Monitoring the Gender Gap with Wikidata Human Gender Indicators

Author Blinded
blind
blind
blind
blind

Author Blinded
blind
blind
blind
blind

Author Blinded
blind
blind
blind
blind

Author Blinded
blind
blind
blind
blind

Author Blinded
blind
blind
blind
blind

ABSTRACT

The gender gap in Wikipedia's content, specifically in the reperesentation of women in biographies is well-known, but has been difficult to measure and monitor its evolution. There exist efforts to address this gender gap, but the impacts they are having have received no attention. To investigate we utilise Wikidata, the database that feeds Wikipedia, and introduce the "Wikidata Human Gender Indicators" (WHGI), an open source, open data, real time, longitudinal, biographical dataset that can provide insights into gender disparities across time, space, culture, occupation and language. Through these lenses we show how women's representation has changed along 11 dimensions. Furthermore, to demonstrate it's more general use in research we present validations of the WHGI against three exogenous datasets: the world's historical population, "traditional" gender-disparity indices (GDI, GEI, GGGI and SIGI), and occupational gender according to the US Bureau of Labor Statistics.

CCS Concepts

•**Computer systems organization** → **Example;** *Redundancy;* Robotics;
•**Networks** → Network reliability;

Keywords

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

Gender inequality is a long-standing social problem which affects many aspects of society. Worldwide, cultural ideologies have created scenarios which make women more prone to health issues [23]. Likewise in education attitudes create a systemic gender bias in opportunity [9]. And, famously, incomes for identical jobs are lower for women [2].

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

Statistical gender indicators are critically important to understanding gender inequality, but their construction is difficult [14]. Many indicators (single measures) and indices (compound measures) have been proposed, such as The Gender Development Index from the United Nations Development and the Global Gender Gap Index from the World Economic Forum, but no academic consensus exists on which is superior [8]. Owing to this varied landscape using a plurality of inhat dicators is recommended for research [12].

We introduce an open source, open data, real-time, intersectional dataset that can provide insights into gender disparities across time, space, culture, occupation and language. "Wikidata Human Gender Indicators" (WHGI) is a dataset consiting of 11 separate indicators on the gender representation of humans in Wikidata.

Measuring and releasing data on Wikipedia's content gender gap attempts to solve two problems. The scope of Wikipedia, and the advent of its machine-readability through Wikidata, gives us an unprecedented look at gender at a scale never seen before. This provides potential for applications not possible until now. Secondly, we wish to shed light on Wikipedia's data under the philosophy of "what gets measured, gets fixed."

The thread of research on Wikipedia's gender biases originates in the finding that Wikipedia's editors are largely not women [10]. This imbalance has been attributed to an internet skills gap [7] and its internal culture [17].

More and more, in addition to investigations of the Wikipedia *editor gender gap* researchers have also been interrogating the character of its *biography gender gap*. Early studies found that Wikipedia excluded notable women more than its counterparts [19]. More recently [22] showed that while coverage of women in large Wikipedias is not less than other reference works, the language with which women are portrayed is different and focuses more on romance and family. Women also tend to be less central in the link graph of Wikipedia [4]. These linguistic and network findings were confirmed by [6], who also showed evidence of stereotyping in metadata.

Yet in popular mindshare there persists a sentiment that denies that any of this is a problem [3]. Luckily, experiments are showing that awareness of Wikipedia's gender issues is a strategy that can alleviate the problem [11], for which more methods are always needed.

Wikidata, the database that feeds Wikipedia, offers new opportunities to analyze culture programmatically. Launched in 2012, Wikidata is designed to host structured data that is *multilingual* (so

there is only one edition) and *plural* (can support many competing facts) [21]. These features make Wikidata the perfect place for all Wikipedias to collaboratively store facts about the world. If an Italian Wikipedian stores information about the population of ancient Rome, that information is then available to every other Wikipedia with a short code snippet. Every language collaborating together has meant that Wikidata has become a massive free open knowledgebase in its own right, containing over 40 million facts [16].

As a knowledgebase, Wikidata is slowly proving its worth for research. For instance, Wikidata has been used to find popular connections between nationalities and occupations [5]. Or take the fact that all human and mouse genes have been imported into Wikidata [18], for an internet-wide community effort to find links between genes, drugs and diseases [1]. All of these tasks would be difficult to do without Wikidata.

1.1 Outline

This paper begins by describing the format of Wikidata and statistics of humans contained in it. We investigate how gender composition and data quality has changed over time and demonstrate the impact that WHGI can have as a metric for content and women-focused Wikipedian communities.

Moving beyond Wikipedia navel-gazing, we present 3 validation measures utilizing ground truths from the US Census Bureau, Bureau for Labor Statistics, and United Nations Development Program. We show that WHGI does in fact relate to the real world. This means that, albeit imperfectly, the WHGI can be a proxy for numerical data about the real word in times and places for which no previous data exists.

2. HUMANS IN WIKIDATA

Wikidata is a general database consisting of *items* which are described by *properties* that take on *values*. Our interest is in biographies of people, that is any item which has the property *instance of* with value *human* [1]). For each human item we find the corresponding values of *gender*, *date of birth*, *date of death*, *place of birth*, *citizenship*, *ethnic group*, *field of work*, and *occupation* [2] In Figure 1, we illustrate the semantics of a Wikidata Human on the item for Aung San Suu Kyi.

For each of the above eight properties we create an "indicator" by aggregating the dataset on that property, but disaggregate by gender. Take for example the date of birth indicator, it has one row per year found as a date of birth, and one column per gender represented in Wikidata. See a sample excerpt of the date of birth indicator in Table 1, and it's visualization in 2. In addition to the eight indicators made directly from properties, we include three more which feature augmented data. There is a indicator based on the Wikipedia languages in which a human is represented. And we include a geographic aggregation of citizenship, place of birth and ethnic group into indicators called *culture*, and *worldmap*. See Figure 3 for the worldmap visualization.

2.1 Snapshots

All of our data is dervied from the official Wikidata database downloads, which represent a cross-sectional "snapshot" of Wikidata as it was at a specific date. Wikidata releases a new snapshot weekly. We re-process each of the 11 indicators for every new

[1]As Wikidata is intentionally multilingual, itesm, properties and values are actually reference by number. So "instance of: human" is "P31:Q5" in Wikidata terms

[2]These correspond to Wikidata properties P21, P569, P570, P19, P27, P172, P101, and P106 respectively.

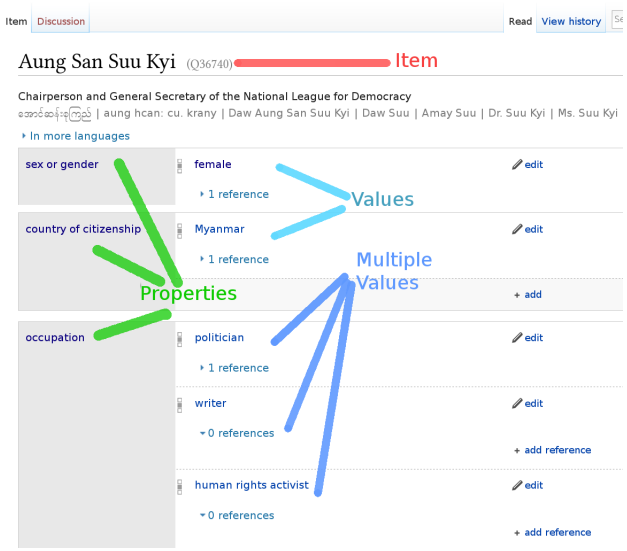

Figure 1: Example Wikidata Human Item of Aung San Suu Kyi

Table 1: A small excerpt of the January 3rd 2016 date of birth indicator illustrating the gender aggregation of Wikidata by birth year. The earliest humans in Wikidata are two men born in 4203 b.c.e., there are many notable births in the 1980's (see Figure 2), and there are thirteen notable 1-year-olds in Wikidata. Notice the inclusion of non-binary genders as they are recorded in Wikidata, as well as biographies without any gender recorded.

date of birth	no gender	trans-gender fe-male	gender-queer	ka-thoey	female	male
4203 (B.C.E.)						2
. . .						
1981	849	1		1	5,042	14,461
1982	861	2			5,132	14,372
1983	864	3			5,078	14,520
1984	830	3	1		5,372	14,558
1985	777	4			5,400	14,664
. . .						
2015	6				4	3

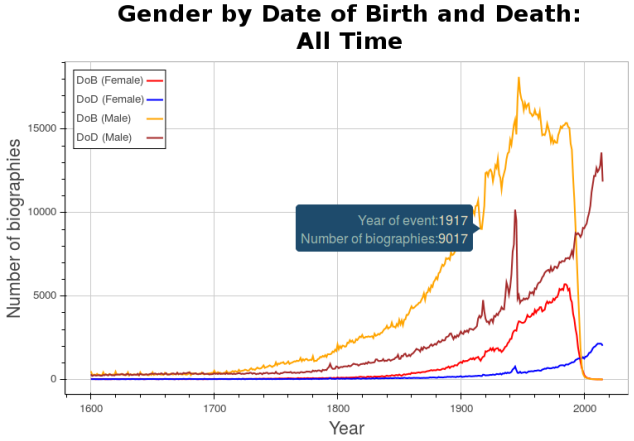

Figure 2: The total of Wikidata gendered biographies aggregated by date of birth and date of death. This data is represents the January 3rd 2016 snapshot. See the noticable spikes in death for men around World War II, and that births of both genders drop about 20 years before the current year, as younger people tend not to be notable.

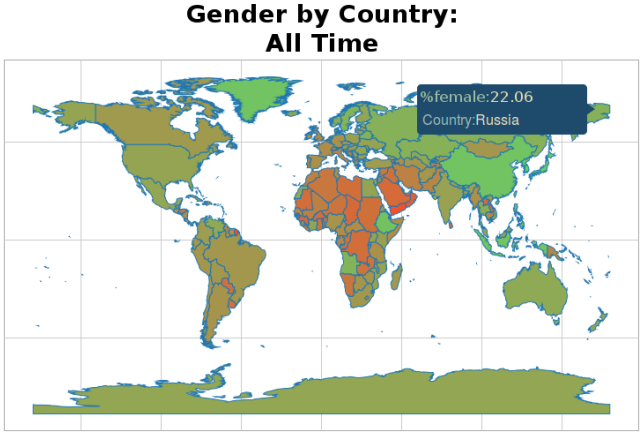

Figure 3: The Wikidata female ratio of biographies aggregated by place of birth and citizenship. This data is represents the January 3rd 2016 snapshot.

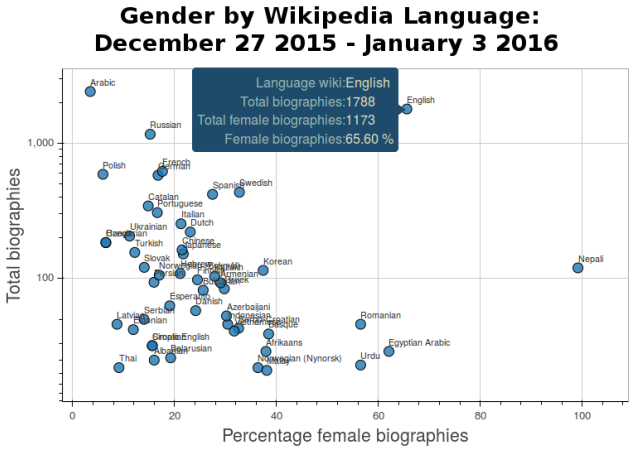

Figure 4: The changes in the size and percentage of female biographies for large Wikipedia languages in the period December 27th 2015 - January 3rd 2016. English Wikipedia Increased 1,788 biographies 65% of which were about women. Meanwhile Nepali Wikipedia increased by 120 biographies, 119 of which were about women.

snapshot, and additionally compute the differences that occured between the newest snapshot and the second-newest. This allows us to monitor activity on Wikidata at a weekly level of granularity. For instance Figure 2 and Figure 3 show the state of the date of birth and country indicators, *for all time*, as of the January 3rd 2016 snapshot. However Figure 4 shows the *changes* of the week between December 27th 2015 - January 3rd 2016.

Therefore we are also generating a dataset of *weekly changes* which allows us to monitor the status of biographies in Wikidata. We can inspect the changes in composition of genders, or date of birth, which can speak to efforts from Wikipedian communities attempting to counter bias in the database.

2.2 Technical Details

Note that for fidelity there is virtually no data-cleaning done, as the point of our project is to display information as faithfully as possible. Our dataset is meant to be used to uncover potential biases in Wikidata and the world at large, and we feel that any cleaning process would introduce further biases. An instructive illustration of this case is that the "gender" property in Wikidata is actually labelled in English as "sex or gender" (no distinction), and not limited to any value. Over our time snapshotting we found 36 values used for "sex or gender", including "male" and "female", but extending to nonbinary genders "transgender female", "intersex", "fa'afafine", "transgender", "Gender fluid", "genderqueer", "kathoey", and "queer". At times the other categories of information are recorded here - perhaps erroneously - such as "gay", or "homosexuality". And even what seem to be mistakes are left in, such as one-offs of "Solanum tuberosum", "Messi", or "sociologist". Cleaning this data would be a disservice, we feel, to communicating how - and how well - Wikidata is used.

Our dataset and code used to generate it is available for free under the CC-BY license online. Our first snapshot is from September 17th 2014, and tracks the official Wikidata data dumps, updating weekly. We archived the January 3rd 2016 version as a quality-checked, canonical version [3]. All our code to make this data and the analyses presented here SVs and generate the following results

[3]WEBSITEBLINDED

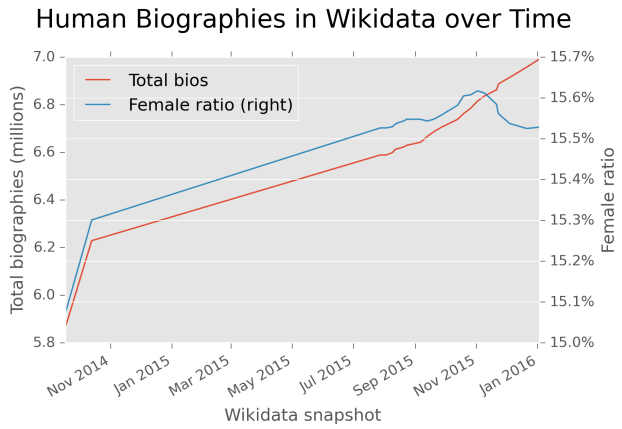

Figure 5: Total number of human biographies (left) and the female ratio of those biographies (right) by Wikidata snapshot. The total number of humans found in Wikidata over time is displaying linear, unconstrained growth. Over the same the female ratio of biographies in Wikidata has risen by 0.5%.

using both *python-pandas* and *R* can be found in our github repository [4].

Note that the missing data in the first half of 2015 is due to the period in which we were building the automation of collecting these statistics.

3. LONGITUDINAL STATISTICS

A main purpose for investigating this dataset is to support and provide metrics for Wikipedians communities attempting to address content gaps. Therefore we turn to focus at statistics of our dataset with regard to how it has changed over time as these Wikipedian communities have been edting.

First we queried the way the total number of biographies and the ratio of women represented as Wikidata has evolved. Total humans in Wikidata increased from 5,869,606 to 6,999,542, and shows linear, unconstrained growth (see Figure 5). Certainly Wikidata is activte and changing, but how? An important measure for content-focused communities is the ratio of biographies which are about women, as espoused in *WikiProject Women in Red*[5]. We looked into the ratio of humans recorded "female" versus all gendered biographies. Similar to total biographies this measure is rising at a fairly linear rate of approximately 0.5% per year5. The final months on record show a slight decline which warrants further investigation.

We took a more granular look at the evolution of the female ratio of biographies by disaggregating by Wikipedia Language in Figure 6. Of the languages that have 100,000 or more gendered biographies, and during the time period July 2015, to January 2016, we see the rate of women in Norwegian (Bokmål), Spanish, and English Wikipedias each increase by more than 0.25%.

Fortunately Wikiproject Women in Red keeps metrics of how many biographies they added on a monthly basis, and we were able to conduct a cross-correlation between the monthly number of biographies created by Women in Red, and the number of female biographies added to English Wikipedia. The correlation between these activities is 0.657, which indicates that they are indeed related. Unfortunately we cannot determine a casual relationship be-

[4]WEBSITEBLINDED

[5]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

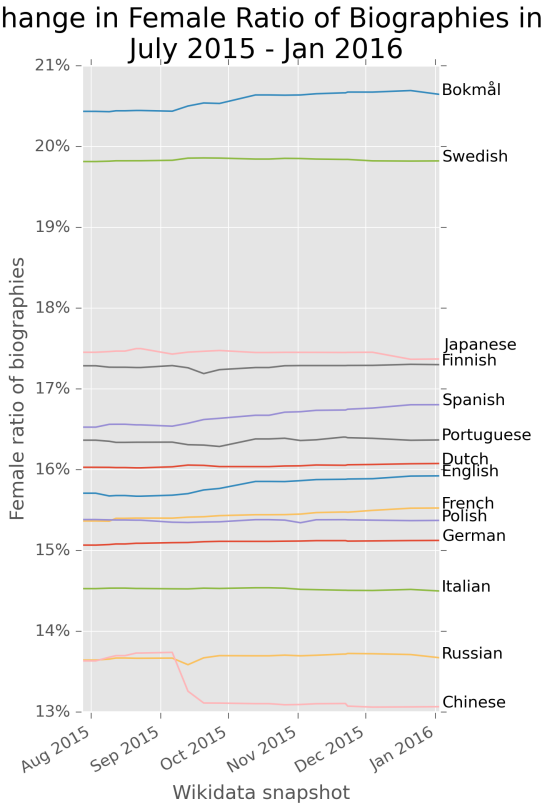

Figure 6: The change in female ratio of biographies over time, by Wikipedia languages with 100,000 or more gendered biographies. Even during this short time period, the Norwegian (Bokmål), Spanish, and English Wikipedias all show visible up-ticks in representing Women.

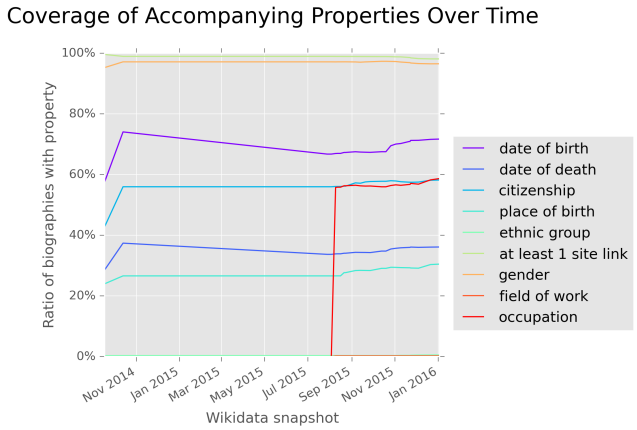

Figure 7: Trends of property coverage by Wikidata snapshot. Most humans have at least one Wikipedia article ("site link") and a recorded gender, other properties are slowly increasing in coverage.

Table 2: Change in rates of property coverage for humans

	2014-09-17	2016-01-03
gender	95.3%	96.5%
date of birth	57.6%	71.7%
date of death	28.6%	36.1%
citizenship	42.8%	58.2%
place of birth	24.0%	30.5%
ethnic group	0.3%	0.6%
field of work	n/a	0.3%
occupation	n/a	58.7%
at least 1 site link	99.6%	98.1%

tween editing efforts and the increase in women's representation in these languages. Still we are able to numerically highlight this trend which was not viewable before.

3.1 Data Quality

Of course the female ratio of biographies is not the only way to characterize the effect of content-focused editing, we might also inquire to the wider quality of biographies in Wikidata. One way to investigate data quality is the coverage of demographic properties on these biographies. Figure 7 and Table 2 show the trend in coverage of all properties at the earliest and latest snapshots. The statistics show that data quality has been increasing across all properties over time. The number of humans with *gender* data increased just 1% point but is close to complete coverage. In the time domain *date of birth* and *date of death* coverage increased by 14% points and 7% points repspectively. Likewise *citizenship* data increased the most, by 15% points. *place of birth* increased by 6% points, and *ethnic group* doubled to 0.6%. *Field of work*, and *occupation* data was not included in our dataset until later, so their growth, while increasing is not precisely comparable.

Curiously the rate of humans having at least on Wikipedia article decreased slightly, but this has an important interpretation. A Wikidata human without a Wikipedia article is known as a "structural item", for instance a member of royalty without a Wikipedia article but is needed to make a family tree complete. With the view that a structural item is an artefact from editors paying attention to Wikidata's structure, the decrease in sitelinked humans can also be seen as an increase in data quality.

There may also be biography articles in Wikipedias that are not recorded as humans in Wikidata, however it is not directly computable how much growth in Wikidata stems from Wikipedia's growth (e.g. a new biography is added), or migration of information to Wikidata (e.g. an existing biography in Wikipedia is marked as a human in Wikidata).

Table 3: Correlation of number of WHGI by date of birth and world population. Significances are $^{**}p \leq 0.01$.

snapshot	Pearson correlation
2014-09-17	0.852**
2016-01-03	0.845**

Table 4: WHGI-country correlation to external indices. Correlation is the Spearman ρ, and significances are $^{*}p \leq 0.05$, $^{**}p \leq 0.01$.

snapshot	GEI	SIGI	GGGI	GDI
2014-09-17	0.417**	0.338**	0.310*	0.278**
2016-01-03	0.457**	0.402**	0.386**	0.299**

4. VALIDATION

Another of our main purposes in creating WHGI was to contribute to the landscape of gender-disparity indicators. In order to gain an idea about how well WHGI reflects the real world we validated our data by comparing it against 3 exogenous datasets. We correlated the WHGI by date of birth versus historical world population trends; WHGI by country versus exogenous gender-disparity indices; and WHGI by occupation versus United States Bureau of Labor Statistics occupation by gender.

4.1 World Population

Our first validation is a "sanity check" to compare the world's population by year to the number of humans in WHGI by year of birth. We conduct this validation even though, the number of people alive and the number of Wikipedia-notable people born are different measures. However if we operate under the assumptions that (a) the proportion of the world population which is Wikipedia-notable is constant over time and (b) that the birth rate is a fixed proportion of the population, then theoretically their curves should share approximately the same shape.

We performed a standard Pearson correlation between the number of people in Wikidata born in a particular year, and the estimated historical world population by the US Census Bureau[6]. We conducted this correlation for our earliest and latest snapshots - the population statistics of Wikidata at September 2014, and again separately at January 2016. The results in Table 3 show a high and significant correlation between real world estimates and Wikidata, at about 0.85. We do see a very minor decrease in correlation over snapshots of 0.007. Overall though the population of Wikidata over time seems very aligned with the World's population over time, so Wikidata at least is a "sane" representation of the world.

4.2 Exogenous Gender-Disparities Indices

WHGI is inspired, in part, by the rich landscape of gender disparity indices. This type of index ranks countries by a measure of gender equality. If we aggregate WHGI by place of birth and

[6]https://commons.Wikipedia.org/wiki/File:Population_curve.svg

citizenship, and look at the female ratio of humans, we too have a sort of country-by-gender equality measure[7]. We correlated the country rankings of this WHGI aggregation with 4 popular exogenous indices to see how well Wikidata reflects real world gender disparities.

The 4 exogenous indices we used were: The traditional United Nations' Gender Development Index (**GDI**) [8] which considers disparity in income, education, and life expectancy. Social Watch's Gender Equity Index (**GEI**) [9] tries to broaden the scope of the variables by incorportaing education and economic participation, but also stretching into economic and political "empowerment". The Global Gender Gap Index (**GGGI**) [10] grows yet wider by covering all previous topics and along more detailed dimensions. And most recently the Social Institutions and Gender Index (**SIGI**) [11] has attempted to capture disparity in norms, values and attitudes.

Additionally we conducted a calibration step, to find the date of birth threshold which maximized our correlations. In each case the maximizing threshold was found to be between 1900 and 1910. We interpreted the found thresholds as a good sign firstly because the exogenous indices are measures of recent history too, and secondly because it shows a robustness in the way that WHGI relates to exogenous indices.

We repeated the index correlations twice, once using the data September 2014 snapshot of Wikidata, and then using January 2016 data. That is, we have correlations between rankings of countries given by exogenous indices and Wikidata - at two separate times.

Table 4 shows the correlations with each index, all of which were significant and ranged from 0.278 to 0.457. Affirmingly, when looking at this information through a longitudinal lens, the correlation with every index is increasing over time at which we sampled Wikidata. On the low end the GDI correlation grew by 7.6%, and on the upper end, the SIGI correlation jumped 24.5% in the about-a-year time frame between Wikidata snapshots. Because we are using the female ratio of biographies by country and not the absoulte number these correlation are not growing simply because of increased number of data points.

Previous analyses showed that WHGI being most closely related to GEI, and least to GDI has implications for Wikipedia's notability policy. Where they both measure gender gap in school enrollment, years of schooling, and earned income, GEI additionally measures positions of power, and GDI life expectancy. That means that notablity in Wikipedia is more related to power in society than it is to health status [15]. This analysis is remains true in 2016, as the order of the strengths of the index correlation has remain unchanged. Still the the strengths of those correlations has increased across all indices, which means that the gender disparities found in WHGI by country are increasingly looking more like these real world gender disparities.

4.3 Occupation Gender

The notion of what a human's job or occupation is, we saw in Table 2, well recorded in Wikidata. To answer the question of how representative of the real world Wikidata's gender by occupation is, we compared it to data from the United States Bureau of Labor Statistics (BLS) [12]. We borrow this ground truth technique from [13] who used it to evaluate the gender representation of Google image search results.

Table 5: Rank correlation of gender ratios by occupation between WHGI and US Bureau of Labor Statistics. Signficances are $^{**}p \leq 0.01$.

snapshot	Spearman Rank Correlation
2015-08-09	0.410**
2016-01-03	0.473**

Approximately 60% of our sample have occupation data, and together over 4,000 occupations are represented. The BLS has 332 occupation categories which are at a higher level ontologically than recorded in Wikidata. Whereas Wikidata might record that someone is a pastry chef, the BLS only has a category for cooks. In order to match the datasets we used Wikidata's internal ontology hierarchy, to generalize the occupation terms. A *subclass of* property exists in Wikidata, that relates items to their more general concept - which we can use for occupations. Wikidata describes that pastry chef is a *subclass of* chef, and that chef is a *sublcass* of cook.

Our method was to raise the generality of Wikidata occupations until there were less than 500 occupations to ease the matching task. Two authors then matched occupations manually for accuracy and confirmation. We resolved disagreements until the sets were matched. However not all occupations could be matched due to the specificity of the BLS, rendering coverage of Wikidata occupations 57% complete. The largest occupations in Wikidata were sportsperson and politician, and neither of them had matches in the BLS. In the reverse, there were many BLS occupations for which Wikidata did not have any matching occupations, such as "lodgings manager". This outlines a limitation of this validation, that being a lodgings manager does not inherently make you notable for inclusion in Wikipedia. It must be acknowledged too that the BLS data describes the United States whereas WHGI has a worldwide scope, which may explain why we found no significant correlation in the *size* of the matching occupations between the two sets.

Finally we correlated the rankings of the list of most gendered occupations according to WHGI to that of the BLS. We did this for early and late snapshots, but because occupation was not a property that we initially recorded, our first snapshot which included occupation was August 9th 2015. Table 5 shows the spearman rank correlation found was a significant 0.410, and since then the correlation has increased to 0.473. These are moderate correlations which we claim support a link that Wikidata reflects the real world.

5. REPRESENTATIVE LIMITATIONS

The WHGI is measuring two phenomena that we do not disentangle: human development and Wikipedia content development. On the one hand the validation of the dataset tells us how well our measurements capture the various dimensions of gender equality and human development. On the other hand, we are also inspecting how Wikipedia's biography articles and notability policy are based in the real world bias.

To some degree WHGI represents the real world. In each of our validation measures we found high or moderate correlations. Certainly WHGI is not random or isolated, but captures real world dynamics with some distortion. We notice both that WHGI validations are rising over time, and that data quality is rising over time. This can be taken to mean that as Wikidata becomes more complete it is modelling the real world more. There is some justification in using WHGI as a proxy for real-world phenomena, but that proxy

[7]Despite having the same by-country unit of analysis with this aggregation WHGI is not an "index" like those we compare it to, since an index weights and combines many indicators [20].

[8]http://hdr.undp.org/en/content/gender-development-index-gdi

[9]http://www.socialwatch.org/node/14366

[10]http://reports.weforum.org/global-gender-gap-report-2014/

[11]http://www.genderindex.org/ranking

[12]http://www.bls.gov/cps/aa2012/cpsaat11.htm

is limited by the worldview of Wikipedia editors, and constrained by its notability policies.

Wikipedia's notability policies require humans to be in positions of power which are systemically biased against women (AUTHOR BLINDED), how then can the rise in women's biographical representation be explained? There could be several possible reasons. At least three factors that affect encyclopedic inclusion are: (1) the rate at which women receive positions of power in the real world, (2) the level of gender bias in Wikipedias' notability policies, and (3) the level of efforts to write about women in Wikipedia. From Figure 6 it seems that there may be some language specific effects.

6. IMPACT OF WHGI

There are many Wikipedian communities who's goal it is is to increase the coverage of Women's biographies, for instance: WikiProject Women Scientists [13], Art + Feminism[14], and Women in Red, just to name a few. One concern of these organizations is if their editing efforts are making large scale impacts on the Wikipedia. Luckily Women in Red have been keeping data on their activity levels.

We compared the number of articles added to English Wikipedia by the Women in Red project to the number of biographies added to English Wikipedia marked as women in Wikidata. The correlation between these two activities was 0.657. This medium-high correlation shows that indeed the editing levels of this group are related to the growth of female biographies in Wikidata. This provides a basis to use changes in WHGI as an effectiveness metrics for editing. We hope to find more activity data on the content-focused editing communities, particularly non-English ones to be able to perform "propensity score matching" between languages. That is, for a language which has much women-focused editing we could compare the way that language's gender compostion changes against a language which was on a similar trajectory before the change, but had less women-focused editing.

Also for the indirect effects of policy changes and other initiatives that are not specifically article-creation focused we could still measure their change on biography gender. As an example, the gendered implications of changes to a notability policy could be tested using WHGI using the same propensity score matching above. We hope now that this high-level metric is being tracked it will provide an incentive to track more specific editing activity.

Furthermore, applications could be built to detect spikes in creation and deletion of specific demographics of humans. Such a tool could also alert to the presence of unplanned activity, good or bad, which affects the macro-level gender of Wikipedia and Wikidata. Take Figure 4, it shows a week where contributions to Nepali Wikipedia are nearly 100% about women. Likewise, if a week were to show a net-subtraction of female biographies a community alert could be generated.

Divorced from Wikipedia entirely, a historian could use the data to determine the gender-disparity levels of a specific place and time. Typically to quantify the gender climate one would rely on the indices like those mentioned in the exogenous indices section. However these indices, are limited to discussing recent history. Our validation showed that our data is in touch with the real world. With this dataset we can quantify a type of gender-disparity of medieval France, ancient Greece, or Ming dynasty China. WHGI is useful in all the same ways that exogenous indices are used, only with a larger timespan. That is certainly a novel approach not possible before Wikidata.

Yet another new avenue this dataset opens is in the gender-disparity of a language. A linguist could use WHGI aggregated by language to quantify the gendered-ness of a language. Furthermore with the date of birth and death information, the linguist could see how languages have focused on gender differently over time. Potentially this could lend evidence to another theory of language that comes from their native methods.

WHGI is, in essence, a biographic database. The data can not only provide insights on gender-related disparity, but also other disparities such as culture disparities, citizen disparities and ethnic group disparities, etc.

7. CONCLUSION AND FUTURE WORK

We made the Wikidata Human Gender Indicators (WHGI), a biographic database for researchers wishing to incorporate gender data along dimensions of time, space and occupation. Based off of Wikidata and Wikipedia it can most obviously be used by those communities to monitor the effects of focused editing and biases in their content. We also validated the indicators with measures of the real world, such as population, country-based gender disparities, and occupations. These validations showed that the WHGI is significantly correlated to real world demographics and gender disparities. We also showed that data quality of Wikidata has been increasing. Data quality and correlations increasing together is particularly encouraging as support for using WHGI as a tool. WHGI is freely available for download, we have outlined some of the potential ways in which it could be used, and hope that many more are thought of by others.

We hope to continue running the open source project in service of the Wikipedia and research communities seeking to statistically describe gender disparities. The ways in which we expand WHGI we hope will be directed by user's feedback.

8. ACKNOWLEDGMENTS

We are especially grateful to the Wikimedia Foundation for funding us through an Individual Engagement Grant. [GRANT BLINDED].

We are also especially grateful to the Wikidata and Wikidata Toolkit teams.

Finally we acknowledge the hard work done by Wikimedians in countering systemic bias.

9. REFERENCES

[1] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. E. Putman, J. Leong, P. Pavlidis, L. Schriml, B. M. Good, and A. I. Su. Wikidata as a semantic framework for the Gene Wiki initiative. *bioRxiv*, page 032144, Nov. 2015.

[2] P. Burstein. *Equal Employment Opportunity: Labor Market Discrimination and Public Policy*. Transaction Publishers, 1994.

[3] S. Eckert and L. Steiner. (Re)triggering Backlash: Responses to News About Wikipedia's Gender Gap. *Journal of Communication Inquiry*, 37(4):284–303, Oct. 2013.

[4] Y.-H. Eom, P. AragÃşn, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PLoS ONE*, 10(3):e0114825, 03 2015.

[5] D. Goldfarb, D. Merkl, and M. Schich. Quantifying Cultural Histories via Person Networks in Wikipedia. *arXiv:1506.06580 [physics]*, June 2015. arXiv: 1506.06580.

[13] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_scientists

[14] http://art.plusfeminism.org/about/

[6] E. Graells-Garrido, M. Lalmas, and F. Menczer. First Women, Second Sex: Gender Bias in Wikipedia. *arXiv:1502.02341 [cs]*, pages 165–174, 2015. arXiv: 1502.02341.

[7] E. Hargittai and A. Shaw. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, Communication & Society*, 18(4):424–442, Apr. 2015.

[8] A. Hawken and G. L. Munck. Cross-National Indices with Gender-Differentiated Data: What Do They Measure? How Valid Are They? *Social Indicators Research*, 111(3):801–838, Apr. 2012.

[9] C. Heward and S. Bunwaree. *Gender, Education and Development: Beyond Access to Empowerment*. Palgrave Macmillan, Feb. 1999.

[10] B. M. Hill and A. Shaw. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE*, 8(6):e65782, June 2013.

[11] M. Hinnosaar. Gender Inequality in New Media: Evidence from Wikipedia. 2015.

[12] J. P. Jütting, C. Morrisson, J. DaytonâĂŘJohnson, and D. Drechsler*. Measuring Gender (In)Equality: The OECD Gender, Institutions and Development Data Base. *Journal of Human Development*, 9(1):65–86, Mar. 2008.

[13] M. Kay, C. Matuszek, and S. A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. pages 3819–3828. ACM Press, 2015.

[14] S. Klasen. Gender-Related Indicators of Well-Being. Technical Report 102, Discussion Papers / UniversitÃd't Göttingen, Ibero-Amerika-Institut fÃijr Wirtschaftsforschung, 2004.

[15] M. Klein and P. Konieczny. Wikipedia in the World of Global Gender Inequality Indices: What the Biography Gender Gap is Measuring. In *Proceedings of the 11th International Symposium on Open Collaboration*, OpenSym '15, pages 16:1–16:2, New York, NY, USA, 2015. ACM.

[16] M. Krötzsch. How to use Wikidata: Things to make and do with 40 million statements - korrekt.org, 2014.

[17] S. T. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 1–10, New York, NY, USA, 2011. ACM.

[18] E. Mitraka, A. Waagmeester, S. Burgstaller-Muehlbacher, L. M. Schriml, A. I. Su, and B. M. Good. Wikidata: A platform for data integration and dissemination for the life sciences and beyond. *bioRxiv*, page 031971, Nov. 2015.

[19] J. Reagle and L. Rhue. Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0):21, Aug. 2011.

[20] R. J. Rossi and K. J. Gilmartin. *The handbook of social indicators: sources, characteristics, and analysis*. Garland STPM Press, 1980.

[21] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85, Sept. 2014.

[22] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv:1501.06307 [cs]*, Jan. 2015. arXiv: 1501.06307.

[23] World Health Organization, editor. *Women and health: today's evidence tomorrow's agenda*. World Organization, Geneva, 2009.