is no other way to mark a staff member of UTS. The category graph is strictly hierarchical. A subcategory cannot be removed from its parent category for the sake of a single page.

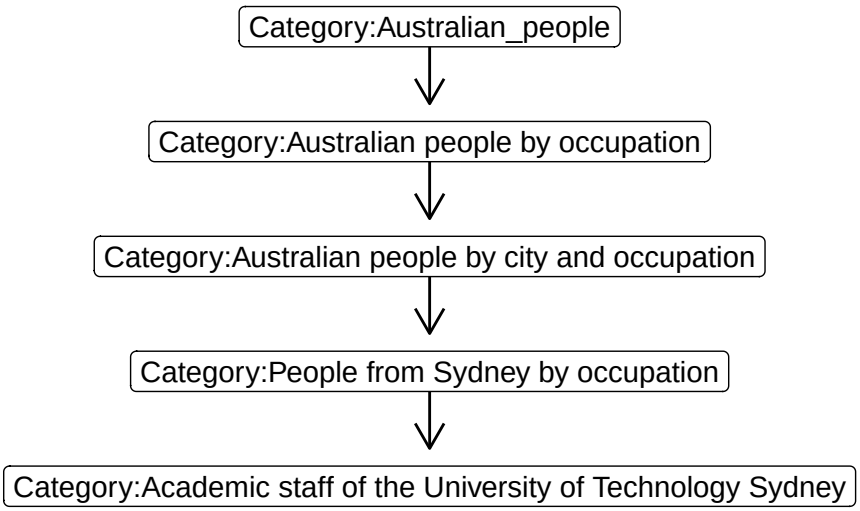

Figure 2.1: The category for UTS staff lies four steps away from Category:Australian people

Category:Australian people has 9,492 subcategories, many of which are even farther from the root than the staff of UTS. We can get a sense of how vast this structure is by taking diameter of the category graph. In this case, the diameter is 11, which means that the most distant subcategory of `Category:Australian people` is 11 steps away. Such distant nodes are called 'eccentric' in graph theory. In this case, the most distant subcategory truly is eccentric: Category:Custard (band) compilation albums does not include articles about 'Australian people' at all, but about indie rock records. This case is extreme. The case of UTS staff members is more typical. If we consider all the articles that have been placed in `Category:Australian people` or its subcategories, we find that the subcategories used are on average 3.17 steps away from the root, with a standard deviation of 0.95. In the typical case, when an editor adds an article to an 'Australian people' subcategory, the subcategory is 2-4 steps away from the root. It seems likely that editors are mostly unaware of the category graph as a whole. Editors work locally,

trying to find the best categorisations for their articles, while Wikipedia's systems work globally, enforcing the hierarchical logic of the enormous and usually invisible category graph.

There is thus a contradiction between the global and local logics of the category graph. Frequently, the category graph also contradicts the actual content of the encyclopaedia. First Ever is categorised as an 'Australian person' even though the text of the article identifies it as 'an Australian manufacturing company'.[1] These contradictions breed a pervasive *categorical anarchy*. Seen from afar, Wikipedia is extremely lax about who or what counts as 'Australian'. A person—or even *not* a person—can be categorised as 'Australian' for whatever reason seems reasonable in the local context. Such laxity is probably necessary to Wikipedia's encyclopaedic ambitions—how else could the vague concept of 'Australian person' ever find a place in Wikipedia? Wikidata is a different beast. Rather than a collection of textual articles that can speak for themselves, it is a database of structured facts (called 'claims') which are supposed to provide a reliable machine-readable knowledge graph. Can categorical anarchy prevail in such a system?

2.1.2 In Wikidata

Wikidata's contributors are well aware of the difficulties of space and time. The simplest way to mark someone as 'Australian' in Wikidata is to give 'Australia' as the person's 'country of citizenship' (P27). Our dataset contains 47,445 'Australian citizens' according to P27. A glance at the Talk page for P27 reveals that Wikidata's editors are well aware of the problems with trying to categorised people as 'Australian' in any stable way.

The user Ghouston clearly articulates the problem of time in a post from 2018:

> It's not easy. Australian citizenship didn't exist before 1949, yet Australia as a federation has existed since 1901 and the term "Australian" was in use in the 19th century. To say that the United Kingdom was created in 1927 or the Netherlands perhaps in 1815 or 1945 or China in 1949 is to ignore much of the history of these countries as states.

Two years later, Ghouston provides additional sources and mentions a possible solution:

> Separate items for citizenships and nationalities have been proposed previously. Perhaps it would help. The term "Australian" for people living in Australia was already used in the 19th century, before Australia had even been federated as a state, e.g., [2]. According to [3], Australia started issuing passports after federation, and didn't even restrict it to British subjects until

[1] In this case, the article falls under Category:Clothing brands of Australia, which is a distant subcategory of Category:Australian designers, which is then linked through a number of other categories to Category:Australian people.

1938, but it seems that passports were not used much prior to WW1, and there was a more flexible attitude to nationality.

At the time of writing, there is no Wikidata property for 'nationality'. Instead, the property of citizenship is limited according to the 'inception' (P571) of the given country and the death date of the given person. The start date of Australia is given as 1 January 1901 in Wikidata. If a person is assigned Australian citizenship in Wikidata, and their death date occurs before 1 January 1901, then a warning appears in their Wikidata page. At the time of writing, the Wikidata item for colonial poet Charles Harpur contains such a warning (Figure 2.2). Harpur, who proudly described himself as as 'An Australian' in newspaper publications from the 1830s, died in 1868.

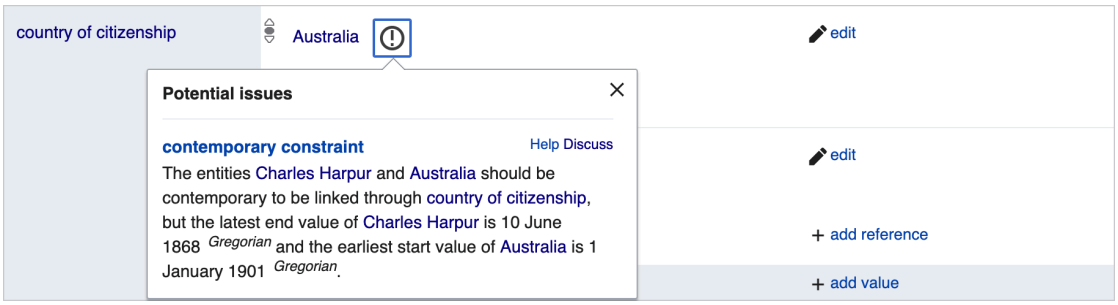

Figure 2.2: Citzenship warning banner for Charles Harpur

This solution combines flexibility with clarity.

- Flexibility: It is possible to mark the 'Australianness' of Charles Harpur and others who have been identified as Australian throughout history.
- Clarity: Wikidata provides clear guidance to its contributors through the warning banner. As far as Wikidata is concerned, Australian citizenship was available to people from 1901.

The price of this clarity and flexibility is inconsistency and quixotism.

- Inconsistency: Wikidata flatly contradicts itself as far as historical figures like Charles Harpur are concerned. Was Harpur a citizen of Australia or not? Wikidata also contradicts, or at least brings into question, its own definition of 'citizenship'. Wikidata defines citizenship (P27) thus: 'the object is a country that recognizes the subject as its citizen' This definition is not reflected in P27's rules: all that is required is that the country exists prior to the supposed citizen's birth date. If Harpur were posthumously recognised as a citizen by the Commonwealth of Australia, this could not currently be reflected in the database. The warning banner would remain.
- Quixotism: As Ghouston himself observes, Australia did not have an official system for conferring citizenship until 1949. By seeming to choose 1 January 1901 as earliest permitted date, Wikidata has invented its own concept of Australian

> citizenship. Is this quixotism consistent with Wikidata's self-description as a 'secondary database, collecting structured data'? Does Wikidata collect structured data, or does it structure collected data?

Wikimedians are aware of the difficulty of time. The state of P27 suggests that they are content for the minute to leave the difficulty unresolved.

The difficulty of space is less clearly articulated on the talk page for P27 ('country of citizenship'). Perhaps this is because the concept 'citizenship' presupposes a deep and publicly sanctioned connection between the person and the country. In a sense, Wikidata delegates the question of a person's spatial connection to Australia, by recording when a person appears in a known database of Australians. These databases make their own decisions about the 'Australianness' of their subjects. The Dictionary of Australian Biography, for example, does not required its subjects to be Australian at all, as explained in their FAQs:

> **Do you have to be an Australian to be in the ADB?**
> The ADB includes anyone who has made a significant contribution to the Australian nation.

The AusStage database, meanwhile, records any person who 'contributed' in some way to a theatrical performance in Australia, or to an overseas performance of an Australian theatrical work. AusStage and the Australian Dictionary of Biography provide very broad markers of Australianness. By contrast, the Re-Member database of the Parliament of Victoria provides a stricter marker of a person's Australianness. To appear in Re-Member, a person must be a former member of the Parliament of Victoria. We identified dozens of other databases that indicated 'Australianess' in Wikidata. When a Wikimedian links a person to the Australian Dictionary of Biography (P1907), AusStage (P8292), Re-Member (P8633), or some other database of Australians, the linked database will encode a different answer to the question 'Who Counts as Australian?'

2.2 Gender

Although it is impossible to provide a unitary definition for 'Australian' in Wikidata and Wikipedia, we have shown that it is possible to derive a large list of people who are marked as 'Australian' in some way across the encyclopaedia and database. What is the composition of this list? We begin with gender.

To determine the gender of a person in our dataset, we rely on the Wikidata property P21:Sex or gender. In our dataset, nearly all records have a sex or gender recorded (Figure 2.3). Only 6 records lack a `P21` statement about their sex or gender.

Weathington and Brubaker (2023) have strongly critiqued Wikidata's representation of sex and gender, observing that "Queer identities are recorded in a database in clunky,

Sex or Gender Label	Total
fa'afafine	1
female	16560
genderfluid	6
genderqueer	1
intersex	3
intersex woman	1
male	66375
non-binary	28
trans man	3
trans woman	27
transgender	2
NA	6

Figure 2.3: Gender breakdown of the Australian biographies dataset

inaccurate, or self-contradictory ways" (p. 2) and that the database ignores the individual's right to determine their own sexual presentation. Nonetheless, we are compelled to rely on Wikidata's representation of individuals' sex and/or gender if we wish to assess its representation of human diversity in Australia. To this end, we categorise our dataset into three broad groups: `male`, `female` and `trans, non-binary or intersex`. While we may assume that most of those marked as 'male' or 'female' are *cis*-male or -female, we cannot determine this from the available data, and so opt for the less specific 'male' and 'female'.

`Female` biographies are clearly under-represented among Wikipedia's Australian articles. (Figure 2.3) shows all the articles in the dataset, categorised by their gender on Wikidata. The gap between `male` and `female` biographies appears to have remained constant over time (Figure 2.4). Viewing the data on a logarithmic scale emphasises the parallel track of `male` and `female` biographies in Australian Wikipedia, while also allowing us to visualise the similar growth of `trans, non-binary or intersex` biographies (Figure 2.5).

Not only are male, female and LGBTQI+ Australians represented in different numbers in the encyclopaedia, they are represented by different words. To show this, we downloaded the introductions of each article in the encyclopaedia, and used the tf-idf statistic to determine which words were most distinctive for each group of biographies. Which words tend to appear more often in the introductions of articles about `male`, `female` and `trans, non-binary or intersex` Australians?

Figure 2.6 discloses two significant patterns:

- *Explicit marking of non-Male genders*: First, Wikipedia articles are more likely to explicitly mention the sex or gender of a `female` or `trans, non-binary or`

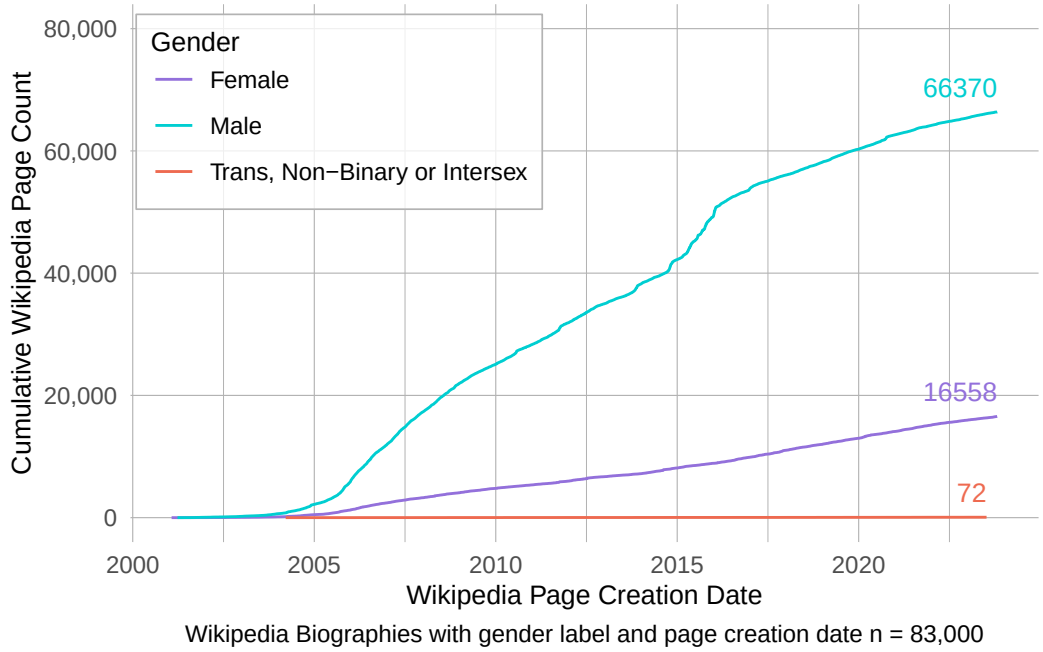

Figure 2.4: Growth of the Australian biographies dataset over time, broken down by gender

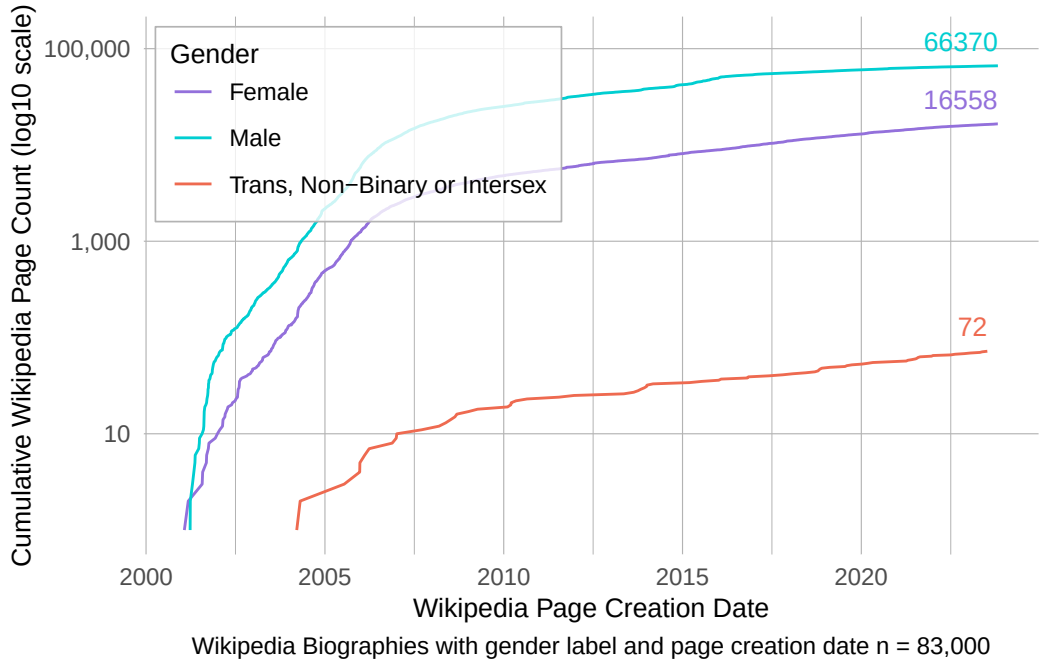

Figure 2.5: Growth of the Australian biographies dataset over time, broken down by gender, on a log scale to compare growth rates for each gender.

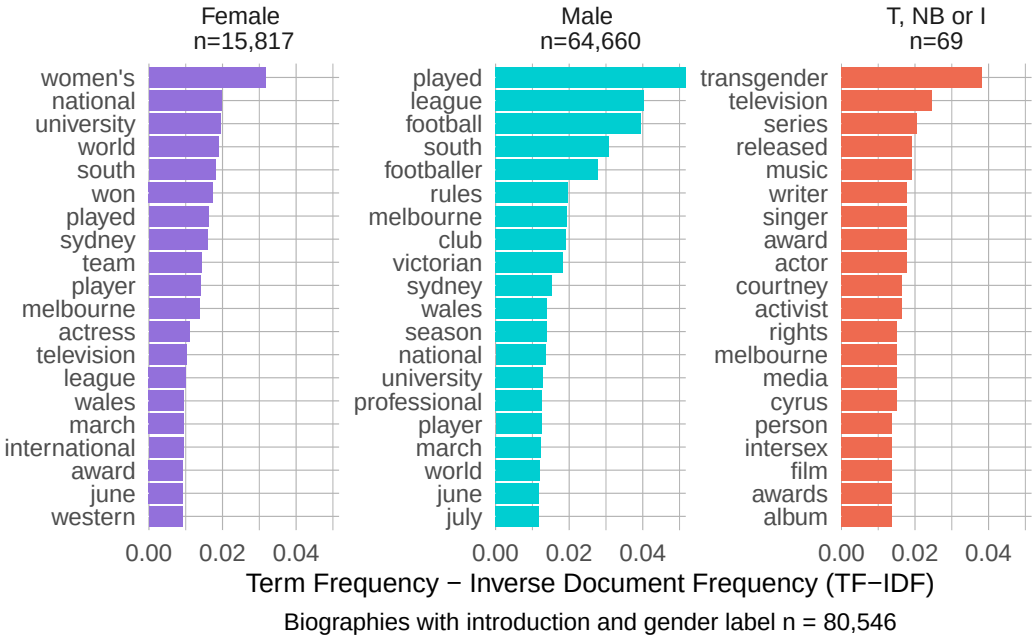

Figure 2.6: Comparing the introductions of each article in the Australian biographies dataset, using the tf-idf metric to reveal the most distinctive words in the introductions for each gender grouping.

`intersex` Australian. The high tf-idf score for "women's" in biographies of Females probably reflects the fact that female sporting teams often have "women's" in their names. Thus Sam Kerr is said to represent Australian in the "national women's team", while Harry Kewell was simply a member of "the Australian national team". The situation for `trans, non-binary or intersex` Australians is even more marked. 'Transgender', 'sex' and 'woman' all appear as distinctive terms in the introductions of Wikipedia biographies in this group.

- *Dominance of sport and entertainment*: Across all three gender categories, sport and entertainment are dominant themes. The `male` biographies seem less likely to describe entertainers than those in the other categories. Unsuprisingly, the `trans, non-binary or intersex` biographies seem more likely to discuss social affairs, 'activist' and 'rights' being prominent terms in the introductions of these biographies.

The data support the views of Weathington and Brubaker (2023). In Australian Wikipedia, maleness is the unmarked default. `Female` and especially `trans, non-binary or intersex` Australians are described using terms such as 'women' or 'transgender'; there is no need to explicitly mention that a person is male or cisgender. In the case of `trans, non-binary or intersex` people, this explicit marking of gender goes a step further. Australians in this group are likely to be described as 'activists', presumbly on gender issues. Biographies of `trans, non-binary or intersex` Australians problematise gender in a manner that biographies of `male` Australians do not.

Wikipedia's defaulting to maleness probably reflects the linguistic habits of broader Australian society, but it could also be exacerbated by the encylopaedia's obsession with sport and entertainment. Both sport and entertainment are highly gendered fields. Men and women compete in gendered competitions such as the soccer or cricket World Cups and receive gendered commendations such as Best Actress at the Academy Awards. While `male` Australians may dominate non-`male` Australians in fields such as politics, business and the arts, these fields are not so explicitly gendered. Sam Kerr may indeed be captain of the "women's national team," but it makes no sense to describe Vanessa Hudson as the CEO of a "women's corporation." Sport and entertainment are fields that reflect society's deepest-held and most immoveable convictions. Wikipedia's powerful focus on these fields may place the encyclopaedia behind other parts of Australia in the evolution of non-gendered language.

Figure 2.7 suggests a silver lining to this story of the gender gap. Although `male` Australians continue to greatly outnumber `female` and `trans, non-binary or intersex` on Wikipedia, many of the `male` biographies are of low quality. In fact, if we consider only C-Class articles and above, the proportion of `female` biographies rises from 19.9% to 33.0%. It appears that editors have devoted considerably more effort to creating pages about men than about women or people of other genders, but that editors devote more effort on average to maintaining pages about non-`male` Australians.

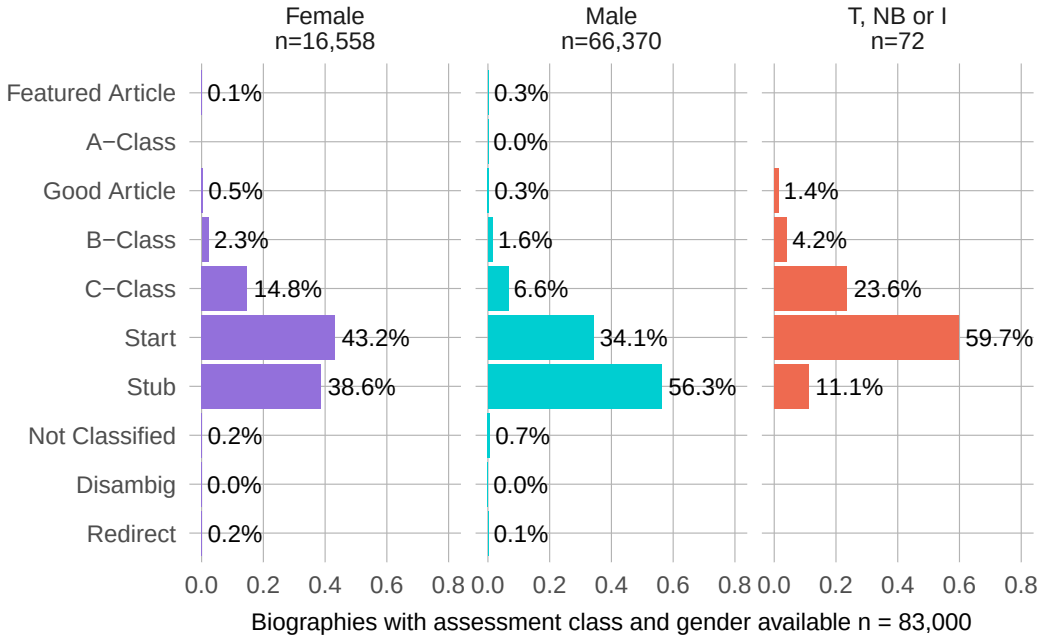

Figure 2.7: Page assessment class by gender in the Australian biographies dataset

2.3 Indigeneity

The distinction between 'settler' and 'indigenous' Australian pervades Wikipedia in a cryptic way. Race and ethncity are portrayed in a lopsided manner in Wikipedia, much like gender. In English Wikipedia, Mandiberg (2023) observes, people are not officially classified by 'race', but whiteness is assumed as a default, and a person is far more likely to be categorised by their ethnic or national origin if they are non-white.

There is no explicit category for 'settler' or 'white' Australian. There are categories for Category:Immigrants to Australia and Category:Australian people of European Descent, but these are applied only haphazardly to indicate the 'settler' status of a minority of Australians. A good example is Tony Abbott. According to his Wikipedia biography, he was born in London to English parents, and 'migrated to Australia at the age of two'. His page is nearly 20 years old, and has been edited more than 4000 times.. Nonetheless, no-one has thought to categorise him either as an 'immigrant' or as an Australian 'of European descent'. Extending Mandiberg's arguments, we can hypothesise that 'settler' status is assumed in Australian Wikipedia, and a person's whiteness or migrant status is considered unremarkable if they are British. We substantiate this hypothesis below.

There may be no single category for 'settler' or 'white' Australian, but there is certainly a single overarching category for Indigenous Australian people. Within this category, editors are instructed to sub-categorise Aboriginal and Torres Strait Islander Australians

as precisely as possible (Figure 2.8). These subcategories fall into three main kinds:

1. Sub-categories for mob or language group, e.g. Category:Pitjantjatjara people.
2. Sub-categories for occupation, e.g. Category:Indigenous Australian linguists (Curiously, neither guide nor tracker is considered an occupation in this scheme.)
3. Miscellaneous subcategories, including Category:Members of the Stolen Generations, Category:Last known speakers of an Aboriginal language, Category:Indigenous Australian feminists and Category:Australian Aboriginal elders.

Figure 2.8: `Template:Category diffuse` instructs editors to categorise Indigenous Australian People more precisely

`Category:Indigenous Australian people` is actually nested quite deeply within the Australian category graph (Figure 2.9). Figure 2.9 may at first be confusing to Australian readers, because it is common in Australia to refer to Aboriginal and Torres Strait Islander people as "First Nations Australians." In English Wikipedia's category graph, however, the term "First Nations" is used the Canadian sense.[2] In this instance, Wikipedia's category graph is potentially incendiary, because it appears to imply there is no special connection between Indigenous Australians and the continent to which they are indigenous. The overarching `Category:Australian people by ethnic or national origin` is organised predominantly by continent or region. There are sub-categories for Australians of African, Asian, Carribean, Latin American, Middle Eastern, North American and Oceanian descent. There is no category for "Australians of Australian descent", nor is `Category:Australian people of Indigenous Australian descent` included as a subcategory of `Category:Australian people of Oceanian descent`.[3] By contrast, `Category:Australians of First Nations descent` is included as a subcategory of `Category:Australian people of Canadian descent`, which is itself a subcategory of `Category:Australians of North American`

[2]This has the notable corollary that there are no categories in Wikipedia for Métis or Inuit Australians.
[3]In other parts of English Wikipedia, Australia is included as part of Oceania. For instance, `Category:British people of Australian descent` is a subcategory of `Category:British people of Oceanian descent`. The same holds true for American, French, and Indian people of Australian descent.

descent. In this local portion of the graph, the link between First Nations Canadians and Canada is formally encoded. The link between Indigenous Australian people and Australia is not. A naïve reader of the graph (such as a computer program) might erroneously conclude that `First Nations` Australians have the same relationship to Australia as `Indigenous` Australians.

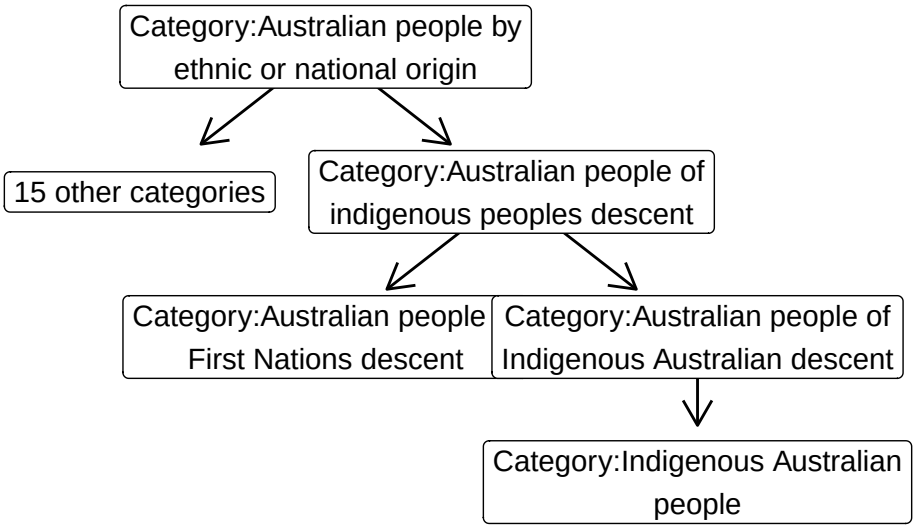

Figure 2.9: Local portion of the category graph for Category:Australian people, showing how Category:Indigenous Australian people fits into the system

In Wikipedia's defence, the indigeneity of Indigenous Australians is encoded in other ways. In this local portion of the graph, a human can see from the *labels* of `Category:Australian people of Indigenous Australian descent` and `Category:Indigenous Australian people` that people in these categories are indigenous *to Australia*. Looking beyond this local portion of the graph, we can also see that `Category:Indigenous peoples of Australia` is a subcategory of `Category:Indigenous peoples of Oceania`. It is nonetheless curious that the local portion of the graph that classifies Australians by 'ethnic or national origin' should de-emphasise `Category:Indigenous Australian people`.

The situation is different for other countries in English Wikipedia. Other settler-colonial countries include categories for their own indigenous peoples directly beneath the super-category for `people by ethnic or national origin`. For example, Category:Indigenous Mexicans is a direct subcategory of Category:Mexican people by ethnic or national origin, while the analogous Australian category, `Category:Indigenous Australian people`, is two steps deeper in the graph. A similar pattern holds for other

settler-colonial states. New Zealand, New Caledonia, South Africa and the United States all include top-level categories for their Indigenous peoples under the `people by ethnic or national origin` super-category.

There is a second notable difference between the category graphs for Australia and Mexico. We have seen that Australia's `Category:Australian people of indigenous peoples descent` is used to classify people who are indigenous anywhere. The analogous Mexican category, `Category:Mexican people of indigenous peoples descent`, is used exclusively to classify indigenous peoples of Mexico itself. Brazil and Argentina similarly reserve their 'indigenous peoples' category for their own indigenous peoples.

Australia is not entirely alone. The subgraph for `Category:Japanese people by ethnic or national origin` also obfuscates the indigeneity of Japan's indigenous peoples. There is no explicit category for Japanese people 'of indigenous descent'. Instead, `Category:Japanese Ainu people` is included as a subcategory of `Category:Japanese people of East Asian descent`. In this way, the link between Ainu people and the region is locally marked in the category graph, but one must look elsewhere in the graph to discover that Ainu people are *indigenous* to Japan.

It is clear that quantifying indigenous representation in Wikipedia is a fraught exercise. Scholars are well advised to carefully consider how indigeneity is understood in different parts of the world, and then to pay close attention to the particular structure of ethnic and national classification in any particular part of Wikipedia.

Nonetheless, we can use the category graph to estimate the number of biographies of Indigenous Australians in Wikipedia. Figure 2.10 suggests that in recent years, Wikipedians may have begun to address the representation gap between Indigenous and settler Australians. Figure 2.10 shows the proportion of biographies in our dataset that fall within `Category:Indigenous Australian people`. The representation gap for Indigenous Australians is not so stark as the representation gap for non-male Australians. According to the Australian Census, the number of Indigenous Australians grew from 2.5% of Australia's population to 3.2% between 2011 and 2021. In our estimate, the proportion of Indigenous biographies in `Category:Australian people`, went from 2.29% on January 1 2011 to 2.05% on January 1 2021. At the time of our data collection, the proportion had risen again to 2.16%. Although comparison with the census would suggest that Wikipedia under-represents Indigenous Australians, the representation gap is currently narrowing.

As for gender see 2.7, we can consider how indicators of page quality alter the picture. Figure 2.11 does indeed suggest that the Aboriginal and Torres Strait Islander biographies repeat the same pattern as for gender. Although biographies of Aboriginal and Torres Strait Islander Australians are less numerous than biographies of other Australians, they are generally of higher quality according to Wikipedia's own quality

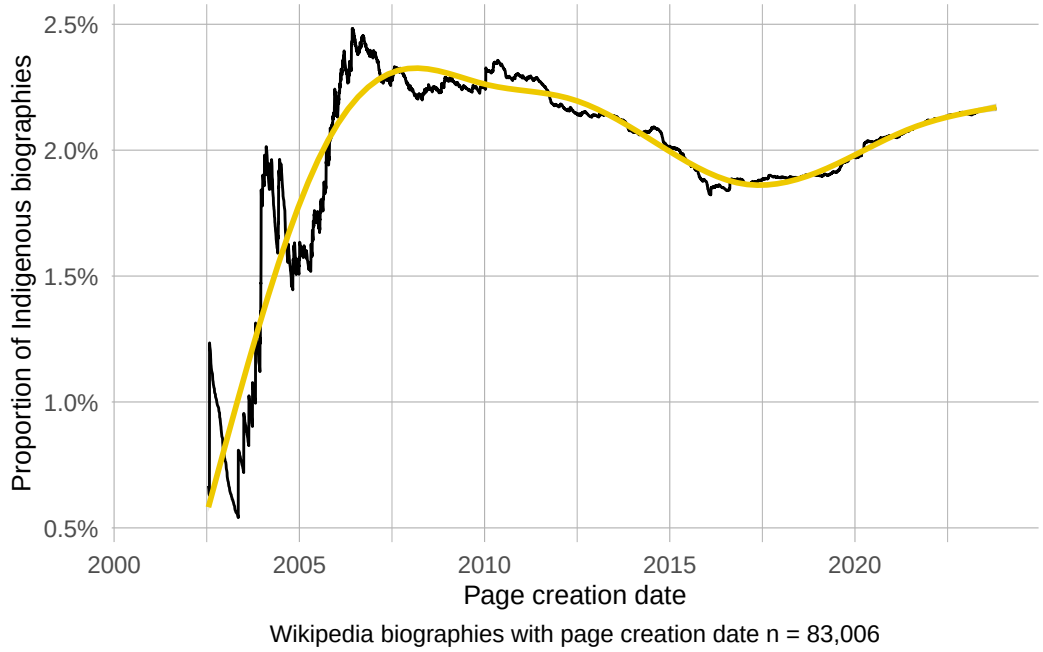

Figure 2.10: Proportion of Indigenous biographies in the Australian biographies dataset over time.

assessment standards. If we consider only articles of C-Class and above, then the proportion of Aboriginal and Torres Strait Islander biographies in the dataset climbs from 2.2% to 3.7%.

These results do need to be taken with a grain of salt. As Thorpe, Sentance, and Booker (2023) show, Wikipedia's articles on Aboriginal and Torres Strait Islander people are dominated by 'settler narratives', and typically rely on public-domain sources that reflect the values of Nineteenth-Century British colonists. Our analysis of Aboriginal and Torres Strait Islander representation in Wikipedia must therefore be supplemented with more detailed analysis of the sources and rhetoric of the articles themselves.

2.4 Historical period

The final aspect of human diversity we consider is historical period. How are people from different periods in Australia's history represented in the encyclopaedia? To estimate this, we rely on date of birth data from Wikidata. Wikidata records a date of birth for 70,741 of the articles in our dataset. This data is displayed in Figure 2.12, which shows the number of biographies in the dataset according to the decade in which the person was born.

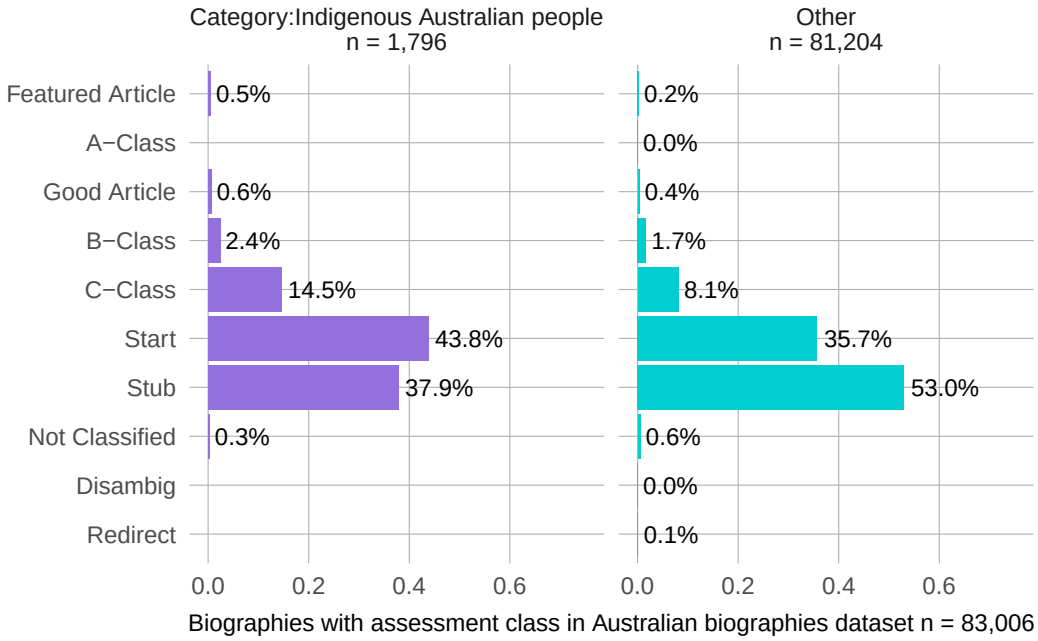

Figure 2.11: Page assessment class broken down by membership of Category:Indigenous Australian people

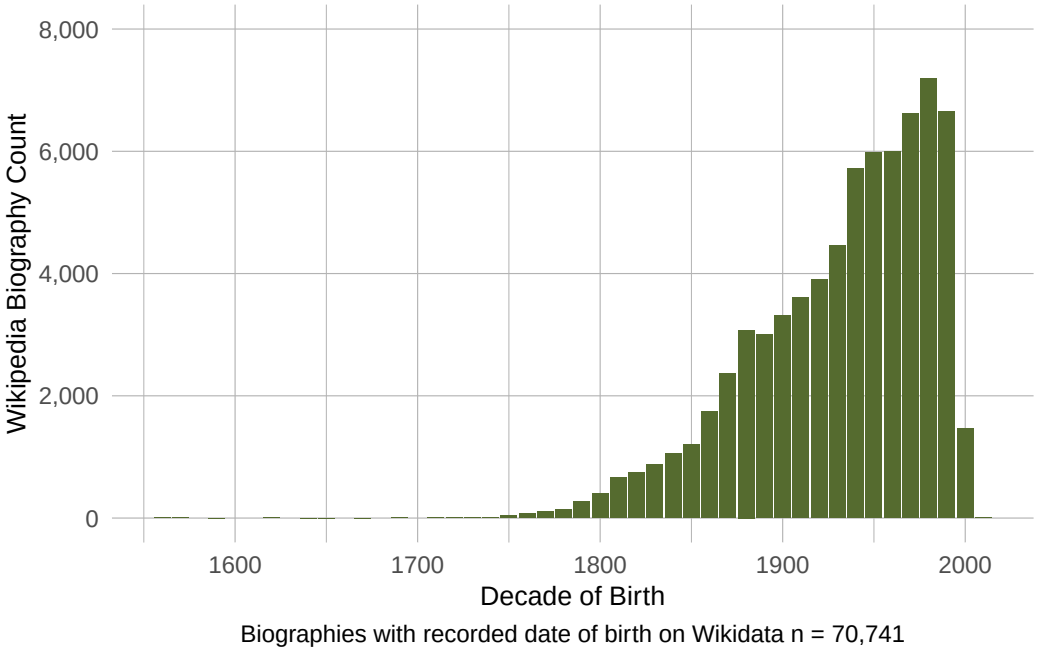

Figure 2.12: Articles in the Australian biographies dataset broken down by the decade in which the subject of the biography was born.

Figure 2.12 displays the obvious colonial bias in the dataset. There are very few biographies of any persons prior to British colonisation of Australia. This is partly a product of our focus on biography. There is very little biographical data available about individual persons living in Australia prior to 1788. But this colonial bias also represents a darker side of Australian history. In the first half of the Nineteenth Century, the majority of Australia's population was probably indigenous. Butlin (1994) estimates that that colonial population only overtook the indigenous population in the 1840s. According to his estimates, the population of Australia hovered around 600,000 from around 1800 until the 1840s, because the number of new migrants into the colony was roughly equal to the decline of the indigenous population due to British depradations (1994, 212). In Figure 2.12, this history is obscured. A naive interpreter of the figure might assume that Australia's population grew continuously from the British invasion of 1788 until the present. In fact, there was a sharp decline in Australia's population in the first two decades of settlement, followed by a period of relative stagnation.

We can conclude that Wikipedia's Australian biographies broadly reflect the growth of colonial society in Australia since 1788. How closely does Wikipedia reflect this growth? Figure 2.13 compares our dataset to historical population growth figures from the Australian Bureau of Statistics (2016). This data records the growth of colonial society, because Indigenous Australians were excluded from the census until 1961. From this data, we calculate the colonial population growth for each decade from the 1780s to the 2010s. For example, in the 1810s, Australia's population grew by 21,977 people. We then find the number of Wikipedia biographies for people whose date of birth lies in this decade, as in Figure 2.12, and calculate this as a percentage of the Australian population growth in that decade. For example, there are 662 biographies for people born in the 1810s in our dataset, which is equal to 3% of the colonial population growth for that decade. The two datasets are not perfectly comparable, because our data does not account for when a person migrated to Australia, but Figure 2.13 nonetheless clearly shows the historical bias of Wikipedia's Australian biographies. There are many more articles per person in the early years of the colony compared to later on in Australia's colonial history. Since the 1850s, the historical coverage of Australian Wikipedia biographies has been more equal.

This analysis by Historical Period complements our earlier analysis of Gender and Indigeneity in the dataset. Generally speaking, Wikipedia's Australian biographies reflect the male-dominated settler narrative of Australia's colonisation, though there are important caveats to this general finding.

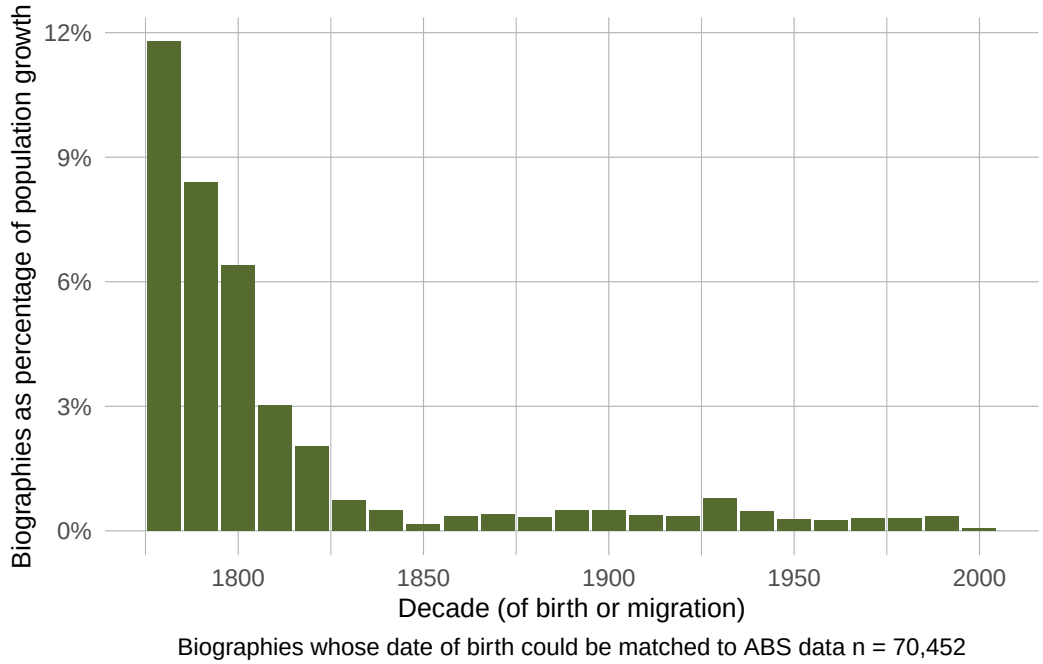

Figure 2.13: Number of articles in Australian biographies dataset compared to historical population growth statistics from the Australian Bureau of Statistics (2016).

3 Conclusion

In the report above we've articulated how Wikipedia represents Australians according to multiple mechanisms for categorizing and defining Australianness in Wikipedia and Wikidata's datasets and we've argued that Wikipedia's datafied representations highlight the problems of taking any overly simplistic view of Wikipedia's biases. The majority of quantitative studies of Wikipedia's biases examine its trajectories in terms of global trends – for example, in the representation of women and men, Westerners and non-Westerners. And yet, what we've shown in this report is that in taking a national perspective we recognize the contingencies and contexts that significantly enhance such analyses.

We hope, then, that this study reveals the importance of examining Wikipedia at a local (national) level. Only by thinking about and with the container of the nation, can the specificities, paradoxes, and discontinuities of Wikipedia's representation become evident. At a national level, we can start to understand how Wikipedia has represented women over time and how such representations compare to global trends. At a national level, we can ask new questions about people in ethnic categories that are particularly discriminated against and question whether the categorization norms are appropriate given alternative models that other nations have followed in relation to similar categories. At a national level, we can examine historical biases in the context of our knowledge about stories that the nation tells itself about who it is and where it comes from.

In this the methods of the discipline of history are helpful. Contingency, embeddedness, and human agency are central to the methods of history. Characteristics like identity or notability are not eternal, static, or natural. Although institutions such as nation states are hugely powerful and condition the experiences and behaviour of people, humans crucially have the power to act on and in the world. But they do so embedded in contexts that are always specific, both temporally and geographically. Moreover, historians, approach people in the midst of things: living in changing times and acting as best they can with the tools they have available to them (Pietsch and Flanagan 2020).

We show that it is possible to estimate how well Wikipedia represents the diversity of a given nation using page-level metadata. Counting pages is a good starting point, and can reveal representation gaps that Wikimedians can then attempt to redress. But counting pages is a difficult exercise, and must be performed sensitively. Rather than simply counting people with the most popular or prominent categories, we argue that is critical to also examine the semantics of the category structure itself. How might the dataset shift and morph depending on different semantic terms? How does the way in

which categories are structured suggest a particular perspective about national identity? How is what the nation values reflected in the characteristics of those who tend to have a Wikipedia page?

4 Appendix: Data

The code to generate this report is contained in two files in the wikihistories reports Github repository:

- who-counts-lib.R contains all the functions to import the data
- who-counts-data.R combines the functions to import the full dataset. This script can be run with `USE_CACHE <- TRUE` to import the data from the saved files for this report, or `USE_CACHE <- FALSE` to regenerate the data afresh from Wikipedia and Wikidata

All the source code for the analysis is included in this report, and can be viewed on Github at who-counts.qmd.

4.1 Wikipedia's category system

From its inception, Wikipedia has provided a category system to 'group together pages on similar subjects', and provide useful 'navigational links to pages in Wikipedia'. Wikipedia's categories form a vast and complex tree. A category may contain infintely many subcategories, and those subcategories may themselves belong to infinitely many supercategories. Nearly all Wikipedia articles are categorised, partly due to the work of a task group and patrol whose job is to hunt down any uncategorised articles.

In this context, the first place we looked for biographies of Australian people on Wikipedia was Category:Australian people, which 'lists notable Australian-born people, or people who identify themselves as Australian'. This category contains 99,626 articles, but not all are biographies. In the first case, there are many articles about groups of Australian people, e.g. Aboriginal Victorians or Australian Chamber Orchestra. In the second case, `Category:Australian people` is a vast category, with 9,492 subcategories. In the trackless expanse of these subcategories, some unexpected articles make their way into `Category:Australian people`, such as the classic glam rock album *Living in the '70s*. We explain the 'eccentricity' of the category graph when we analyse Wikipedia's definition of Australianness below. For now, we simply describe our method for filtering all the non-biographies out of `Category:Australian people`.

To filter out the non-biographies, we rely on Wikidata. We fetch the Wikidata item for each article in `Category:Australian people`, and check to see if it has property

P31(Q5), 'instance of:human'. While the possession of P31(Q5) is not a perfect signal that an article is a biography, we find that this criterion filters out nearly all the non-biographies. Applying this criterion, we identify 81,493 biographical articles in `Category:Australian people`.

4.2 Wikidata's query service

Wikidata provides an alternative method for finding Australian biographies: its query service. Using the query service, it is possible to search for Wikidata items that possess given properties. If these Wikidata items have a corresponding Wikipedia page, the link can be automatically retrieved.

It is more complex to find the 'Australians' in Wikidata than in Wikipedia. Wikipedia provides a single master-category for 'Australian people'. Wikidata does not. Instead, Wikidata can mark a person as 'Australian' in two main ways:

1. It can explicitly link a person to Australia (Q408) through properties such as citizenship (P27), place of birth (P19) or place of death (P570); or
2. It can implicitly link a person to Australia by identifying them in a database of Australians, such as the Dictionary of Australian Biography (P1907), Indigenous Australia (P9246), or Convict Records of Australia (P9919)

We queried Wikidata by 45 such properties suggesting Australianness. Using these properties, we were able to identify 85,773 Wikidata items representing Australian people. Of these Wikidata items, 63,692 items had a corresponding Wikipedia article.

The two datasets overlap substantially, but not completely (Figure 4.1). There are thousands of biographical articles in `Category:Australian people` whose Wikidata items do not clearly identify them as 'Australian'. Likewise, there are hundreds of Wikidata items whose corresponding Wikipedia articles do *not* appear in `Category:Australian people`, even though the Wikidata item has P31(Q5), 'instance of:human', and also contains one or more 'markers of Australianness'.

Dataset	Number of Articles
Person only found by querying Wikidata	1520
Person only found in `Category:Australian people`	19321
Person found in both datasets	62172

Figure 4.1: There is substantial, but incomplete overlap between the biographies in the two datasets

4.3 Metadata

We combine our two datasets into a single larger dataset of 83,013 articles. To enable analysis of these articles, we then enrich them with three kinds of metadata:

1. Using the Wikidata Action API, we find each person's dates, birthplace and gender.
2. Using the Wikipedia Action API, we download the introduction of each article for text analysis.
3. Using the XTools API, we download page usage data for each page, including the page creation date, assessment class (e.g. Good, Featured), pageviews and number of edits.

References

Australian Bureau of Statistics. 2016. "Historical Population." ABS Website. https://www.abs.gov.au/statistics/people/population/historical-population/2016.

Beytía, Pablo. 2020. "The Positioning Matters: Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia." In *Companion Proceedings of the Web Conference 2020*, 806–10. WWW '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3366424.3383569.

Butlin, Noel. 1994. *Forming a Colonial Economy, Australia 1810-1850*. Cambridge: Cambridge University Press.

Ford, Heather, Tamson Pietsch, and Kelly Tall. 2023. "Gender and the Invisibility of Care on Wikipedia." *Big Data & Society* 10 (2): 1–17. https://doi.org/10.1177/20539517231210276.

Konieczny, Piotr, and Maximilian Klein. 2018. "Gender Gap Through Time and Space: A Journey Through Wikipedia Biographies via the Wikidata Human Gender Indicator." *New Media & Society* 20 (12): 4608–33. https://doi.org/10.1177/1461444818779080.

Mandiberg, Michael. 2023. "Wikipedia's Race and Ethnicity Gap and the Unverifiability of Whiteness." *Social Text* 41 (1 (154)): 21–46. https://doi.org/10.1215/01642472-10174954.

Pietsch, Tamson, and Frances Flanagan. 2020. "Here We Stand: Temporal Thinking in Urgent Times." *History Australia* 17 (2): 252–71. https://doi.org/10.1080/14490854.2020.1758577.

Thorpe, Kirsten, Nathan Sentance, and Lauren Booker. 2023. "Wikimedia Australia and First Nations Metadata: ATSILIRN Protocols for Description and Access." University of Technology Sydney. https://doi.org/10.57956/B05F-CF08.

Weathington, Katy, and Jed R. Brubaker. 2023. "Queer Identities, Normative Databases: Challenges to Capturing Queerness On Wikidata." *Proceedings of the ACM on Human-Computer Interaction* 7 (CSCW1): 84:1–26. https://doi.org/10.1145/3579517.