Wikidata Gender Diversity (WiGeDi)

Wikimedia Research Fund Proposal

Title

Title of proposal.

Wikidata Gender Diversity (WiGeDi)

Dates

Provide the anticipated start and end dates of the project.

01.06.2022 – 31.05.2023 (12 months)

Username

Applicant's Wikimedia username. If one is not provided, then the applicant's name will be provided on Meta with the proposal for community review.

Mushroom (username of the first applicant)

Description

The Stage II Description should provide a clear statement of the work to be undertaken and must include the objectives for the period of the proposed work and expected significance; the relationship of this work to the present state of knowledge in the field. It should incorporate, build on, and replace the material in the Description, Impact, and Dissemination sections of Stage 1. The Project Description should outline the general plan of work, including the broad design of activities to be undertaken, and, where appropriate, provide a clear description of methods and procedures. Proposers should address what they want to do, why they want to do it, how they plan to do it, how they will know if they succeed, and what benefits could accrue to Wikimedia and to others if the project is successful. The project activities may be based on previously established and/or innovative methods and approaches, but in either case must be well justified. These issues apply to both the technical aspects of the proposal and the way in which the project may make broader contributions. This description should provide a discussion of the potential impacts of the proposed activities. Broader impacts may be accomplished through the research itself, through the activities that are directly related to specific research projects, or through activities that are supported by, but are

complementary to the project. They may be direct benefits to Wikimedia projects and/or the Wikimedia movement or indirect. As in Stage I, we aim to prioritize proposals that aim to enable the Wikimedia communities in making decisions or taking actions of significant impact as a result of the research conducted. We will give special consideration to proposals that directly address the Wikimedia 2030 Strategic Direction (including but not limited to the Movement Recommendations) as well as proposals that attempt to answer research questions in less commonly studied languages of Wikipedia. We will prioritize proposals that will impact Wikimedia user groups, affiliates, and developer communities. The Community Wishlists from 2019, 2020, and 2021 for ideas of problems raised by these groups.

The Wikidata Gender Diversity (WiGeDi) project aims to investigate the issue of gender diversity in the Wikidata knowledge base, focusing in particular on the marginalized identities of trans, non-binary, and other gender-diverse people.

All previous studies about this subject in Wikimedia projects have focused mainly on the *gender gap*, defined as the gap in the representation of women versus that of men. Some of these studies (e.g. the ones by Konieczny and Klein) have acknowledged the existence of trans and non-binary people, but no research has looked specifically at how marginalized gender identities are represented, or how fair and inclusive the current representation is.

The Wikidata Gender Diversity project aims to adopt a different perspective, which centers marginalized gender identities. Since every edit and every user discussion throughout the history of Wikidata is archived in the project itself and made publicly accessible, our project will allow a unique and comprehensive overview of *how* marginalized gender identities have been represented in Wikidata, *what* exactly has been represented, and *why* the users have made certain modeling and implementation choices.

Our preliminary study about this topic (*Metilli D. & Paolini C., Non-binary gender representation in Wikidata*, to be published in *Ethics in Linked Data*, Litwin Books, 2022; draft available here: https://docs.google.com/document/d/1It-iyomHPlKT4Zwh5a6dyepCnUuOAAdGRye-3zQ877U/edit?usp=sharing) shows that gender modeling in Wikidata has a very complex history, from which important lessons can be learned about how the representation of marginalised gender identities have been handled by the user community, and which steps remain to be taken to make Wikidata a fully inclusive project with regard to gender.

WiGeDi will perform a broad analysis of gender diversity in Wikidata, from three different

research perspectives: *modeling*, *data*, and *community* (see section “Research Questions”). The project will also develop a web application to present the collected data in a user-friendly way (see section “Development”). Finally, the project will benefit Wikidata itself and other Wikimedia projects by identifying issues with the current and previous modeling of gender and suggesting improvements (see section “Impact”).

Research Questions

The WiGeDi project aims to center marginalized gender identities by performing a broad analysis of gender diversity in Wikidata from three different — and complementary — perspectives: *modeling*, *data*, and *community*.

Modeling

The *modeling* question will look at the Wikidata ontology of gender, including the classes and properties that make up the current model, their taxonomies, and any relevant contextual data.
We also plan to analyse the evolution of the model over time to understand if, and how, the model has evolved over time to support a more inclusive representation of gender.

In particular, we aim to answer the following research questions:

1. How does the current Wikidata ontology model represent gender?
2. Does the Wikidata ontology model represent gender in a fair and inclusive way?
3. How has the model evolved to account for gender identities that fall outside the traditional binary view of gender?

Data

The *data* question will look at the actual biographical data that is represented in the knowledge base, and compute statistics about gender representation in the knowledge base. These data will then be analyzed to understand how fair and inclusive the Wikidata gender model is, and how well it represents society.

In particular, we aim to answer the following research questions:

4. What kinds of marginalized gender identities, such as trans and non-binary identities, are described in Wikidata, and how?

5. How are these gender identities distributed through time, space, and other metrics, and how do they compare to the whole biographical dataset?
6. Is the current gender data contained in Wikidata representative of society as a whole?

Community

The *community* question will look at how the Wikidata community has developed the current ontology of gender, and through which processes the model has evolved over time. We plan to look especially at user discussions about the topic throughout the years, adopting computational linguistics techniques to perform quantitative analyses on the data.

In particular, we aim to answer the following research questions:

7. How did the current Wikidata model of gender come to be adopted?
8. What issues related to gender modeling did the community discuss over the years, and how have these discussions evolved over time?
9. How has the multilingual nature of Wikidata affected the user discussions?

We believe that only by answering all the research questions reported above it will be possible to obtain a comprehensive overview of gender diversity in Wikidata.

Related Works

To date, there have not been many studies about gender in Wikidata. The first scholars to approach the subject were Klein et al. (2016) and Konieczny and Klein (2018), who carried out an in-depth analysis of the Wikidata gender gap. The authors applied several gender gap indexes to the data contained in the knowledge base, showing that women are under-represented compared to men, and in general that Wikidata appears to be affected by the same gender disparities that exist in society at large. The most recent project by the authors is Humaniki,[1]a tool showing the gender gap among all Wikimedia projects.

Hollink, Van Aggelen, and Van Ossenbruggen (2018) measured gender differences in a subset of Wikidata entries. Zhang and Terveen (2021) recently conducted a case study about the Wikidata gender content gap. The Wikidata Community Survey 2021 has looked at gender metrics in the community of Wikidata editors, showing that the Wikidata community is overwhelmingly male (75%), while female users make up just 16%, and non-binary users 2.9%. The remaining users (6%) opted not to answer the question.

There have been many studies about the gender gap in Wikipedia, which is a sister project to

Wikidata. Antin et al. (2011) were the first to study gender differences in Wikipedia editing, while Reagle and Rhue (2011) looked at gender bias in the content of the encyclopedia. Wagner et al. (2015) analyzed how men and women are portrayed in Wikipedia. Johnson et al. (2020) looked at gender differences among the readers of the encyclopedia. Field, Park, and Tsvetkov (2020) analyzed social biases in Wikipedia biographies, while Tripodi (2021) investigated the frequent deletion of biographies about women.

More recently, Redi et al. (2020) have created a taxonomy of knowledge gaps found in Wikimedia projects. Miquel-Ribé & Laniado (2021) have developed the Wikipedia Diversity Observatory, a project that tracks the content gaps that are present in Wikipedia, making it easier to remedy them. Miquel-Ribé, Kaltenbrunner & Keefer (2021) have looked specifically at LGBT+ content, comparing gaps among different language editions of Wikipedia.

While some of the previous studies about gender in Wikimedia projects acknowledged the existence of marginalized gender identities, they did not investigate specifically how their identities are represented, or how this representation has evolved over time. Our project will differ from the previous ones because it will be the first to center trans, non-binary, and other

[1] https://humaniki.wmcloud.org

gender diverse identities. Furthermore, it will adopt a holistic view of the topic that does not merely focus on statistical data, but also looks at modeling, community processes, and contextual events, to build a comprehensive overview of gender diversity in Wikidata.

Project Schedule

The project will be composed of five work packages:

1. Data Collection

This work package will feature the collection of three kinds of data from the Wikidata knowledge base: about the *ontology model*, about *biographical data* about people who are represented in the knowledge base, and *user discussion* data.[2] These last two tasks will depend on the completion of an ethics statement, since they may involve personal or sensitive data (WP5). A further task of *data cleaning and integration* will complete the work package.

2. Research

This work package will contain the three main research tasks of the project, which are related to the *modeling*, *data*, and *community* research questions. The *modeling* question will be investigated from M2 to M9. The *data* question will be investigated from M3 to M10. Finally, the *community* question will be investigated from M4 to M11. Each of these three broad research questions will lead to at least one publication. See section "Research Deliverables" below for a description of the anticipated outputs.

3. Development

This work package will consist of three tasks, each dedicated to the implementation of three development outcomes: the Wikidata Gender Timeline, to be previewed at M6 and completed at M10, the Wikidata Gender Dashboard, to be previewed at M9 and completed at M12, and the Wikidata Gender Talk corpus, to be released as a dataset at M6 and then as a browsable online repository at M10. See section "Development Deliverables" below for a description of the anticipated outputs.

[2] Some data was collected for a preliminary study before the start of the project, but the focus of WiGeDi is much broader. Furthermore, we need to make sure that our data collection practices comply with the ethics requirements that will be agreed with UCL.

4. Ethics

This work package will investigate ethical questions surrounding our data collection, research, and development practices, and in particular the re-publication of user-generated content (such as user discussions) in our web application. Subsequently, at M8 we will complete an ethics review to ensure that we are compliant with any applicable laws and regulations, and that we fulfill any ethical requirements.

5. Community Engagement

This work package will be dedicated to engaging with the Wikidata community itself. We plan to inform the community of our progress throughout the whole duration of the project, and produce a detailed report at M6 (to be published on Wikidata itself), and then a final report at M12. In addition, we also plan to organize a final seminar event which will feature both scholars and community members (M12).

Gantt Chart

The following Gantt chart illustrates the main activities to be carried out in the project, including their duration in months and the related deliverables.

WP #	TASK	M1 06/22	M2 07/22	M3 08/22	M4 09/22	M5 10/22	M6 11/22				
1	Data Collection										
1.1	Ontology Model				D1.1						
1.2	Biographical Data				D1.2						
1.3	User Discussions				D1.3						
1.4	Data Cleaning & Integration						D1.4				
2	Research										
2.1	Model							D2.1			D2.4
2.2	Biographical Data							D2.2			
2.3	User Discussions								D2.3		
3	Development										
3.1	Wikidata Gender Timeline							D3.2			
3.2	Wikidata Gender Dashboard							D3.3			D3.5
3.3	Wikidata Gender Talk (WiGeTa) Corpus						D3.1		D3.4		
4	Ethics										
4.1	Ethics Statement		D4.1								
4.2	Ethics Review										D4.2
5	Community Engagement										
5.1	Wikidata Engagement						D5.1				D5.2

5.2	Seminar Organization									D5.3

Project Outcomes

The project will result in several research and development outcomes. In the following, we describe each of them by referencing the above Gantt chart.

Data Collection Deliverables

The data collection work package will produce the following deliverables: D1.1, a dataset related to the Wikidata gender model; D1.2, a dataset of biographical data about people, and in particular gender data, contained in Wikidata; D1.3, a dataset of gender-related user discussions.[3] These datasets will go through a data cleaning and integration phase, resulting in D1.4, the WiGeDi knowledge base that subsequent tasks will be based on.

Research Deliverables

Our project aims to answer all research questions reported above, and produce at least one publication for each. In particular, we are considering the following publication venues:

- For the *modeling* research questions, we are at least one publication in an open access journal such as the Semantic Web Journal (IOS Press), or the Journal of Web Semantics (Elsevier), about the Wikidata gender model and its evolution over time (D2.1). This research will be led by Daniele Metilli, who has a background in knowledge representation.
- For the *data* research questions, we anticipate participation in at least one relevant conference or workshop (D2.2), such as the Wiki Workshop. This research will be co-led by Marta Fioravanti and Beatrice Melis, and will be tightly integrated with the work on the Wikidata Gender Dashboard (see section “Development Deliverables”).
- For the *community* research questions, we intend to publish our findings on user discussions in a computational linguistics journal (D2.3). This research will be led by Chiara Paolini, who has a background in computational linguistics.

[3] These datasets are considered internal project deliverables. The dataset of user discussion will be published as D3.1.

- Finally, at the end of the project, we aim to publish a journal article (D2.4) presenting the final results of the three research tasks from a broad perspective, discussing our findings, and investigating future avenues of research.

UCL's Department of Information Studies will support us in our research activities, in particular by providing review and advice from experts on the subjects of data modeling, linguistics, and gender studies, which are affiliated to the Department. Moreover, we will be able to leverage the existing APC agreements between UCL and open access publishers, thereby greatly reducing publication costs incurred by the project (see section "Budget").

Development Deliverables

In addition to research dissemination, we also plan to develop a web application through the work of two freelance digital humanists.[4] One will assume the role of *Creative Technologist*, while the other will assume the role of *Data and Narrative Scientist*. The resulting web application (hosted at https://wigedi.com) will contain:

- A *Wikidata Gender Dashboard* about the current status of gender diversity in Wikidata, centered on marginalized gender identities and featuring notable trans and non-binary people, including new additions to the knowledge base. The dashboard will be released in two stages. A preliminary version (D3.3) will be published at M8, while the final version (D3.4) will be published at M12. The dashboard will also feature statistics about Wikimedia projects connected to Wikidata, and in particular the different Wikipedia language editions. Most of the work on the dashboard will be carried out by the *Creative Technologist*, with the support of the *Data and Narrative Scientist*.

- An annotated *Wikidata Gender Timeline* (D3.2), featuring a detailed report of how gender has been modeled since the beginning of the project, contextualized with information about real-world events that impacted changes in the model, and tightly integrated with the *Wikidata Gender Talk* repository (see below). The timeline will be developed mainly by

the *Data and Narrative Scientist*, who will be tasked with building a coherent and cohesive view over the gender-related events since the beginning of Wikidata. The *Creative Technologist* with support the design of the Timeline.

[4] Unfortunately, due to institutional requirements, we are not able to select and name the two freelance digital humanists before the beginning of the project.

- A browsable *Wikidata Gender Talk* repository of gender-related Wikidata user discussions, interconnected with the timeline (D3.3); initially published as a corpus to enable further study (D3.1). These two development outcomes will enable the investigation of the *community* research questions, leading to the publication of the related research deliverables. The repository will be developed jointly by the *Creative Technologist* and *Data and Narrative Scientist*.

The final outcomes of the project will provide a complete overview of the topic of gender representation in Wikidata, enabling further studies and comparisons with other knowledge bases and Wikimedia projects, including Wikipedia. Furthermore, the development deliverables will be tightly integrated with the research deliverables. For example, the timeline and the repository of user discussions will be referenced in subsequent research publications to allow the reader to access the primary sources that the research is based on.

Ethics Deliverables

The ethics deliverables will be the initial ethics statement related to data collection practices (D4.1), and an ethics review at the end of the project (D4.2) that will ensure the fulfillment of the requirements identified in the ethics statement.

Community Engagement Deliverables

The community engagement deliverables will be a mid-project detailed report (D5.1), a final report (D5.2), and the outcomes of the community seminar (D5.3). We also plan to update the community on the project's progress at least once a month. In addition, we will participate in relevant community-led events, such as the Wiki Workshop and WikidataCon.[5] Finally, we plan to establish ties with existing WikiProjects related to gender diversity (see section "Impact" below).

[5] We have opted not to consider participation in community events as separate project deliverables, however we strongly believe that participating in relevant community-led events will be crucial to make the project successful.

Sustainability

To ensure the project's sustainability, we plan to design the web application such that it is automatically updated based on changes in Wikidata content, and that it requires minimal maintenance. Server hosting costs have been adjusted to account for at least 4 years of self-hosted operation. However, we will also investigate the possibility of relying on Wikimedia Cloud Services (WMCS) after the end of the project, to ensure a successful operation of the web application for years to come.

After the end of the project and end of the freelance work contracts, the team intends to continue maintenance of the web application with the help of volunteers from the Wikidata community. Through community engagement and discussions with Wikimedia, we will investigate the best way to accomplish this goal. In any case, the four applicants and the institution backing the project (UCL) are fully committed to ensuring the sustainability of the project, and will consider this a crucial project requirement.

Impact

The WiGeDi project will have a significant impact both on Wikidata and on other Wikimedia projects. It will be the first research project to investigate gender diversity in Wikidata by centering marginalized gender identities.

The project will address the 2030 Wikimedia Strategic Direction by fostering safety and inclusion of marginalized gender identities (recommendation #3), improving user experience through the monitoring of project activities (recommendation #10) and the identification of potential improvements in gender modeling (recommendation #2), ensuring equity by analyzing past instances of non-inclusive decision processes (recommendation #5), and our research will

also help identify further topics of impact (recommendation #8).

We plan to coordinate with existing WikiProjects dedicated to gender issues, such as the Wikidata Project LGBT, Humaniki, Art+Feminism, and WikiWomen in Red. At the beginning of the project, we will establish ties with these communities and collect useful input and suggestions from them. We also plan to involve them in any project events.

Our research focuses mostly on the "Gender" content representation gap from the Wikimedia knowledge gap taxonomy, but it also involves other content gaps (Language, Structured data, Sexual orientation, Age/Recency, Cultural background) and on the related contributor representation gaps and reader representation gaps.

Another important impact is related to the insights that the project will produce with regard to the best practices for handling the representation of marginalized gender identities in knowledge bases in general. We believe that our study will produce useful feedback by analyzing what has been handled successfully by Wikidata over the years, where the community has stumbled to make progress, and what still remains to be done. Any collaborative project facing similar issues will potentially benefit from such feedback, including Wikipedia and other projects hosted by the Wikimedia Foundation.

Applicants' Backgrounds

Prior contributions to related academic and/or research projects and/or the Wikimedia and free culture communities. If you do not have prior experience, please explain your planned contributions.

Daniele Metilli is a research fellow in advanced architectures for Digital Humanities at University College London. They have previously contributed to Wikimedia projects as an administrator of the English Wikipedia and Wikidata (User:Mushroom), is a member of the Italian Wikimedia chapter (Wikimedia Italia), and was previously a Wikipedian in residence at a major Italian museum. Their research activity has often involved Wikimedia projects, in particular Wikidata (e.g. *A Wikidata-based tool for the creation of narratives*, Master's Thesis, University of Pisa, 2016; *A Wikidata-based tool for building and visualising narratives*, IJDL, 2019; *Populating narratives using Wikidata events: An initial experiment*, IRCDL, 2019), but also Wikipedia user discussions (*Talking Wikipedia: Mining the network of coordination interactions in Wikipedia discussion pages*, Bachelor's Thesis, Milan Polytechnic University, 2010).

Chiara Paolini is a PhD candidate in Linguistics, working in the Quantitative Lexicology and Variational Linguistics research group at KU Leuven. She holds a Master's Degree in Linguistics and Translation from the University of Pisa, and has previously won two Erasmus scholarships to work as a trainee at CNRS (Toulouse, France) and Utrecht Institute of Linguistics UiL-OTS (Utrecht, The Netherlands). Starting in 2020, Daniele Metilli and Chiara Paolini have begun researching gender diversity in Wikidata, participating in the WikidataCon 2021 with the presentations titled *Non-binary gender identities in Wikidata*. Their first publication about the subject is *Non-binary gender representation in Wikidata*, to be published in *Ethics in Linked Data*, Litwin Books, 2022.

Marta Fioravanti is a creative technologist at oio studio (London) and a Master's student in Digital Humanities at the University of Pisa. Her research interest involves the design of accessible and visual tools for analysis guidance. Her MA thesis, titled *Treemob: Expressive mobility data representation through tree-based structures*, is about finding expressive representations about the personal mobility of a user, embracing the individual perspective on movement data and presenting a new methodology to represent and study mobility through the aid of tree-shaped structures. With oio studio, she works for international clients like Space 10 and Ikea to imagine how the industry will change with the diffusion of AI, and to create interactive experiences explaining the creative potential of machine learning.

Beatrice Melis is a Master's student in Digital Humanities at the University of Pisa. Her study and research interests concern the use of computational methods and supports to preserve human cultural and emotional heritage — with a particular focus on the field of Italian poetry and literature. The topic of Beatrice's BA thesis *La primavera hitleriana di Eugenio Montale: saggio di critica letteraria con strumenti di storytelling digitale* (University of Pisa, 2020) was the construction of narratives using knowledge from Wikidata. The results of this study have been presented at the WikidataCon 2021, with the presentation titled *Wikidata-based narratives for research and education*.

Budget Stage II

Total amount in USD including any overhead, indirect costs, or administrative fees (up to a maximum of 15% of the direct costs).

USD 43,500

Entity Receiving Funds

Provide the name of the individual or organization that would receive the funds.

University College London (UCL)

Budget Description

Provide the link to your budget sheet. We recommend that you make a copy of the following template and share a link to your copy at the submission time.

The filled Wikimedia budget template is available here: https://docs.google.com/spreadsheets/d/1s-vWBeQNsTgpR7v-SwumVh6vSi6_eP3f/edit#gid=2082344268. Below, we provide a more detailed description of the budget expenses.

Budget Synthesis

Fees for freelance workers $24,000 (2 people per 20 hours per 40 weeks)
Conference and travel expenses $7,000 (attending 3 conferences for 2 people; travel costs for 2 project meetings) Seminar organisation $3,600 (final project event; speaker fees plus organization and promotion)
Computer equipment and software $1,000
Open access publishing costs $700 (for journals not covered by UCL open access agreement with publishers)
Server hosting $700 (including domain fees) Institutional overhead $6,500
Total $43,500

Fees for Freelance Workers

The project will pay two digital humanists freelancers to fulfill the following roles:

- The *Creative Technologist* will work on website development, graphic design, and data visualisations, handling in particular the presentation of data about gender diversity in the Wikidata Gender Dashboard, but also the graphics of the Wikidata Gender Timeline and of the Wikidata Gender Talk repository.
- The *Data and Narrative Scientist*, who will work on data collection and cleaning, and especially on the development of the interactive Wikidata Gender Timeline, including data, user discussions, and links to the relevant historical context, and also on the collection and data cleaning of the Wikidata Gender Talk corpus.

Each of the two digital humanists will be paid $12,000 as freelance work, based on a salary of $15 per hour per 20 hours per 40 weeks of work.[6]

Conference and Travel Expenses

We anticipate paying for participation in at least 3 conferences or workshops. Examples of relevant events include OpenSym, the International Semantic Web Conference, and the Queer in AI workshop. The project will cover conference registration, travel expenses, and accommodation for one to two people participating in each event. We also anticipate international travel and accommodation costs for two in-person meetings for the four applicants, to be held at M6 and M12.

Seminar Organisation

We will organise one event to present the final results of the project at M12, at which we will invite scholars, Wikidata users, and members of relevant communities. The event will be held virtually, but we intend to pay 4 invited speakers for participation. We anticipate an expense of $500 for each speaker (total $2,000), plus $1,600 for promotion of the event, production of videos, and other related expenses.

Computer Equipment and Software

At present, the only equipment expenses that we foresee are related to software licenses and computer accessories (e.g. hard disks). None of the co-applicants need to acquire computers or other expensive equipment. We anticipate the total expenses to be lower than $1000, but we

have set such amount to have room for contingencies in case of unforeseen circumstances.

[6] This is an estimate. The actual duration of the contracts may vary based on the project's needs. Compared to the previous draft of the budget, we have removed benefits expenses, as based on UCL guidance these are not applicable to freelance work in the UK.
Compared to the previous budget, we have significantly lowered equipment-related expenses due to changes in circumstances.[7]

Open Access Publishing Costs

We intend to publish at least three open access journal papers as outcomes of the project. Wherever possible, UCL will cover publication costs through existing APC agreements, but we have set aside an amount of $700 to account for one possible publication in a relevant journal that is not included in the agreements (such as *Semantic Web Journal*, IOS Press). For conference-related costs, see "Conference and travel expenses" above.

Server Hosting

Server hosting costs have been adjusted to account for at least 4 years of self-hosted operation (at a cost of 12€/month, plus domain registration). See section "Sustainability" above.

Institutional Overhead

Institutional overhead expenses have been increased to the allowed maximum ($6,500, or 15% of the total budget) to cover expenses incurred by UCL for managing the administrative and financial aspects of the project.

References

- Antin, Judd, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. "Gender Differences in Wikipedia Editing." In Proceedings of the 7th International Symposium on Wikis and Open Collaboration, 11–14.

- Field, Anjalie, Chan Young Park, and Yulia Tsvetkov. 2020. "Controlled Analyses of Social Biases in Wikipedia Bios." arXiv Preprint arXiv:2101.00078.

● Hollink, Laura, Astrid Van Aggelen, and Jacco Van Ossenbruggen. 2018. “Using the Web of Data to Study Gender Differences in Online Knowledge Sources: The Case of the European Parliament.” In Proceedings of the 10th ACM Conference on Web Science, 381–85.

[7] Our institutions (and UCL in particular) have provided us with equipment that we deem sufficient for the needs of the project, therefore there is no need to acquire additional hardware and software.

● Johnson, Isaac, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. 2020. “Global Gender Differences in Wikipedia Readership.” arXiv Preprint arXiv:2007.10403.

● Klein, Maximilian, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. 2016. “Monitoring the Gender Gap with Wikidata Human Gender Indicators.” In Proceedings of the 12th International Symposium on Open Collaboration, 1–9. ACM.

● 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.

● Miquel-Ribé, M., & Laniado, D. 2021. “The Wikipedia Diversity Observatory: Helping Communities to Bridge Content Gaps Through Interactive Interfaces”. Journal of Internet Services and Applications, 12(1), 1-25.

● Miquel-Ribé, M., Kaltenbrunner, A., & Keefer, J. M. 2021. “Bridging LGBT+ content Gaps Across Wikipedia Language Editions”. The International Journal of Information, Diversity, & Inclusion (IJIDI), 5(4), 90-131.

● Reagle, Joseph, and Lauren Rhue. 2011. “Gender Bias in Wikipedia and Britannica.” International Journal of Communication 5: 21.

● Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. 2020. “A Taxonomy of Knowledge Gaps for Wikimedia Projects” (second draft). arXiv preprint arXiv:2008.12314.

● Tripodi, Francesca. 2021. “Ms. Categorized: Gender, Notability, and Inequality on Wikipedia.” New Media & Society, 14614448211023772.

- Wagner, Claudia, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. “It’s a Man’s Wikipedia? Assessing Gender Inequality in an Online Encyclopedia.” In Ninth International AAAI Conference on Web and Social Media.

- Zhang, Charles Chuankai, and Loren Terveen. 2021. “Quantifying the Gap: A Case Study of Wikidata Gender Disparities.” In 17th International Symposium on Open Collaboration, 1–12.