

humaniki

User Research

Understanding how data powers diversity efforts in Wikimedia Movement.

October 2020

Sejal Khatri
sejal.khatri5@gmail.com

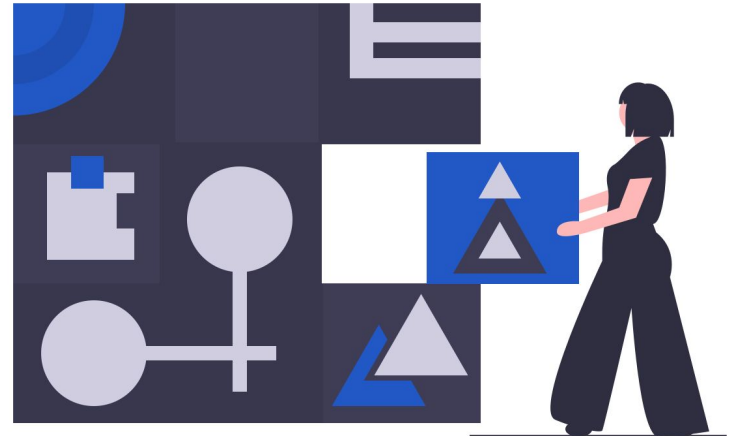


Illustration from undraw.co

01 Introduction

02 Research Background

03 Methods - Interviews

04 Findings - Personas

05 Discussion - Themes

06 Implications for Design - Next Steps

07 Appendix

Project / Context

Introduction

Introduction

Problem Space

Wikimedia projects hold several million biographies, which form an incredible dataset about *WHO* the project finds important. However without further analysis and tools, it's difficult to answer some prominent questions about those biographies like:

- how diverse are they?
- which diversity gaps would be easy to fill?
- how are these statistics changing over time?

humaniki

humaniki is a merger of two previous Wikimedia projects [Wikidata Human Gender Indicators \(WHGI\)](#) and [Denezh](#), that provide data to evaluate aspects of the **diversity in biographies**. This project generates and tracks statistics about who is represented in Wikimedia projects, helping to address diversity gaps with data tools.

This project is Wikimedia Foundation grant-funded, being built in collaboration with the Wikimedia community.

Project team

Meet our team:



Maximilian Klein
Data Scientist



Sejal Khatri
UX Researcher/Designer



Eugenia Kim
Front End Developer



Envel Le Hir
Data Engineer

Project / Research

Research Background

Research Background

Objectives

The objective of this research study is to elicit feature requirements by understanding the challenges community members face with the current tools, along with identifying the opportunities of collaboration and integration with teams leading diversity efforts in the community.

The objectives of research are guided by previous discussions and community engagements on the talk pages and other channels.

Research Question:

How might we provide usable and actionable insights for diversity focused editors and community members so that they can inform diversity driven efforts for Wikimedia Projects?

Key Objectives

- Enhance the user experience of current tool
- Identify integration opportunities
- Identify other diversity metrics needed by the community

Literature Review

We followed the participatory design approach wherein we involved a defined set of stakeholders in our research process to democratize the process of the designing the new tool. We conducted a [literature review](#) of existing community efforts in the similar direction to inform our research strategy and identify stakeholders for our research study.

Literature review sections:

1. [Relevant user groups - Wikiprojects that tackle systemic bias](#)
2. [Current tools that support user groups](#)
3. [Previous feedback and suggested features on WHGI and Denelezh](#)
4. [User Research Approach - Process Flow](#)

Project / Research / Methods

Methods - Interviews

Interview design and recruitment

We used statistically non-representative stratified sampling to build a sample of **23** community stakeholders which also involved users of the WHGI and Denelezh tools. We made effort to recruit people from all genders, as well as both leaders and non-leaders from each wiki project. Our final sample included 8 Men and 13 Women, with 2 undisclosed participants.

We recruited participants by posting screening surveys on each community's discussion channels including on each community's "Village Pump" community announcement page, social media pages, and news channels. We used snowball sampling and purposive sampling to increase our reach to participants who are the primary user base for our tool.

Survey Tool:

We used a secure online survey tool, Qualtrics, for our [screener survey](#). We collected information about respondents' demographics, wikipedia tenure and wiki project involvement to help us shortlist a inclusive and balanced subsets of participant groups. We provided Interview [consent form](#) and [information sheet](#) to all participants who signed up for interview study.

Interview design and recruitment

Response Rate

Amongst the 24 survey respondents, 23 consented to participate in survey research, 13 consented to participate in interview research (4 Men, 9 Women).

Interview Protocol

[For Editors](#), [For Researchers/Organizers](#).

Reimbursement for survey participants

We reimbursed community members who signed up through the screener survey for their time. The reason for reimbursing participants was to motivate members to participate in our research and in turn, to increase response rate from diversity focused editors in the community.

Participants Summary

Following is a summary about the number of different participant profiles that we interviewed:

No.	Participant Profile
8	Diversity-focused editors (editors who participate/organize diversity events, make contributions to improve representation of underrepresented groups)
7	Community representatives from several diversity organizations such as Women in Red, WikiDonne, Wiki Gap, WhoseKnowledge, and Art+Feminism.
3	Researchers from gender and ethnic studies
4	Technical experts and developers of other similar tools

Research Methodology

We used qualitative research methodology to understand community frustrations and needs with the goal to elicit feature requirements for the new tool. Our **23 interviews** lasted between 50 mins to 70 mins for an average of 60 mins and a total of 23 hours of interview data. We employed grounded theory approach to analyse our interview data, using Taguette, a open source qualitative data analysis tool. We used axial coding to identify our key themes mapped with feature requirements.

We used three types of interview sessions to learn from participants:



Semi-Structured Interviews

In the interviews with technical experts, community representatives, and researchers, we focused on discussing potential integrations, community needs, and interface guidelines to better inform our design.



Open Card Sorting

In the interviews with editors, we conducted classic co-design activities like open card sorting where they suggested new diversity statistics that should be tracked along with ways in which current statistics can be made more usable.



Group Interviews

In group interviews with editors from same wikiprojects, we were able to identify common challenges and gauge the current state of community discussions by facilitating a co-operative discussion environment.

Project / Research / Findings

Findings

Personas

From interviews, we developed a deeper understanding of our 4 participant user groups, which guided our definition of 6 identified archetypes of humaniki's potential users.



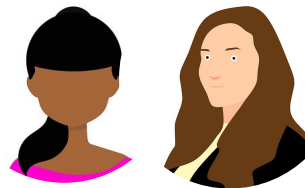
Carina and Amina

Carina and Amina are community leaders, Amina is also an Activist.



Mia

Mia is a community developer from Swedish Wikipedia.



Gurleen and Stef

Gurleen and Stef are diversity focused editors. Gurleen is multilingual and is from non-English language edition.



Emilia

Emilia is a academic researcher and has background in studying underrepresented communities.

* The names are fictional

Carina

Wikiproject Representative

Carina is long term Wikipedia and has been working towards improving content diversity on Wikipedia, focussing on women articles. She is closely connected with local and international community and is aware of the tools available in the space. She strives to increase the awareness of gender gap on Wikipedia projects to direct resources (events, meetups) in that direction.

“We are sitting with spreadsheets, manual counting that takes forever. There is really a problem with a lack of sufficiently pro content tracking tools.”



Concerns

The labels used for structuring the data on wikidata vary based on the language edition.

There is conflicting notability criteria by language editions - “We face notability problems, we have to control name by name if it is notable or not.”

Observations/Needs

Key point: Level of relevance - I want to find relevant diversity statistics based on location and field of interest.

Key point: I find it hard to follow the progress in the content growth of underrepresented groups.

Goals

- Identify and bridge the knowledge gaps on Wikimedia projects
- Communication with community and general people to create awareness.

Opportunities

- Track knowledge gaps
- Provide easy to use tools to search for relevant (local) content recommendations/stats.

Tools in use

- Programs and Events Dashboard
- Wikidata query
- Listeria Bot
- Wikidata Human Gender Indicators
- Denezh
- WikiGapFinder

“

It is hard to do a good query search which will rule out the objects from Wikidata that don't meet relevance criteria on Wikipedia. This ends up with a list where quite a few redlinks are actually relevant.

– [Wikiproject Representative](#)



Technical Challenges

We found that developers and data scientists working on tools in the diversity domain face challenges in identifying good ways to index data based on categories of relevance like country, language, field of interest, etc. This could be either due to information gaps in structured data sources like Wikidata or shortage of resources to support an architecture for this purpose.

With the global scale usage of wikimedia infrastructure and diversity of development efforts, identifying ways to support architecture for different initiatives would play a key role in sustainable development.

Amina (activist)

Wikiproject Representative

Amina is an activist who works towards improving the representation of LGBTIQ people on Wikimedia projects. Amina started by making “snail trails of revolution”, and has organized grassroots level activities in the community. Amina works towards creating a space, where activists can think out loud, and map out our thoughts to brainstorm ideas.

“This is alarming but it needs to be translated for people to understand that it is alarming otherwise its like its happening in another world, in English and nobody is thinking about it”



Concerns

Receiving consent from individual members to identify and categorize them as a member of the LGBTIQ community. Also, stringent citation rules on Wikipedia with most of available sources being local while Wikipedia only accepting articles from high profile publications.

Observations/Needs

Key point: We are not aware of the tools available in the community

Key point: Our efforts are grassroots level, conducted manually without using a structured system.

Key point: It has been only 2 or 3 years that the space was opened to actually discuss visibility of women or queer people in our country.

Goals

- Compiling and writing biography from most important persons from LGBTQ community.
- Identifying resources and tools that can help inform or support this movement.

Opportunities

- Create awareness and provide accessible tools to members from underrepresented communities

Tools in use

- No Tools

Mia

Community Developer

Mia is a community developer who helps community representatives and event organizers curate lists for diversity editathons. Mia also helps volunteers navigate the resources and access tools available for automating manual processes.

“I don't want to write a sparql query that 'times-out' anymore”



Concerns

Relevant list making: On tools that support article recommendations for article translations between language editions (like WikiGapFinder), developers find it difficult to generate relevant lists because currently filters like citizenship and occupation aren't available in conjunction with translation tools.

There is lack of structured data availability for non-gender bias like race data.

Observations/Needs

Key point: We use [listeria](#) list with [SPARQL query](#) to scale list generation.

Key point: We want to create tools that can be used by volunteers.

Goals

- Standardize statistics for non-gender bias
- Provide seamless interface for campaigns
- Provide relevant interwiki translation opportunities

Opportunities

- Improve usability of tools
- Provide completion metadata like (en->swedish women) has 25% overlap
- Customized list making based on relevant factors like country and language.

Tools in use

- Wikidata query
- Listeria Bot
- WikiGapFinder
- Wikidata Human Gender Indicators
- Denezh

“

We would not like to make people write SPARQL in the first place. We want a tool that can be used by volunteers.

– Community Developer



Technical Accessibility

We found that most of tools available to facilitate diversity driven efforts like list making with listeria bot and SPARQL for wikidata query search have technical barrier of participation. Developers from our sample set recognize this and are constantly looking for tools that are easy to use for volunteers who don't have a technical background.

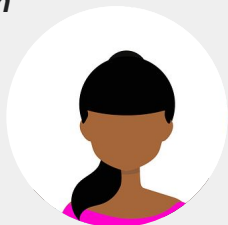
Gurleen (non-native English user)

Diversity focused Editor

Gurleen is a editor from a non-English Wikipedia language edition. She occasionally attends and organizes diversity focused events. Gurleen is a **inclusionist**.

Gurleen came to know about gender gap on Wikipedia by her own experiences with her women article contributions being rejected, as well as by awareness campaigns led by Wikiproject representatives.

“Sometimes we feel that tools with English Interface only track the English written information on Wikidata, it's a knowledge gap!”



Concerns

Tools not available in native languages.

Notability conflicts for women articles and high deletion rate. Leadership dominated by Men in small size language editions.

Observations/Needs

Key point: I am not a coder, I look at other people's queries and adapt them.

Key point: If data could be shown interactively, it would be much easier to convince grant making organizations.

Key point: I have observed that Men editors want to focus on needs of readers i.e mainstream subjects like politics, media, etc rather than writing about women. (Hindi Wikipedia user)

Goals

- Tools that can help create awareness of knowledge gaps on Wikipedia to drive resources in that direction

Opportunities

- Awareness of tools amongst non-English user base.
- Internationalization - make tool interface available in other languages.

Tools in use

- Wikidata query
- Wikidata Human Gender Indicators

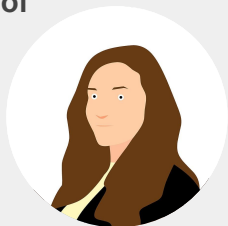
Stef (native English user)

Diversity focused Editor

Stef is a editor of English Wikipedia language edition and has participated in diversity editathons and events. She is a **inclusionist**.

Stef actively engages in the English and International wikimedia communities. She also contributes to wikidata by updating items with relevant gender categories.

“I see a lot of efforts to create new articles on women becoming a magnet for deletion, women of color being nominated for deletion”



Concerns

Preserve diversity: Notability conflicts and large scale article deletions.

As we fill the article gaps from past century, the gender gap would further inflate as most articles are of white men.

Observations/Needs

Key point: I have gone through big rejected/deleted piles of articles to restore deleted articles.

Key point: I want to be able to filter out languages that I am most interested in, languages I am interested in are really big in terms of global usage.

Goals

- Track article deletions to preserve diversity

Opportunities

- Provide information about gender representation on Wikidata
- Provide options to customize visualizations based on individual use cases
- Enable tracking or provide data of article deletions along with article creations

Tools in use

- Wikidata query
- Listeria Bot
- Wikidata Human Gender Indicators
- Denezh
- WikiGapFinder

Emilia

Academic Researcher

Emilia is a subject matter expert and is a academic researcher in a recognized institution. She conducts research in the diversity space and also teaches in a similar domain. Emilia thinks data made available by humaniki would help her identify new research themes to understand diversity trends on Wikimedia Projects and study factors curbing the growth of marginalized content.

“Knowledge is contested actively in places today that claim to be democratic and accessible”



Concerns

The same patterns of power are just being replicated within the space that is supposed to be democratic and more equity focused.

Observations/Needs

Key point: I am interested in looking at the incremental change over time.

Key point: Ethnicity/race are socially constructed and they are constantly negotiated, including on-wiki.

Key point: I want access to an open dataset by indicators such as country, profession, disciplines, region, caste.

Goals

Use the data made available by humaniki for identifying key research areas and spaces for intervention.

Some research ideas included:

1. Compare wikipedia data with other platforms like social media to identify patterns/trends and generate research questions.
2. Create an anomaly detection data model to identify bias in deletion actions against underrepresented groups.
3. Identifying coverage of indigeneous peoples

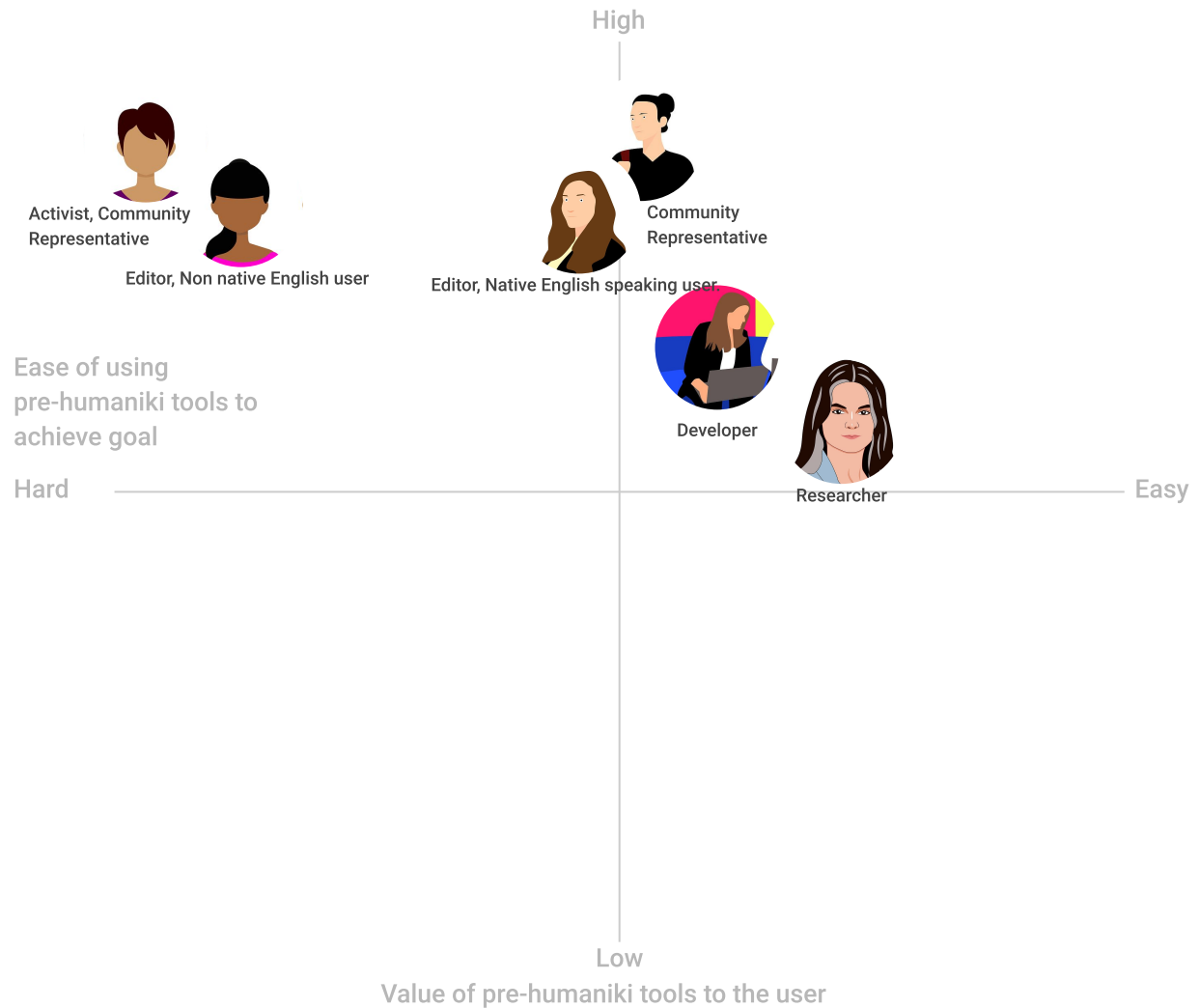
Opportunities

- Make data accessible to researchers in research friendly format.

Personas

We created a prioritization matrix to understand the ease of use in using the previous tools (WHGI, Denelezh) in comparison with the value those tools provide to them. Based on the fact that humaniki is a merge of those tools, we can say that humaniki would fall under the high value index for all our identified personas.

* These personas are mapped based on researcher's interpretations of their background and challenges.



Key Themes

We identified four key themes from the list of elicited feature requirements from the community.

- 1 Improving the usability of the data we already collect** by making it more shareable, more searchable and more re-usable.
- 2 Expanding the analysis dimensions of the existing data** including different attributes of humans, the snapshot date of the data was collected, and the interface language.
- 3 Providing actionable insights by highlighting editing opportunities** on wikidata and wikimedia projects.
- 4 Maintain highly used features and statistics** already in use by the community

1. Improving the usability of the data we already collect by making it more shareable, more searchable and more re-usable.

1a. Publication ready presentation

1b. Customizable and comparative visualizations

1c. Enabling third party applications via data API

1a. Publication ready presentation for creating awareness

Providing high level gender gap statistics for press release.

Description

Previous tools achieved their best success at producing high-level statistics that people found useful to cite (BBC, NYT, Bloomberg), users requested to make making these statistics pretty and easy to copy/cite.

Example: Provide high level gender gap stats for press release



We need high level metrics to create gender gap awareness and to initiate the discussion in the community. When a women meetup is organized, the community members reject the idea and say 'Why are we putting so much money for women's party?'

– P25, [Indian Language Wikipedia Editor](#)

Use cases

As a wikiproject representative,

1. I want information about high level statistics like latest percentage of women biographies to keep the women in red wikiproject page stats updated.
2. I want to be able to easily view high level metrics to be able to say something at a press release, stats that can be turned into a narrative.

`</>` Shortlisted Feature for Development

1b. Customizable and comparative visualizations to enable data exploration

Provide custom selectors for filtering most relevant information based on user needs

Description

In the current version, we allow users to view the static visualizations but without dynamic exploration of the data. Customizable visualizations would enable interactive exploration of gender gap on different wikiprojects.

Example: Provide custom selectors for filtering most relevant information based on user needs



We need Nicer tool and more interactive interface then it could be used to do so many things to find what is missing

– P18, Diversity focused Editor

Use cases

As a wikiproject representative,

1. I want to filter out information based on Wikiproject and Language.
2. I want to filter out most significant geographic list with real countries.
3. I want to be able to compare two groups of countries and be able to aggregate a custom country list.
4. I want to be able to compare high level occupation trends between different countries, by grouping subsets of occupation categories like football players, dancers as sports people
5. I want to be able to aggregate some language wikipedias to compare situation only some language wikipedians, e.g. african language wikipedians

As a diversity focused editor,

1. I want to be able to filter out languages that I am most interested in, languages I am interested in are really big in terms of global usage.

</> Shortlisted Feature for Development

1c. Enabling third party applications via data API

Deliver a rich data pipeline that can support integration with similar tools and research in the domain

Description

With the humaniki's rich data pipeline, further applications could be made that are out of our project scope.

Example: Anomaly detection to alert community members about the large scale deletions targeting a particular gender, language, or occupation, or event (date range).



I want access to an open dataset by indicators such as country, profession, disciplines, region, caste to explore research opportunities in the diversity domain.

– P8, Researcher

Use cases

As a researcher,

1. I want to export data about article deletions for further analysis.

As a community representative,

1. I want to share precise link to customized visualization for others to duplicate my pattern of data exploration.

As a developer:

1. I want to identify good seed articles, like ukrainin scientists to feed into GapFinder, to improve its search accuracy.

2. Expanding the dimensions of the existing data

Including different attributes of humans, the snapshot date the data was collected, and the interface language.

2a. Gender gap evolution

2b. Maximizing currently collected gender gap data

2c. Internationalization

2a. Gender gap evolution to track diversity trends

Enabling tracking of content diversity over time.

Description

While current gender gap statistics are useful, users requested to be able to view how they have changed over time to get insights about community efforts.

Example: Provide Monthly, weekly and yearly tracking of article creation and article deletions



I want to track deletions to make a strong argument of how knowledge is being suppressed.

– P22, [Researcher](#)

`</>` Shortlisted Feature for Development

Use cases

As a diversity focused editor,

1. I want to track article deletions (24hrs - week, to 3 months after creation) mapped by ethnicity to try and keep articles of underrepresented groups from being deleted.
2. I want to view how much female related information is available with at least one sitelink (any wiki project/language edition), So that we can work on the percentage that is not represented.
3. I want to view differences in women scientists articles (particular occupation field) being created over the course of a year, upto today.
4. I want to view evolution of gendered content representation on Wikidata.

As a researcher,

1. I want to see an all up view of additions and deletions over time, mapped with editathons and events.

As wiki project representative,

1. I want to measure the evolution from one point to next time, and be able to look data from specific years.

2b. Maximizing currently collected gender gap data

Providing statistics about other categories like readership mapped with gender.

Description

Previously this project has just been concerned with the basic dimensions of humans in Wikidata, but users requested to be able to view other categories mapped with gender.

Example: Occupations are described in Wikidata, but their subclass-trees are messy, we can aggregate by greatest-common-denominator occupations.



I want to choose specific women based on high level occupation fields like Women scientists that should automatically encompass researchers and other subcategories so that I don't have to write Sparql queries. Also, health care professionals (optometrist, audiologists, nurses, doctors) and STEM category.

– P18, Diversity focused Editor

Use cases

As a diversity focused editor,

1. I want to identify articles on women that have fewer sitelinks.
2. I want to get suggestions for relevant links that can be added between different articles about women.

2c. Internationalization to adapt software to different languages

Make tools available in non-English language interfaces.

Description

This tool has only been available in English, yet the gender gap exists across different language projects and users from non-English language editions requested for tool to be made available in their language.



Sometimes we feel that tools with English Interface only track the English written information on Wikidata, it's a knowledge gap!

– P15, Diversity focused Editor from Indian Language Wikipedias

Use cases

As a **wiki**project representative ,

1. I want to create awareness about gender gap on local language Wikimedia communities to encourage participation in bridging the gap and develop empathy amongst the dominating administrators for such activities.
2. I am identifying resources and tools that can help inform or support the diversity movement.

As a **diversity** focused editor,

1. I want to create gender gap awareness to motivate participation at grassroots levels.

`</>` Shortlisted Feature for Development

3. Providing actionable insights by highlighting editing opportunities on wikidata and wikimedia projects.

3a. Customized List-making

3b. Data Completeness

3a. Customized List Making

Providing advanced search features not only to view diversity trends but also to extract a list of relevant articles for improvement.

Description

List making already occurs for many wikiprojects with listerobot and other tools. But there are still some advanced search criterias that can't be made that way, and users requested to expand WHGI and Denezh's advance filtering features to enable list making.

Examples:

1. Finding relevant articles by language edition, country, and occupation.



“It is hard to do a good query search which will rule out the objects from Wikidata that don't meet relevance criteria on Wikipedia. This ends up with a list where quite a few redlinks are actually relevant.”

– P4, [Wikiproject representative](#)

Use cases

As a developer,

1. [Integration Opportunity] I want to identify good seed articles, like Ukrainian scientists to feed into WikiGapFinder, to improve its search accuracy.

As wikiproject representative,

1. I want to identify article overlaps between languages.

As diversity focused editor,

1. I want to identify women in wikidata with no sitelink, so that we can create articles for them.
2. I want to be able to see how many sitelinks a given Wikidata item has because it allows them to have confidence about notability checks
3. I want the tool to give me all women who don't have an article on my language Wikipedia who have an occupation that is a subclass of artists that have articles on at least one wikipedia, and want to be able to sort based on no. of sitelinks.

3b. Data Completeness to improve structured data source

Highlight information gaps on our data source - Wikidata to create high impact improvement opportunities for knowledge gaps on Wikidata.

Description

While we do cite the source of the data for humaniki, users requested to be able to view the limitations of wikidata and highlight editing opportunity. This can help direct user traffic to improve gaps on wikidata.

Example: Highlight proportions of articles that don't have gender.



“Currently I have to manually calculate the wikidata item count without gender”

– P16, Diversity focused Editor

Use cases

As diversity focused editor,

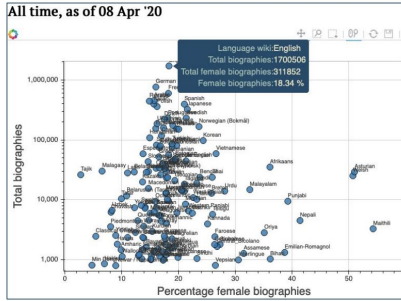
1. I want to identify women in wikidata with no sitelink, so that we can create articles for them.
2. I want to view evolution of gendered content representation on Wikidata.
3. I want to view proportion of articles that don't have gender [wikidata humans without gender].
4. I want to view statistics about sitelinks for gendered content on wikidata to get high level of opportunities for improvement.

</> Shortlisted Feature for Development

4. Retain highly used features and statistics already in use by the community



- New articles created:
Jul 18, 2015 - Mar 31, 2020:
127,685
- Percentage of biographies about women on English Wikipedia:
 - Oct 2014: **15.53%**
 - Apr 2020: **18.34%**



Women in Red (<https://w.wiki/347>) (<http://whgi.wmflabs.org/gender-by-language.html>)

Women in Red using Women Representation by Language Plot (WHGI)

Retain highly used features

Polishing existing tools to retain highly used features and statistics by the community

Description

While our research goal was to elicitate new features requirements based on community needs in the diversity space, we also identified the high usage level of some WHGI and Denelezh features that the community wants to retain.

Some visualizations/data that provide high value to the community: gender by language, filter by project family (pedia, quote, source), timeline view (by date of birth), denelezh advance view helps in understanding evolution with the horizontal bar graphs, gender by country.



I will use the language scatter plot visualization as long as it is made available by the tool. It helps start conversations between members from different language editions.

`</>` Shortlisted Feature for Development

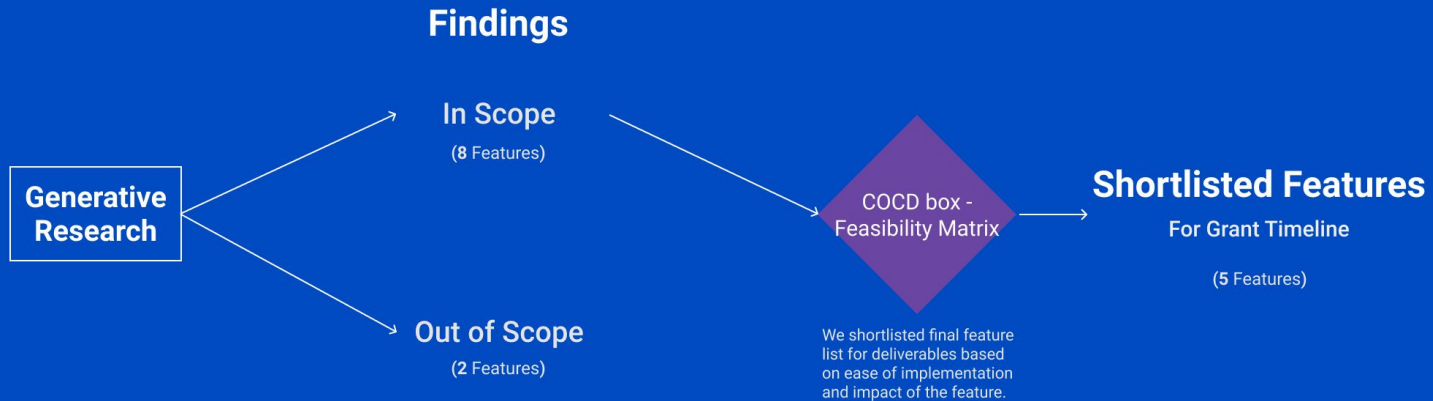
Occupation a	Total	Total with gender -	Females -	% Fem. -	Gap	Males -	% Mal. -	Others -	% Oth. -
agent	3,311,455	3,209,217	573,452	17.86%		2,634,587	82.09%	1,178	0.037%
worker	1,642,656	1,577,779	215,334	13.64%		1,361,943	86.32%	502	0.032%
creator	1,568,603	1,520,948	365,985	24.06%		1,154,148	75.88%	815	0.054%
professional	1,119,574	1,073,605	150,530	14.02%		922,753	85.94%	322	0.030%
sports figure	807,432	788,712	117,139	14.66%		681,493	85.32%	80	0.010%
artist	795,516	778,898	219,895	28.31%		556,321	71.61%	540	0.070%
athlete	784,450	776,549	115,705	14.80%		660,765	85.06%	78	0.010%
author	612,323	595,067	120,367	20.22%		474,300	79.72%	310	0.050%
competitive player	536,164	530,903	52,640	9.91%		478,234	90.07%	29	0.005%
position	515,678	500,024	76,387	15.27%		423,523	84.70%	114	0.023%
politician	513,578	483,366	56,182	11.62%		427,072	88.35%	112	0.023%
entirety	507,892	491,380	78,387	15.85%		412,865	84.02%	113	0.023%
actor	447,592	434,423	93,445	21.51%		340,702	78.43%	253	0.058%
researcher	444,189	429,036	65,941	15.29%		363,604	84.74%	91	0.021%
scientist	294,872	283,041	41,699	14.73%		241,280	85.24%	62	0.022%
musician	256,387	250,804	61,144	24.37%		189,496	75.55%	164	0.065%
association football player	251,162	248,117	11,388	4.57%		237,725	95.42%	4	0.002%
actor	245,492	241,279	108,133	44.81%		132,879	58.07%	267	0.111%

Women's biographies by occupation (<https://www.denelezh.org/>)

Women in Red using Women's biography by occupation plot (Denelezh)

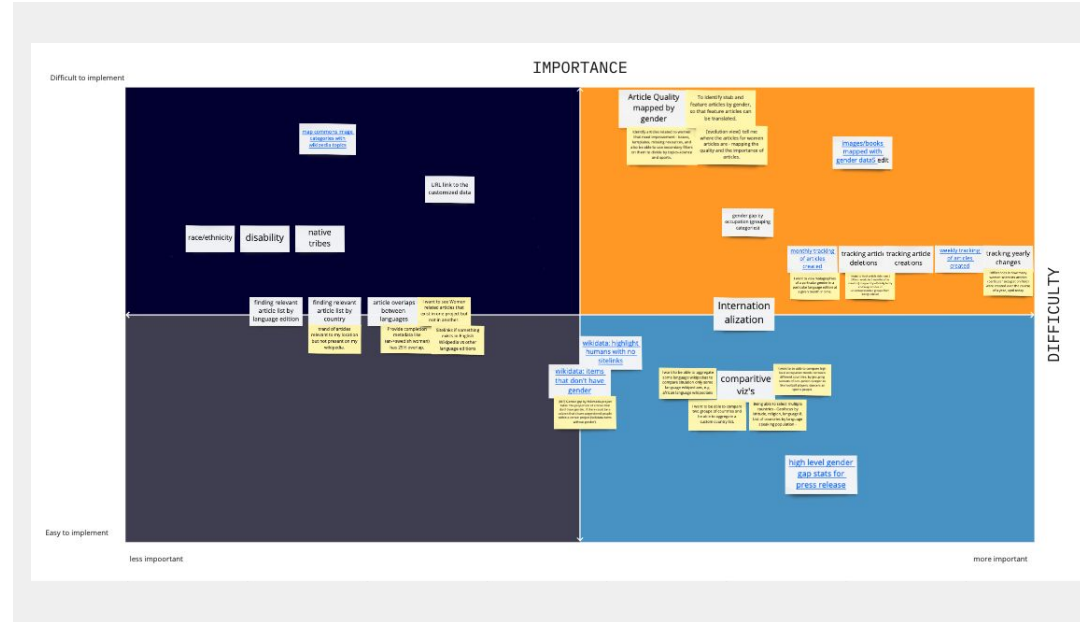
– P12, Women in Red Representative

Implications for Design - Next Steps



Feasibility and Deliverables

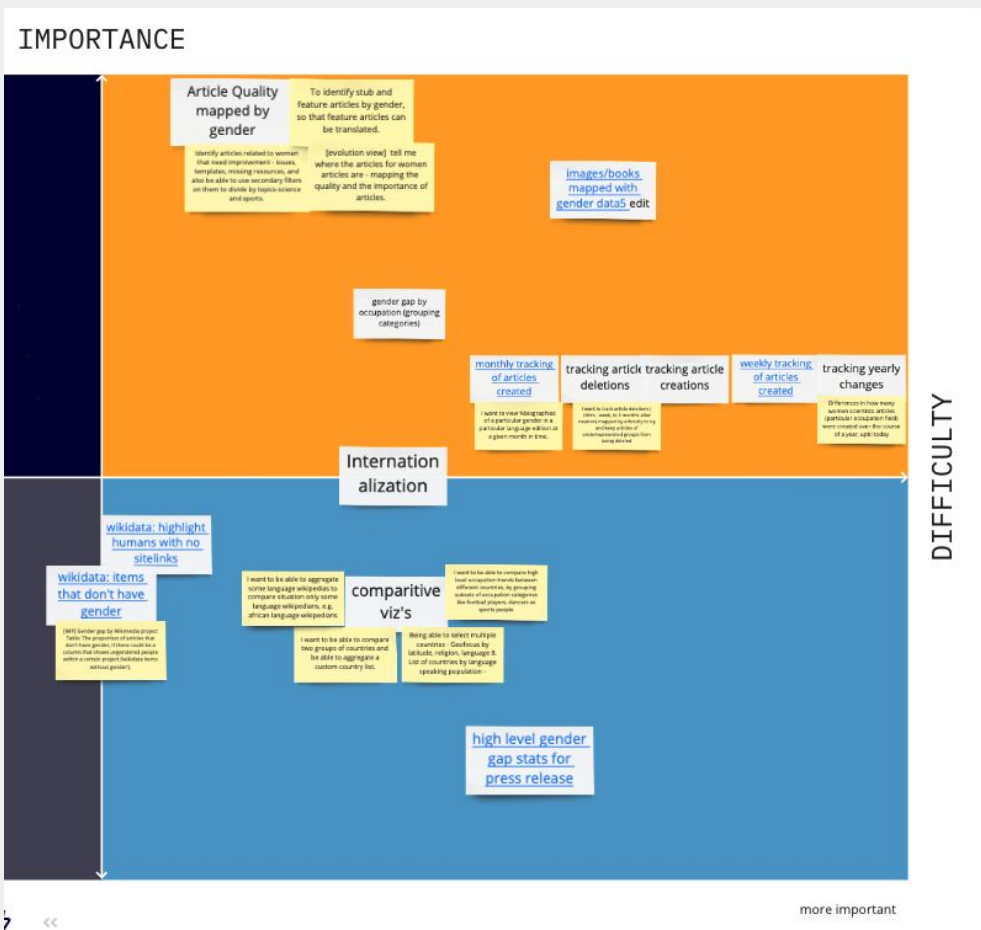
We converged our ideas into a list of deliverable features by using a tweaked version of [Centre for Development of Creative Thinking's matrix \(COCD box\)](#), which allowed us to select final feature list based on ease of implementation and impact of the feature.



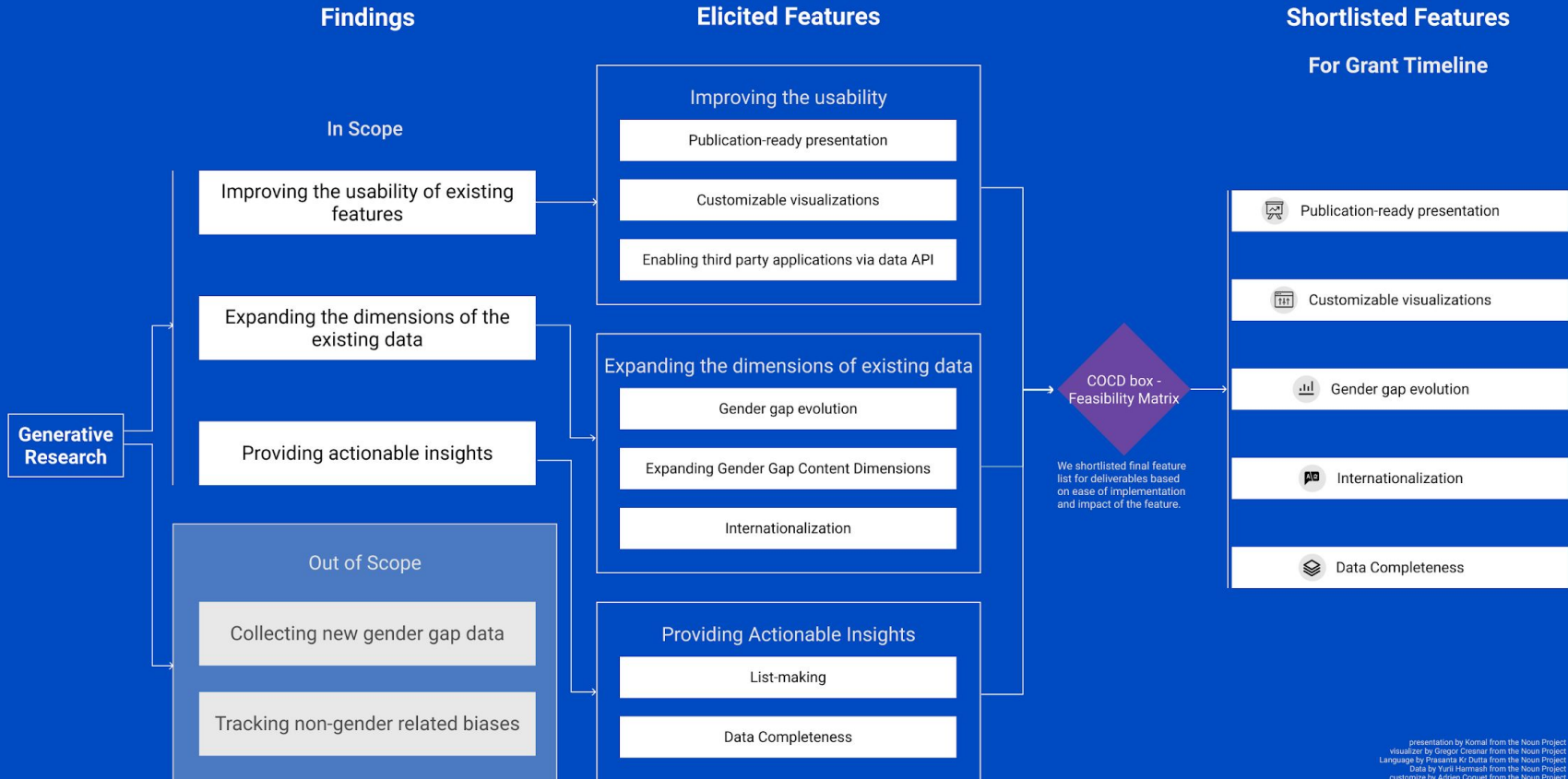
Final Feature List

The features mapped in the lower right quadrant are easy to implement and of high importance to our identified personas. The level of difficulty increases as we go from lower to upper quadrant. We shortlisted high feasibility, high impact features for our final set of deliverables, they are listed below:

- 1a. Publication ready presentation for creating awareness
- 1b. Customizable and comparative visualizations to enable data exploration
- 2a. Gender gap evolution to track diversity trends
- 2c. Internationalization to adapt software to different languages
- 3b. Data Completeness to improve structured data source



humaniki Project Process Flow



Design

Designing the elicited features and creating a working prototype

We plan to implement designs for the above identified 5 features and deliver mockups and prototypes that can further be tested with the community. We plan to employ co-designing exercises using the [Five Design Sheets](#) method with the engineers, scientists and designers in the team to create effective visualizations to best address the identified use cases.

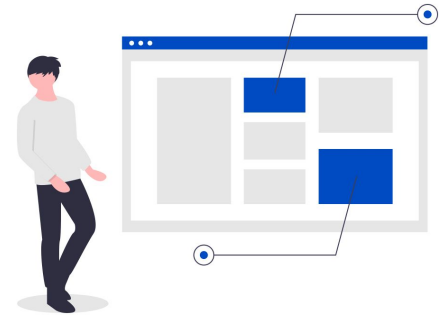


Illustration from undraw.co

Alpha and Beta Testing

Alpha testing

Our development team is prioritizing to release alpha features, a functioning application with merger of denezh and WHGI. We plan to test the functioning website with a technical team to get feedback on architecture and technical design.

Beta testing

We then plan to direct our efforts towards implementing a beta version, delivering the tool with the the final feature list identified from our generative user research.



Illustration from undraw.co

Project / Appendix

Appendix

Out of Scope Opportunities

Important identified themes that we are not going to pursue

1 Collecting other attributes of gender gap data

2 Tracking non-gender related bias

User:WP 1.0 bot/Tables/Project/Women writers

From Wikipedia, the free encyclopedia
< User:WP 1.0 bot | Tables | Project

Women writers articles by quality and importance							
Quality	Importance						Total
	Top	High	Mid	Low	NA	???	
★ FA	8	16	47	36			107
★ FL		1	9	3			13
ⓘ A			2				2
⊕ GA	8	24	91	178			301
B	48	141	465	859		1	1,514
C	36	304	3,196	5,817	4	699	10,056
Start	3	110	4,320	10,834	22	5,111	20,400
Stub		40	2,913	8,726	62	1,247	12,988
List	2	2	130	107	1		242
Category					2,020		2,020
Disambig					1		1
File					65		65
Project					17		17

Articles by quality and Importance on Wikipedia

Collecting other attributes of gender gap data

Description

We identified, community members found statistics of gender mapped with different categories very useful and of high impact. But they also requested to enable below listed categories to be mapped with gender to identify biases in various other dimensions. We do not have the data in place to execute this.

1. Statistics and trends about article quality and importance mapped by gender.
2. Show editor gender gap
3. Track readership of male and female articles (using the new pageview API).
1. Semantic analysis on text to see if it is a biased text or not (identify context), check if women perspective is missing
2. Identify images/books mapped with gender data
3. Provide a tool to update gender data available on infobox to wikidata automatically? (possibility: via lua)
4. Map commons image categories with relevant wikipedia topics of women to improve visual presentation of the content.

Tracking non-gender related biases

Description

Although we identified the community's request to track non-gender related biases, the tracking of which would enable understanding the content coverage of other underrepresented groups, we do not have the data in place to execute this.

We identify the need for tracking/uncovering following non-gender related biases:

1. Race/Ethnicity
2. Disability
3. Native Americans

Research Ideas (suggested by participants)

1. How reliable sources, notability affect marginalised communities on Wikipedia?
2. How different things shape deletions actions - gender, sexuality, global north|global south, race, ethnicity.
 - a. Why and how the content was deleted?
 - b. Biographies created during publicized events could get deleted!
 - c. When thinking about knowledge gaps, really attending to intersectionality and thinking it not only as what's missing but also what's being actively attacked i.e mass deletions.
3. Compare English and Malayalam, Swedish and Malayalam, what proportion of doctors in Malayalam language Wikipedia vs in English Wikipedia. And see if one community has better representation of health information vs others.
4. Semantic analysis of gendered content: Are articles on women about children and their relationships whereas the articles on men are about career achievements?
5. Is there a gender skew in Bureaucrats on English Language Wikipedia? When Bureaucrats were first created in 2003, administrators and bureaucrats had gender skews.
6. Another editor survey is needed with attributes: Age skew, gender skew, ethnicity skew, Smartphone usage, Age representation on Wikipedia (Interesting to know if over 50's age group is overrepresented), when did editor first have access to internet, leisure time to use internet, Desktop vs laptop.