Non-Binary Gender Representation in Wikidata

DANIELE METILLI & CHIARA PAOLINI

Introduction

In the era of big data, new ethical questions have arisen from the creation of large knowledge bases, whose data is produced, consumed, and shared by millions of users, both humans and machines. These knowledge bases often contain biographical information about people, including sensitive data such as gender, sex, ethnicity, or sexual orientation. Implicit biases in such data can generate unfairness[1] and lead to discriminatory applications that impact marginalized communities.[2]

This is particularly true for the trans and non-binary communities, who experience discrimination on the basis of gender identity. Digital projects have struggled to cope with the wider

1. Michael Veale and Reuben Binns, "Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data," *Big Data & Society* 4, no. 2 (2017): 1–17; Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 1–35.

2. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 2018 ACM Conference on Fairness, Accountability and Transparency*, ACM, 2018, 77–91; Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2018, 610–23.

societal acceptance of the fact that gender is not binary,[3] and in many cases they have perpetuated—or even amplified—the misgendering and erasure of trans and non-binary people that has occurred in society throughout history.[4]

In this chapter we present a preliminary quantitative analysis of non-binary gender identities in a large-scale knowledge base: Wikidata.[5] Wikidata is a collaborative project that allows the editing of knowledge—and even the data model itself—by a broad community of users.[6] The present research constitutes the first step of our project, Wikidata Gender Diversity (WiGeDi),[7] which aims to investigate the issue of how gender identities are represented[8] in the knowledge base.

This study aims to contribute to the growing area of data ethics by offering, for the first time, an empirical exploration of the representation of non-binary gender identities in a large knowledge base, and by providing fresh insights and data to gender studies scholars interested in more qualitative approaches to research.

Since every edit and every user discussion throughout the history of Wikidata is archived in the project itself and made publicly accessible, this study allows us to have a unique and comprehensive overview of *how* non-binary identities have been represented in Wikidata, *what* exactly has been represented, and

3. Suzanne J. Kessler and Wendy McKenna, *Gender: An Ethnomethodological Approach* (Chicago:University of Chicago Press, 1985).

4. Os Keyes, "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." *Proceedings of the 2018 ACM Conference on Human-Computer Interaction* 2 (2018): 1–22.

5. Denny Vrandečić and Markus Krötzsch, "Wikidata: A Free Collaborative Knowledgebase," *Communications of the ACM* 57, no. 10 (2017): 78–85.

6. Alessandro Piscopo, Chris Phethean, and Elena Simperl, "What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata," in *Proceedings of the 2017 International Conference on Social Informatics*, (Springer, 2017), 305-22.

7. Wikidata Gender Diversity will be hosted on https://wigedi.com.

8. Whenever we use the terms "represent" or "representation" throughout the chapter, we are referring to the concept of "knowledge representation" (Davis, Shrobe & Szolovits, 1993).

why the users have made certain choices. We performed our analysis from three different—and complementary—perspectives:

1. the modeling question, looking at how the Wikidata ontology has evolved to support non-binary representation, e.g., by updating the properties that directly or indirectly express gender; we aim to analyze the Wikidata ontology to identify representational issues and potential areas of improvement;
2. the data question, computing statistics about non-binary gender representation in the knowledge base, and analyzing it from a quantitative point of view; also, by comparing non-binary people described in Wikidata to the general population(s) of non-binary people in society;
3. the community question, looking at how the Wikidata community has handled the evolution towards a more inclusive non-binary representation, by analyzing user discussions about the topic in a quantitative way; indeed, gender representation is often intrinsically connected to language.

We believe that only by answering all three questions it will be possible to obtain a comprehensive overview of non-binary gender representation in Wikidata. Previous studies on the topic, such as Klein et al. and Konieczny and Klein,[9] have mostly focused on the gender gap in the data, without looking in detail at the model or at the community's interactions and decision processes. Furthermore, to the best of our knowledge, no published research has yet specifically been centered on the modeling of non-binary identities in Wikidata.

9. Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu, "Monitoring the Gender Gap with Wikidata Human Gender Indicators," in *Proceedings of the 12th International Symposium on Open Collaboration*, ACM, 2016, 1–9; Piotr Konieczny, and Maximilian Klein, "Gender Gap Through Time and Space: A Journey Through Wikipedia Biographies via the Wikidata Human Gender Indicator," *New Media & Society* 20, no. 12 (2018): 4608–33.

Based on the research questions listed above, we discuss the theoretical background (Section 2, *Background*) and the state of the art of studies about gender in Wikidata and other knowledge bases (Section 3, *State of the Art*). Then, we present an overview of the current Wikidata model of gender, and a timeline of its historical evolution (Section 4, *The Model*); a set of statistics about gender representation in Wikidata, and non-binary identities in particular (Section 5, *The Data*); a corpus of user discussions about gender called *WiGeTa-En*, and a preliminary analysis of it based on computational linguistics techniques (Section 6, *The Community*). Finally, we conclude with a discussion of the current status of non-binary representation in Wikidata.

Background

We begin our study from the fact that gender is not binary. The binary view that has been prevalent in most of the world until the current century is in fact quite recent[10] and not universal.[11] In this traditional view, gender consisted of a binary classification that allowed only two slots, "man" and "woman," corresponding to two sexes, "male" and "female." Sex was assigned to each person at birth by a doctor based on the person's external anatomy, without regard for genetics, hormonal factors, or identity.

Since the 1970s, the view that gender is a social construct has become prevalent in the scientific community.[12] Gender studies scholars distinguish between *sex assigned at birth* (e.g., female), *gender identity* (e.g., woman), and *gender expression* (e.g., androgynous).[13] Another term that has recently been proposed

10. Leah DeVun, *The Shape of Sex: Nonbinary Gender from Genesis to the Renaissance* (New York: Columbia University Press, 2021).

11. Gilbert Herdt, *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History* (Princeton, NJ: Princeton University Press, 2020),.

12. Judith Butler, *Gender Trouble: Feminism and the Subversion of Identity*. (Routledge, 1990).

13. Julia Serano, *Whipping Girl: A Transsexual Woman on Sexism and the Scapegoating of Femininity* (UK: Hachette, 2016).

is *gender modality*, to describe the correspondence between sex assigned at birth and gender identity (e.g., cisgender).[14] A person's gender identity may or may not correspond to the sex assigned at birth, and a person's gender expression may not reflect the gender roles associated with their gender identity by society.

A wide spectrum of identities exists outside the traditional binary view of gender. We hereby provide a few definitions for the reader's convenience. However, it should be noted that the reality of these identities is much more complex and varied than these broad definitions may suggest.

First of all, the term *transgender* (or *trans*) indicates any person whose gender identity is different from the one assigned to them at birth.[15] *Trans women* are women who are assigned male at birth, while *trans men* are men who are assigned female at birth. On the contrary, the term *cisgender* (or *cis*) describes people who identify with the same gender that is assigned to them at birth.

Non-binary people have a gender identity that falls outside the gender binary. A non-binary person may or may not necessarily identify as trans; therefore, we prefer to talk about *trans and gender-diverse* identities to refer to these communities in an inclusive way.

Intersex people are people whose sex is not classifiable in a binary way. The term has traditionally referred to sexual characteristics; however, it is also used for self-identification.[16]

In recent decades, the LGBTQIA+[17] movements have worked to reclaim as valid the identities of people who do not fit into their assigned gender.[18] This work has resulted in a wider societal

14. Florence Ashley, "'Trans' Is My Gender Modality: A Modest Terminological Proposal" *Trans Bodies, Trans Selves* (2021).

15. Serano, *Whipping Girl.*

16. Jens M. Scherpe, Anatol Dutta, and Tobias Helms, *The Legal Status of Intersex Persons* (Intersentia, 2018).

17. Lesbian, Gay, Bisexual, Trans, Queer, Intersex, Asexual, and other sexual orientations and gender identities.

18. Lisa M. Stulberg, *LGBTQ Social Movements* (John Wiley & Sons, 2018).

acceptance and some limited legal recognition, but significant erasure and discrimination persist.[19]

It is important to note that knowledge bases can directly or indirectly contribute to erasure of trans and gender-diverse people. As discussed in Sandberg,[20] the people who are tasked with modeling and cataloging biographical data have important ethical responsibilities that should not be overlooked. This is also true for Wikidata, where the user community holds a collective responsibility over the data.

Given the complex and interlocked nature of gender identity, sex assigned at birth, and gender expression, we cannot study each of these concepts in an isolated way, but rather we need to consider them holistically when looking at gender modeling in Wikidata.

State of the Art

To date, there have not been many studies about gender in Wikidata. The first scholars to approach the subject were Klein et al. and Konieczny and Klein,[21] 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 that, in general, Wikidata appears to be affected by the same gender disparities that exist in society at large. The most recent project by

19. Michelle Dietert and Dianne Dentice, "Gender Identity Issues and Workplace Discrimination: The Transgender Experience," *Journal of Workplace Rights* 14, no. 1 (2009): 121–140.

20. Jane Sandberg, *Ethical Questions in Name Authority Control* (Sacramento, CA: Litwin Books, 2009).

21. Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu, "Monitoring the Gender Gap with Wikidata Human Gender Indicators," in *Proceedings of the 12th International Symposium on Open Collaboration* (ACM, 2016), 1-9; Piotr Konieczny and Maximilian Klein, "Gender Gap Through Time and Space: A Journey Through Wikipedia Biographies via the Wikidata Human Gender Indicator," *New Media & Society* 20, no. 12 (2018): 4608–33.

the authors is Humaniki,[22] a tool showing the gender gap among all Wikimedia projects.

Hollink, Van Aggelen, and Van Ossenbruggen[23] measured gender differences in a subset of Wikidata entries. Zhang and Terveen[24] recently conducted a case study about the Wikidata gender content gap. The Wikidata Community Survey 2021[25] has looked at gender metrics in the community of Wikidata editors which show that the Wikidata community is overwhelmingly male (75%), while female users make up just 16% and non-binary users 2.9%.[26] The remaining users (6%) opted not to answer the question.[27]

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

Antin et al.[28] were the first to study gender differences in Wikipedia editing, while Reagle and Rhue[29] looked at gender bias in the content of the encyclopedia. Wagner et al.[30] analyzed how men and women are portrayed in Wikipedia. Johnson et

22. https://whgi.wmflabs.org.

23. Laura Hollink, Astrid Van Aggelen, and Jacco Van Ossenbruggen, "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* (ACM, 2018):381–85.

24. Charles Chuankai Zhang and Loren Terveen, "Quantifying the Gap: A Case Study of Wikidata Gender Disparities," in *17th International Symposium on Open Collaboration* (2021): 1–12.

25.https://commons.wikimedia.org/wiki/File:Wikidata_Community_Survey_2021.pdf.

26. The specific question that was asked was "What is your Gender?", and the possible answers were "Woman", "Man", "Non-binary", "Prefer not to disclose", and "Prefer to self-describe" (see page 15 of the Wikidata Community Survey).

27. See page 23 of the Wikidata Community Survey.

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

29. Joseph Reagle and Lauren Rhue, "Gender Bias in Wikipedia and Britannica," *International Journal of Communication* 5 (2011): 21.

30. Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier, "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia," in *Ninth International AAAI Conference on Web and Social Media* (2015).

al.[31] looked at gender differences among the readers of the encyclopedia. Field, Park, and Tsvetkov[32] analyzed social biases in Wikipedia biographies, while Tripodi[33] investigated the frequent deletion of biographies about women.

More recently, Redi et al.[34] have created a taxonomy of knowledge gaps found in Wikimedia projects. Miquel-Ribé & Laniado[35] 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[36] 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 differs from the previous ones because it is the first to center trans, non-binary, and other gender-diverse identities. Furthermore, we 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.

---

31. Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia, "Global Gender Differences in Wikipedia Readership." *arXiv Preprint arXiv:2007.10403*,

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

33. Francesca Tripodi, "Ms. Categorized: Gender, Notability, and Inequality on Wikipedia." *New Media & Society* (2021), 14614448211023772.

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

35. M. Miquel-Ribé and D. Laniado, "The Wikipedia Diversity Observatory: Helping Communities to Bridge Content Gaps Through Interactive Interfaces," *Journal of Internet Services and Applications* 12, no. 1 (2021), 1-25.

36. M. Miquel-Ribé, A. Kaltenbrunner, and J.M. Keefer, "Bridging LGBT+ content Gaps Across Wikipedia Language Editions," *International Journal of Information, Diversity, & Inclusion (IJIDI)* 5, no. 4 (2021): 90-131.

The Model

In this section we describe the Wikidata modeling of gender and its evolution through time. Due to the open and collaborative nature of Wikidata, the model is fluid and constantly changing.[37]

The Wikidata model defines two basic types of entities: *items* and *properties*. Each Wikidata page describes a single item through one or more *labels* (multilingual strings of text representing the name of the item), one or more *aliases* (alternative names), one or more *descriptions* (multilingual strings of text), and one or more *statements*.

Each Wikidata statement expresses a fact that is known about the item, and is composed of a property, a value, and, optionally, one or more qualifiers and one or more references. For example, the Wikidata item *Q173399 Elliot Page* is connected by the property *P27 country of citizenship* to the value *Q16 Canada*.[38]

Since the term "item" is not widely used in the field of data modeling and may create confusion, in the following we will use the more general term "entity" to refer to Wikidata items.

Ontological Representation of Gender

In this section, we will describe the ontological representation of gender that has been adopted by Wikidata.[39]

In Wikidata, gender is modeled using the property *P21 sex or gender*, connecting a person (an entity that is an instance of

37. Alessandro Piscopo, Chris Phethean, and Elena Simperl, "What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata," in *Proceedings of the 2017 International Conference on Social Informatics*, (Springer, 2017): 305–22.

38. For more details about the data model, see https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer.

39. This is based on an analysis conducted in December 2021. To avoid reporting outdated information, we plan to publish an overview of the model that is frequently updated, in an automated way, on the website of our project (https://wigedi.com).

Q5 human) to one or more entities representing sex or gender.[40] Some of these entities are explicitly allowed as values of the property, while others are not. This distinction is evaluated through community discussions, and it is implemented through property constraints.[41] However, these constraints do not prevent the users from setting any value of their choice and are simply used to check for possible errors *after* the sex or gender has already been set.

Since its creation in 2013, the *P21* property has conflated the concepts of *sex* and *gender*, and this ambiguous nature of the property has led to many discussions and controversies (see Section 6, *The Community*). However, no significant changes to the definition of *P21* have been made in the last eight years since the creation of the property.

In the following, we will look at the possible values of *P21* and at their taxonomy. At this stage, we focus only on the labels of each entity without looking at its description. The reason for this is that the label of a Wikidata entity is often stable and consistent across different languages, while the description may change significantly over time and across languages. Therefore, listing the current English description of each entity would be quite misleading.

The allowed values for *P21* include:

1. instances of Q48264 gender identity. At present, there are 59 instances of gender identity, of which 23 are currently in use, and 3 more are reported as allowed on the property's discussion page. These are reported in Table 1.
2. instances of Q290 sex. At present, there are 24 instances of sex, of which 8 are currently in use, and 1 more is reported as allowed on the property's discussion page.

40. P21 can also applied to fictional human or other entities. It should be noted that each entity, including humans, is allowed to have multiple values of P21.

41. https://www.wikidata.org/wiki/Help:Property_constraints_portal.

These are reported in Table 2.

3. 8 other values from the set reported in Table 3, of which 7 are currently in use. These are values that are neither instances of Q48264 gender identity nor instances of Q290 sex, and their classification is highly varied (see below).
4. the unknown value.

In total, 31 entities are currently used as values of *P21*, and 40 entities are explicitly allowed as values of the property according to the constraints listed on *P21*'s discussion page. One entity is currently in use but not allowed.[42]

Fig. 1 shows the current class taxonomy of the gender entities reported in Table 1, including only those that are presently used as values of *P21*.

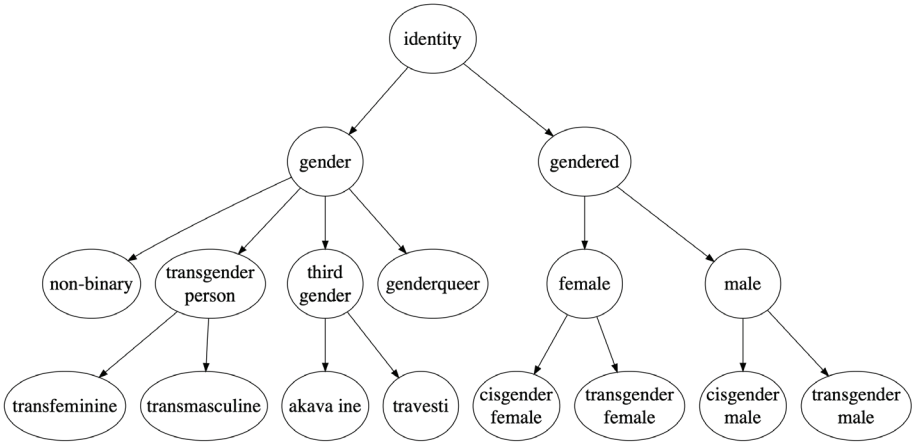

Figure 1: The taxonomy of gender in Wikidata

42. The entity is hermaphrodite, which is widely considered a derogatory term when applied to humans.

Table 1: Gender values of P21

Wikidata ID	English Label	Allowed Value	Usage
Q108876763	abinary	No	–
Q505371	agender	Yes	27
Q104838508	alyha	No	–
Q97595519	androgyne	No	–
Q97577404	aporagender	No	–
Q859614	bigender	Yes	6
Q107144810	binabinaaine	No	–
Q56388896	calabai	No	–
Q65212675	calalai	No	–
Q15145779	cisgender female	Yes	21
Q107785560	cisgender gay male	No	–
Q15145778	cisgender male	Yes	12
Q1093205	cisgender person	No	–
Q93954933	demiboy	Yes	2
Q63715923	demigender	No	–
Q93955709	demigirl	Yes	–
Q1399232	fa'afafine	Yes	4
Q107427210	fakafifine	No	–
Q350374	fakaleitī	Yes	–
Q6581072	female	Yes	1825280
Q11491595	gender identity disorder	No	–
Q56314793	gender incongruence	No	–
Q106647285	gender modality	No	–
Q99485732	gendered	No	–
Q106781857	genderfaun	No	–
Q18116794	genderfluid	Yes	43
Q12964198	genderqueer	Yes	40
Q660882	hijra	Yes	1
Q11713472	intergender	No	–
Q104717073	intersex person	No	–
Q106990131	isogender person	No	–
Q746411	kathoey	Yes	2
Q6581097	male	Yes	5741522
Q82028886	maverique	No	–
Q24886035	mudoko dako	No	–
Q3277905	māhū	Yes	6
Q1289754	neutrois	Yes	2
Q48270	non-binary	Yes	510
Q69990794	non-binary human	No	–
Q48796147	nádleehi	No	–
Q7130936	pangender	Yes	2
Q64606208	polygender	No	–
Q3404005	questioning	No	–
Q106647045	sekhet	No	–
Q27679684	transfeminine	Yes	8
Q1052281	transgender female	Yes	1125
Q2449503	transgender male	Yes	295
Q189125	transgender person	Yes	29
Q27679766	transmasculine	Yes	8
Q107502361	transneutral	No	–
Q17148251	travesti	Yes	14
Q7841680	trigender	No	–
Q301702	two-spirit	Yes	17
Q108266757	vakasalewalewa	No	–
Q104834145	waria	No	–
Q8025501	winkte	No	–
Q96000630	x-gender	Yes	–
Q108854353	xenogender	No	–
Q8053770	yinyang ren	No	–

Table 2: Sex values of P21

Wikidata ID	English Label	Allowed Value	Usage
Q4700377	akava'ine	Yes	1
Q59592239	altersex	No	–
Q4849481	bakla	No	–
Q2904759	bissu	No	–
Q106610856	endosex	No	–
Q6581072	female	Yes	1825282
Q43445	female organism	Yes	4530
Q1054122	futanari	No	–
Q106647285	gender modality	No	–
Q430711	gynandromorph	No	–
Q303479	hermaphrodite	No	1
Q1097630	intersex	Yes	133
Q28873047	intersex organism	Yes	–
Q1062222	khanith	No	–
Q25035965	koekchuch	No	–
Q6538491	lhamana	No	–
Q6581097	male	Yes	5741522
Q44148	male organism	Yes	8927
Q30689479	meti	No	–
Q24886035	mudoko dako	No	–
Q3333006	mukhannathun	No	–
Q3177577	muxe	Yes	1
Q20577996	sex reassignment	No	–

Table 3: Other values of P21

Wikidata ID	English Label	Allowed Value	Usage	Instance of
Q207959	androgyny	Yes	10	gendered expression/identity
Q179294	eunuch	Yes	251	social status/job/physiological condition/occupation
Q64017034	cogenitor	Yes	1	fictional sex
Q52261234	neutral sex	Yes	13	no class
Q16674976	hermaphroditism	Yes	7	reproductive system
Q48279	third gender	Yes	2	subclass of sex; subclass of gender
Q56315990	assigned female at birth	Yes	1	assigned gender
Q25388691	assigned male at birth	Yes	0	assigned gender

As shown in the figure, the top gender classes are *gender* ("range of physical, mental, and behavioral characteristics distinguishing between masculinity and femininity") and *gendered* ("state of having gender identity"). The class *gender* has subclasses *non-binary*, *genderqueer*, *third gender* and *transgender person*, while the class *gendered* has subclasses *male* and *female*.

The class *female* has subclasses *cisgender female* and *transgender female*, while the class *male* has subclasses *cisgender male* and *transgender male*. The class *transgender person* has subclasses *transfeminine* and *transmasculine* (which are not widely used as values). The class *third gender* has subclass *travesti* (a Latin American gender identity often considered a third gender). The class *non-binary* is the superclass of most of the remaining identities.

The current Wikidata gender taxonomy is unusual and, to the best of our knowledge, not based on any model of gender that is described in the literature. It is unclear why the distinction between *gender* and *gendered* exists. This distinction appears to have been added in 2020 to replace a previous sex-based classification of *female* and *male* (e.g., *female* was a subclass of *female organism*) with a gender-based one but, supposing that this was the intention, it would have been more consistent to simply make *female* and *male* subclasses of *gender*.[43]

Let's now look more in detail at the subclasses of *non-binary* that are reported in Fig. 2. The class *genderqueer* is not a subclass of *non-binary* but, rather, is connected to it through the property *P460 said to be the same as*, indicating that this class is "said to be the same as that item, but it's uncertain or disputed." The subclasses of *non-binary* are: *agender*, *bigender*, *demigender, fa'afafine*, *genderfluid*, *hijra*, *kathoey*, *māhū*, *neutrois*, *pangender*, and *two-spirit*. We report statistics about the usage of these non-binary identities in Section 5, *The Data*.

43. Even though, as we said in the Background section, it is more common to use the terms woman and man to refer to binary genders and female and male when referring to sex assigned at birth.

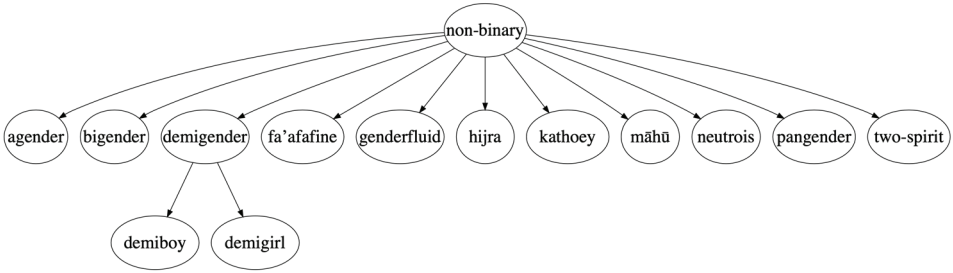

Figure 2: The taxonomy of non-binary gender in Wikidata

The current Wikidata model of non-binary identities is very simple, having only one main class and several subclasses, with no further levels except for *demigender*, which is a superclass of *demiboy* and *demigirl*. This simple taxonomy is consistent (meaning that there are no obvious contradictions); it can be considered accurate only insofar as *non-binary* is recognized as an umbrella term by every person who is described in Wikidata and identifies as one of the identities listed as subclasses of *non-binary*. Unfortunately, it would be very difficult to verify whether this is true or not.[44]

Fig. 3 shows the class taxonomy of the sex entities reported in Table 2, excluding *male* and *female*, which are currently not connected to the sex-based class tree.

As shown in Fig. 3, the current classification of sex contains *female human* and *male human* classes; however, these are not allowed as values of *P21*. The classes *female organism* and *male organism* are used for animals, while the class *intersex* is used for both humans and animals. The class *hermaphrodite* (a term generally considered offensive when applied to people) is a subclass of *intersex*.[45]

44. It is interesting to note that according to Wikidata's verifiability policy (https://www.wikidata.org/wiki/Wikidata:Verifiability), every statement that is collected in the knowledge base should be properly sourced; however, very often the structural subclass of relations that make up the model are not sourced at all.

45. At present, this value is used to express the gender of a single entity (a snail).

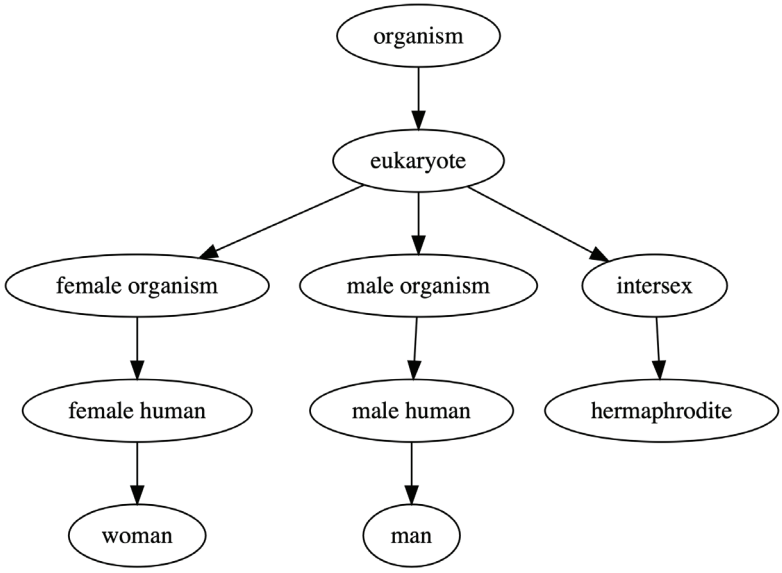

Figure 3: The taxonomy of sex in Wikidata

The remaining values of *P21*, which we reported in Table 3, have highly varied classifications that fall outside of the main ones reported in Fig. 2 and Fig. 3. Providing a detailed description of each is beyond the scope of this chapter; however, we will list them briefly for the reader's convenience. The entity *androgyny* is an instance of *gendered expression* and *identity* but, at the same time, it is also a subclass of *sexual diversity*. The entity *eunuch* is an instance of *social status*, *job*, *physiological condition* and *occupation*, but it is also a subclass of *man*[46] and *castrated creature*. The entity *cogenitor* is an instance of *fictional sex*. The entity *neutral sex* had no class assigned at the time of data collection. The entity *hermaphroditism* is an instance of *reproductive system* and a subclass of *sexual reproduction*. The entity *third gender* is a subclass (not an instance) of both *sex* and *gender*. Finally, the entities *assigned female at birth* and *assigned male at birth* are instances of *assigned gender* but also, strangely, subclasses of *assigned gender* and subclasses of *female* and *male* (respectively).

46. Confusingly, this entity Q8441 man is not an allowed value for P21.

Gendered Properties

Apart from *P21*, gender is also expressed through other properties, albeit in a more implicit way. In particular, the properties *P25 mother* and *P22 father* express gendered family relations.[47] For a brief historical overview of gendered family properties, see section *Gendered Properties for Family Relations* below.

In addition, gendered properties exist for the representation of athletes and sports competitions, which are often divided by gender (for example, the many properties linking athletes to their descriptions in external databases). At this stage, we are not aware of other non-familial properties that have been divided by gender.

A related issue is that of the property labels, which, in several languages, are affected by the lack of gender-neutral terms. For example, *sibling* in Italian can only be rendered as *fratello o sorella* (brother or sister). We plan to investigate property labels in a future study.

A Timeline of Wikidata Gender Modeling

This section reports a timeline of the main events related to the modeling of gender in Wikidata.

The Beginning

Wikidata opened to the public on October 25, 2012. In the initial stage of the project, the users focused on importing data from the existing Wikipedia. The items Q44148 male and Q43445 female were imported on November 13, 2012. On the same day, a user created the item Q48270 genderqueer, which would later be renamed "non-binary." The concept of non-binary identities was thus present in Wikidata since a very early date.

47. A property P8810 parent (unspecified) has recently been created to express a generic parent-child relationship, but it is not meant to replace mother and father.

On November 28, 2012, the item *Q189125 (transgender)*, later renamed *transgender person*, was imported from Wikipedia. The item's description received immediate transphobic vandalism from an anonymous user, which went unremarked upon for a whole year. Similarly, the items *Q1052281 (trans woman)* and *Q2449503 (trans man)* received incorrect English descriptions (e.g., "a person born male but identifying as female" for trans woman) that were not fixed for more than a year.

The first mention of "gender" in the Project Chat, the main English-language discussion page, was made on December 6, 2013. The users discussed the representation of gender and how to source it properly. In this initial stage of the project, there were often tensions between users who favored *completeness* (i.e., Wikidata should grow as fast as possible) versus those who favored *accuracy*. (i.e., every Wikidata statement should be properly sourced).

Looking specifically at gender, some users in the early Wikidata community did not understand the complexity of the gender modeling issue at all ("A sex property needs only male/female options. Demanding reference for that [makes] it look funny"), but there were a few who acknowledged the existence of non-binary people ("There are people [whose] sex can't readily be described as male or female").[48]

Creation of P21

The history of *P21* began on February 2, 2013, when a proposal was made in the Property Proposal section of Wikidata to create a property to represent human gender. After a short discussion, shown in Fig. 4,[49] the property was created on February 4, 2013.

48. https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2013/02#Reference.

49. The names of the Wikidata users participating in the discussion have been redacted.

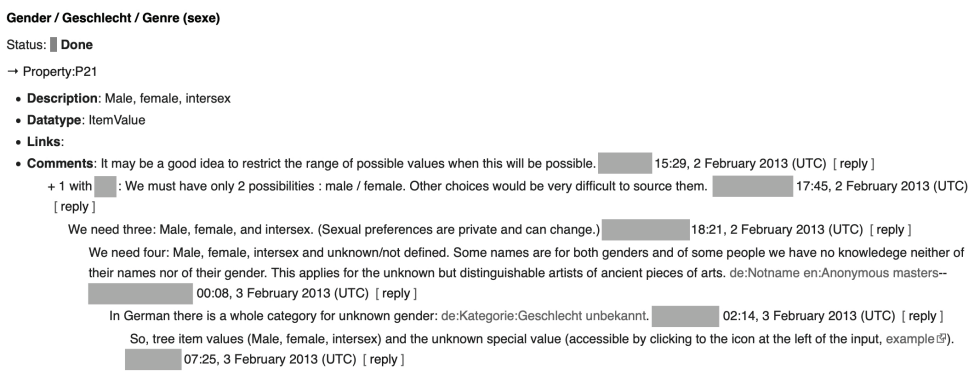

Figure 4: The discussion about the creation of P21

Initially, the property was called *P21 gender*, and the only allowed values were *male*, *female*, *intersex*, and *unknown*. However, on the day of its creation, the labels for the property that were set in various languages (Italian, Portuguese, Czech) referred to *sex*, not gender. This created an initial confusion that was not resolved until December 2013, when, after a long discussion, Wikidata users decided to conflate the concept of sex with the concept of gender and change the property labels in all languages to *sex or gender*.[50]

However, the conflation of the two concepts generated additional ambiguities; for example, with regard to the representation of transgender people (see later section on *Representing Trans Identity*), because in many cases, the sex assigned at birth and the gender identity of a person are different.

The Rise of Bots

On February 6, 2013, shortly after the creation of *P21*, a user requested permission to use a bot to add some statements, including gender statements, to Wikidata items, based on information from

50. https://www.wikidata.org/wiki/Property_talk:P21/Archive_1#Transgender_/_Cisgender_changes.

Wikipedia categories: "A good example is en:Category:Women physicists. We can safely assume that all members of that category are female."[51] The proposal was approved on February 17, and the gender data started being populated automatically.

From this point onward, the automatic addition of gender data to people became routine. In the first two years of the project, at least seven bots,[52] each developed by a different user, added gender to Wikidata from various sources. Through these means, millions of entities representing people were assigned a gender. In June 2015, it was announced that gender data completeness had reached 93.8%.[53]

The main sources used by the bots were as follows:

1. Wikipedia categories, i.e., the gendered categories found in some Wikipedia language editions (for example, the German Wikipedia category Mann for men, or the English Wikipedia category Women Physicists).
2. External databases, such as VIAF (Virtual International Authority File) and GND (Gemeinsame Normdatei, the German Integrated Authority File).
3. Personal names, as listed in the Wikidata label of the entity or in the title of the corresponding Wikipedia article(s) (e.g., all people named Alice would be marked as women).
4. Personal pronouns, by counting the occurrences of each pronoun in the corresponding Wikipedia article(s), the reasoning being that the most frequent pronoun would be correlated to the gender.[54]

51. https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Legobot_.

52. The names of the bots are: Legobot, Dexbot, Sk!dbot, JAnDbot, VIAFbot, SamoaBot, and Reinheitsgebot. Three of these are still active today, but they are performing different tasks unrelated to gender.

53. https://www.wikidata.org/wiki/Property_talk:P21/Archive_1#Pie_chart.

54. This was initially proposed on June 10, 2013, in the following discussion: https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/SamoaBot_33.

Unfortunately, the extraction of gender data from the latter two sources was highly problematic. While Wikipedia categories and external databases likely received at least some overview from human users, the extraction of gender data from personal names relies on the mistaken assumption that a personal name can be applied only to men or only to women. This is not true in general, as even with the most gendered names there are often exceptions, but furthermore, some names are applied to different genders in different languages.

The creation of gender data from personal pronouns, which luckily was performed on a more limited scale, is also flawed, as it relies on the incorrect notion that *he/him* pronouns are applied only to men, and *she/her* only to women. In fact, personal pronouns can be wholly independent of both sex assigned at birth and gender identity.

The systematic process of gender data population through bots introduced significant errors in Wikidata, which then had to be manually corrected by users through the effort of projects focused on gender diversity, such as Wikiproject LGBT[55] and Art+Feminism.[56] However, given the wide scale of gender-related bot activity, it is likely that a significant number of errors are still present in the knowledge base.[57]

Gendered Properties for Family Relations

The second big issue that the Wikidata community had to solve was the use of gendered properties; for example, *mother/father*, *brother/sister*, *uncle/aunt*, etc. This is an issue because it makes it impossible to include non-binary people in family relations. Users began questioning this model from the early days of the project, but it took a long time to bring meaningful change.

55. https://www.wikidata.org/wiki/Wikidata:WikiProject_LGBT.

56. https://artandfeminism.org

57. It should be noted that some (more limited) semi-automatic additions of gender data are still being performed today through newer tools, such as PetScan and QuickStatements. These are more difficult to track, but we intend to do so as future work.

The properties *uncle/aunt* were replaced with *relative* in 2013.[58] The properties *brother/sister* were replaced with *sibling* in 2017, and the same was done for *stepfather/stepmother*, replaced with *stepparent*. The replacement of *brother/sister* took extensive discussions and faced significant opposition from a subset of the Wikidata community, especially due to linguistic issues (several languages lack a word for *sibling*).

The replacement of *mother/father* with *parent* was proposed several times throughout the years (in 2013, 2015 and 2016) but unfortunately, the proposal repeatedly failed to reach the wide consensus required for its approval. As of March 30, 2022, the gendered properties *mother* and *father* still remain.

Representing Trans Identities

Once the gender data was populated, the discussion shifted to the representation of trans identities. This first became an issue in the Wikidata community in August 2013 when the American activist and whistleblower Chelsea Manning publicly announced her trans identity.[59]

Before that time, trans men and women had been quietly added to the knowledge base, sometimes using the *transgender male/female* values for P21, other times simply using *male/female*. The required changes in Chelsea Manning's name and gender identity faced significant opposition, leading to edit wars (i.e., disputes where opposing editors continually change the statements without significant discussion), deadnaming (i.e.,

58. This may appear to result in a loss of information but, in fact, the model that was adopted involves the use of a qualifier kinship to subject to list the specific familial relation.

59. The issue was raised in a brief discussion on the Wikidata item's talk page (https://www.wikidata.org/wiki/Talk:Q298423) while the item itself was in the middle of an edit war. (https://www.wikidata.org/w/index.php?title=Q298423&offset=20130901000000&limit=100&tagfilter=&action=history).

when users referred to a trans person by a name they used prior to transitioning), and transphobic comments.

Several early Wikidata items about trans people faced the same issues, after which the community ultimately started labeling trans men and women more consistently as *transgender male/female*. However, in 2014, the classification of these two gender identities was changed in a highly problematic way: *transgender male* was no longer a subclass of *male*, and *transgender female* was no longer a subclass of *female*, meaning that a user querying Wikidata for all women, for example, would not receive any transgender woman as output. Trans men and women were effectively made invisible. This issue was solved only after two years, in 2016, when the correct classification was restored.

Discussion

Unfortunately, due to space limitations we are not able to provide a complete timeline of the evolution of the modeling of gender in Wikidata and of how the model was populated by the users. However, from the abridged timeline reported above, we can already gather the following important facts:

1. the Wikidata gender model has evolved over time;
2. such evolution has been the product of extensive user discussions;
3. such evolution has been influenced by historical events, and in particular by changes in societal acceptance of gender diversity (as evidenced by user discussions);
4. the actual production of gender data has been influenced by inaccurate assumptions about gender held by the users who were participating in the project.

We plan to report a more detailed timeline as an outcome of the first phase of our project, publish the full timeline on the project website, and discuss our findings in a future publication.

The Data

In this section, we report statistical data about non-binary people described in Wikidata.[60] As explained in the Introduction, we have decided to focus this quantitative analysis on people whose gender is explicitly reported as *non-binary* (or any of its subclasses) in Wikidata, based on data collected in April 2022. We plan to perform a wider, more extensive study of gender diversity at a later stage of our project.

Gender Identity Distribution

First of all, as a general point of reference, we will look at the distribution of the values of *P21 sex or gender* among humans represented in Wikidata. This distribution is reported in Fig. 5.

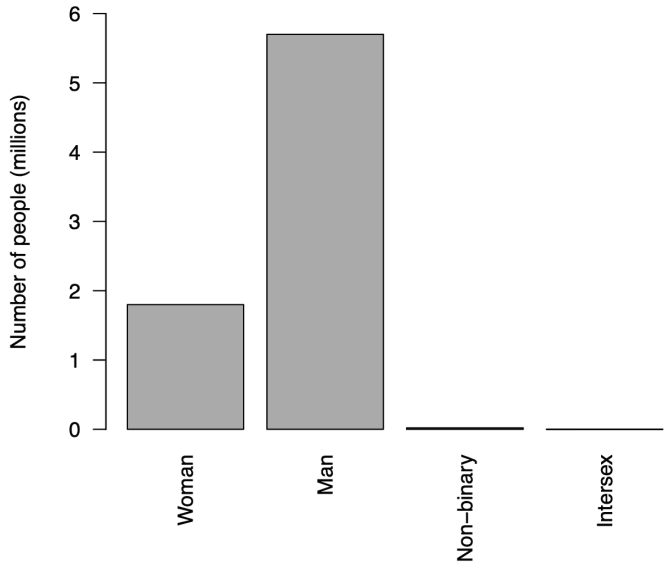

Figure 5: Distribution of sex and gender in Wikidata

60. Disclosure: The first author of this paper is represented in this dataset.

Wikidata contains approximately 1.8 million entities representing women and 5.7 million entities representing men. In Fig. 5, we have included both trans and cis women in the *woman* category, and both trans and cis men in the *man* category.[61] The number of trans women is 1089 (about 61 per 100,000 women), while the number of trans men is 282 (about 5 per 100,000 men).

The most glaring fact that emerges from the chart is the very large gender gap between men and women. The number of men in Wikidata currently outnumber women by almost 5-fold. The other significant result is that the percentage of non-binary people represented in Wikidata is extremely low (585, or about 8 per 100,000), as is the percentage of intersex people (132, or about 2 per 100,000).

These values indicate a severe underrepresentation of gender-diverse identities, given that the percentage of non-binary people is estimated to be about 360 per 100,000,[62] while the prevalence of intersex people is at least 18 per 100,000.[63] The number of trans people is also significantly lower than their actual prevalence in society, which is at least 355 per 100,000.[64]

The underrepresentation that we notice is likely influenced by the fact that the retroactive assignment of a non-binary identity to historical people is very difficult to do, and often impossible

61. It should be noted that Wikidata uses female and male labels instead of woman and man, but given the conflation of sex and gender in the model, it is impossible to know whether any specific entity has been classified based on sex assigned at birth or based on gender identity.

62. BDM Wilson and IH Meyer, "Nonbinary LGBTQ Adults in the United States." (Los Angeles, CA: UCLA, Williams Institute, 2021)/

63. The number of intersex people in Wikidata is difficult to compare to statistics about intersex people in society due to the fact that the term intersex can refer both to sexual characteristics and to gender identity, and it is impossible to know which definition has been adopted by each Wikidata user who marked an entity as intersex. However, the large gap that we have identified suggests an actual lack of representation of intersex people in Wikidata; Leonard Sax, "How Common Is Intersex? A Response to Anne Fausto-Sterling," *Journal of Sex Research* 39, no. 3 (2002): 174–78.

64. Lindsay Collin, Sari L Reisner, Vin Tangpricha, and Michael Goodman, "Prevalence of Transgender Depends on the 'Case' Definition: A Systematic Review," *The Journal of Sexual Medicine* 13, no. 4 (2016): 613–26.

to verify with absolute certainty. Indeed, as we will see in the later section, *Non-binary Identities Over Time,* the overwhelming majority of non-binary people in Wikidata were born in the 20th century.

Non-Binary Identity Distribution

The distribution of gender identities among non-binary people is shown in Fig. 6. Unlike in the data reported in Table 1, we include only real humans, excluding all fictional characters and other entities that may have a gender.[65]

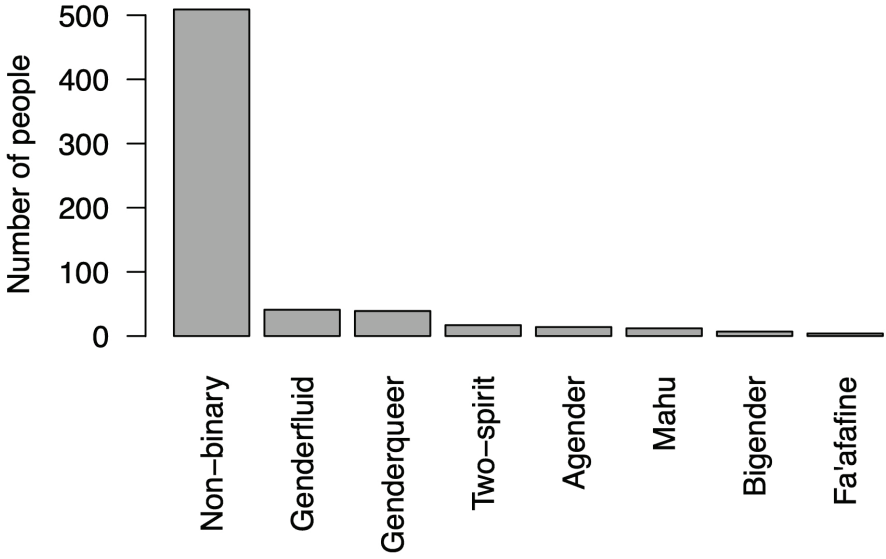

Figure 6: Distribution of non-binary gender identities in Wikidata

Among the 585 non-binary people described in Wikidata, the most represented gender identity is simply *non-binary* (509

65. Fictional characters with gender are less than 1 per thousand people with gender. At present, 5 fictional characters are listed as non-binary, 2 as intersex, 1 as trans, and 4 have one of the values of P21 reported in Table 3.

people), followed by *genderfluid* (41 people), *genderqueer* (39 people), *two-spirit* (17 people), *agender* (14 people), *māhū* (12 people), *bigender* (7 people), and *fa'afafine* (4 people).[66]

In the following sections, we will analyze the distribution of non-binary people based on time, space, and other characteristics. These statistics provide a broad overview of which non-binary people are currently described in the knowledge base.

Non-Binary Identities Over Time

First of all, the distribution of non-binary people based on their birth date is reported in Fig. 7.

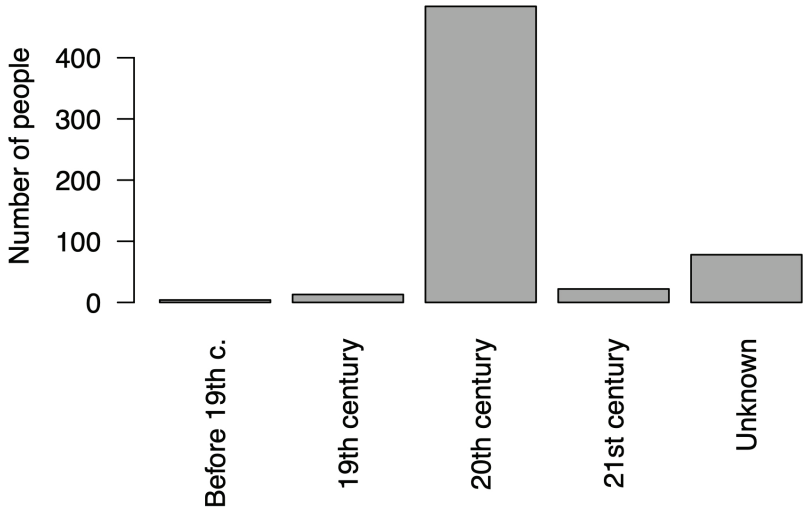

Figure 7. Distribution of non-binary gender identities over time

Fig. 8 reports instead the density of birth years of non-binary people, starting from 1603, which is the first available birth year for a non-binary person in Wikidata.

66. The total is greater than 585 due to the presence of people with multiple non-binary identities.

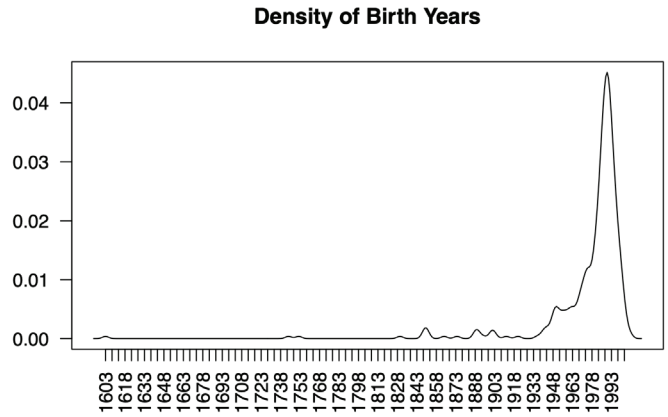

Figure 8. Density of birth years of non-binary people

The oldest person to be considered non-binary in Wikidata is Xu Deng, a Chinese doctor who lived around the year 200.[67] Interestingly, this person is described only in the Chinese and Swedish Wikipedias. Other historical people recognized as non-binary in Wikidata are Thomas/Thomasine Hall (17th century), Theodora de Verdion (18th century), and the Public Universal Friend (18th century).

Thirteen people (2.2%) were born in the 19th century. The vast majority of the people, i.e., 484 (80.4%) were born in the 20th century. 22 people (3.6%) were born in the 21st century. 79 people (13.1%) lack a date of birth or death.

Non-Binary Identities by Country and Language

Looking now at the distribution of non-binary identities by country (Wikidata property *P27 country of citizenship*), we have plotted the distribution on a world map in Fig. 9.[68]

67. This person is not represented in Fig. 8 due to the lack of a birth date. Only an approximate death date of "200s" is reported in Wikidata.

68. We have decided to report birth countries instead of citizenship due to incompleteness of the citizenship data; however, the picture that emerges by plotting citizenship is very similar.

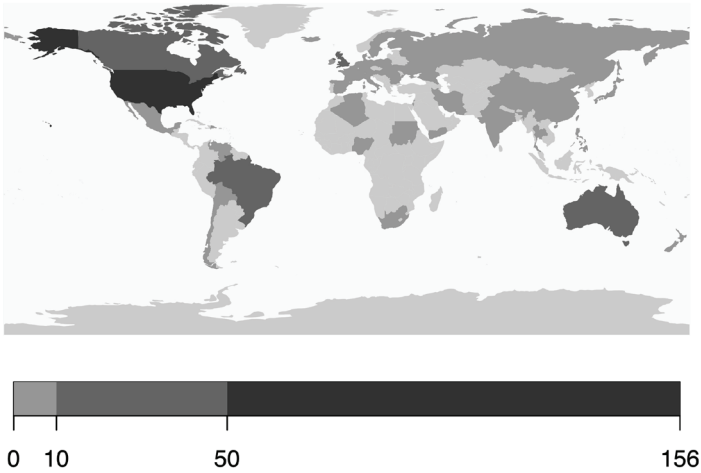

Figure 9: Distribution of non-binary gender identities by country

As shown in the figure, non-binary people represented in Wikidata are distributed throughout the whole world. Most non-binary people represented in Wikidata, however, were born in the United States (156 people), followed by the United Kingdom (30), Canada (27), Australia (13) and Brazil (11). All other countries have fewer than 10 people.

This high prevalence in the Global North is, perhaps, not too surprising given the current legal recognition of non-binary identities, but it is interesting to note how frequently non-binary people from countries outside of the Anglosphere are under-represented in Wikidata. The whole European Union has 56 non-binary people; i.e., about a third of those in the United States and not even twice those in the United Kingdom.[69]

This fact is confirmed by looking at the language(s) spoken by each person. While the data provided by Wikidata is quite

69. It should be noted that the high prevalence of American, British, and Canadian people does not reflect the general statistics about citizenship in Wikidata (https://www.wikidata.org/wiki/Wikidata:WikiProject_Q5/numbers/country_of_citizenship), where the United States is indeed 1st, but the United Kingdom is 5th, and Canada is only 10th. Moreover, people from countries such as France, Germany, or Japan are very highly represented in Wikidata, but the percentage of non-binary people from these countries is extremely small.

incomplete, English is the first language by a factor of 10, as shown in Fig. 10.

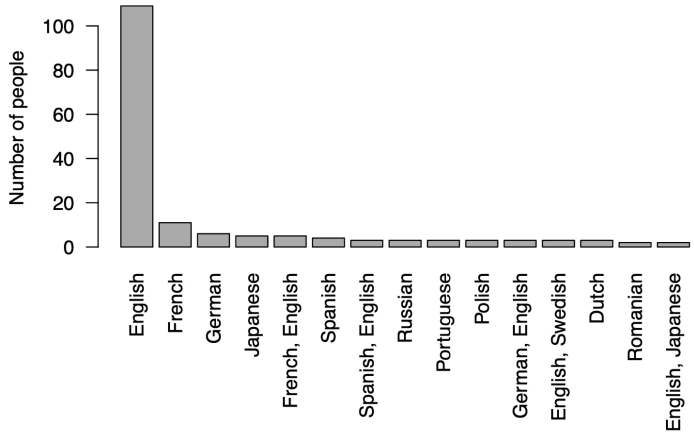

Figure 10: Distribution of non-binary gender identities by language

Non-Binary Identities by Occupation

Finally, we will look at non-binary people by occupation. We believe that this perspective is interesting, because it allows us to compare the distribution of occupations among different gender identities and identify possible gaps in the data (i.e., people who are completely missing from Wikidata, or occupations where non-binary people have been misclassified). The distribution of the 15 most common occupations is reported in Fig. 11.

The most common occupation for non-binary people in Wikidata is *actor* (98 people), followed by *writer* (90 people), *singer* (58 people), and *artist* (37 people).[70] The occupation *LGBTI rights activist* is also common (35 people), as is the more general *activist* (34 people).

70. Some occupations, e.g., "actor" and "film actor" may appear duplicated, but this is the way they are listed in Wikidata, and we decided not to alter them.

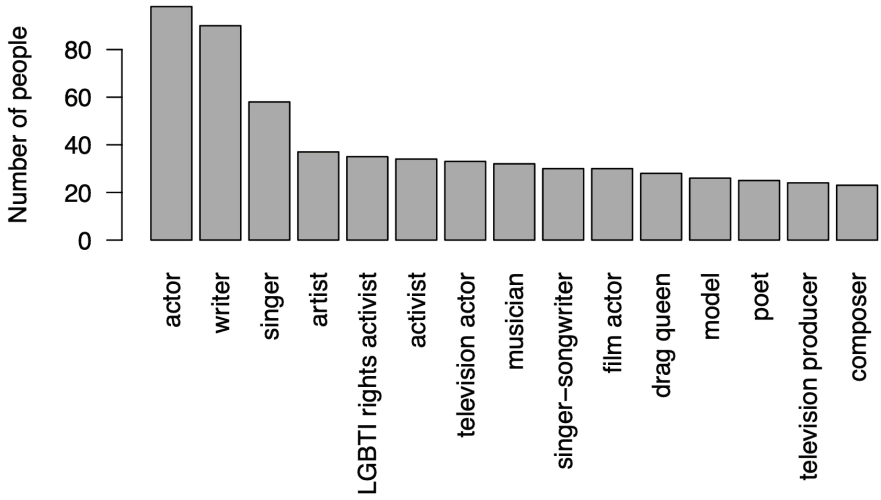

Figure 11: Distribution of non-binary gender identities by occupation.

It is interesting to note that the distribution of occupations for non-binary people is much different from that of men and women. Among men and women, the top occupations include researcher (due to the import of large publication datasets into Wikidata), politician, various subclasses of athlete, and teacher.

In some cases, this under-representation of non-binary people in specific occupations is probably a reflection of societal barriers (e.g., politicians, athletes), but in other cases (e.g., researchers), it may instead be due to incorrect assignment of gender. We plan to conduct a more detailed analysis of these disparities and report it in a future publication.

Discussion

The initial quantitative analysis reported above allows us to have a broad overview about the non-binary people that are described in the Wikidata knowledge base. Our main findings are as follows:

- the number of non-binary people who are described as such in Wikidata is very small, which suggests that they may be significantly under-represented. The same is true for binary trans people and for intersex people.
- the distribution of non-binary people over time is highly skewed towards the present, with an overwhelming majority of people born in the 20th century; this is probably due to the fact that, in the past, fewer people identified as non-binary but, also, that assigning a non-binary identity to a historical figure is often difficult to justify. We intend to explore this topic further in a future study.
- the distribution of non-binary people over space is highly skewed towards the Global North and, in particular, the Anglosphere, despite the fact that Wikidata is a highly multilingual project. In our assessment, this may be due to one or more of the following factors: (i) non-binary may actually be over-represented in such countries (but this explanation seems simplistic); (ii) non-binary people from these countries may declare more openly their gender identity, thus increasing the chance that this information ends up in Wikidata; (iii) Wikidata users who edit from the Anglosphere may be more eager to represent the identity of gender-diverse people.
- the distribution of non-binary people by occupation suggests that most of them work either in the creative arts or as gender rights activists. Some professions that are common among men and women are significantly under-represented in non-binary people, which may be explained by one or more of the following factors: (i) an actual difference in society; e.g., there are few non-binary politicians and non-binary sportspeople due to gatekeeping that excludes them; (ii) non-binary people who hold such

occupations may declare less openly their gender identity, thus preventing it from ending up in Wikidata; (iii) Wikidata users who focus on editing certain occupations may be more eager to represent the identity of gender-diverse people.

The Community

This section describes our work on the Wikidata community—more specifically, looking at user discussions. The main goal is to analyze how the narrative around gender-related topics has changed during Wikidata's nine years of existence. For this purpose, we created a specialized corpus of Wikidata discussions related to gender, composed of 613 Wikidata English discussions from October 2012 till September 2021, and performed an unsupervised topic analysis on this corpus.

We have chosen this unconventional approach for studying the community because it is very difficult to elicit gender identity data from the Wikidata community of users—namely, Wikidata does not require users to declare any identity features, such as nationality, age, and of course gender identity, at the time of registration.[71] Therefore, we can only study what the users do on Wikidata—that is, engaging in discussions and editing the knowledge base—and not who the users are.

Corpus-based studies are currently very popular in linguistics, and we believe that building a corpus, and corpus analysis, could allow us to keep track of the narrative about gender identities in the Wikidata community, see how the narrative has changed over time, study the impact of LGBTIQ+ movements and how they have brought awareness in Wikidata discussions, and detect cases of hate speech, sexism, misgendering, etc.

71. Wikidata also allows users to participate without registering at all. In this case, they will be identified by their IP address.

The WiGeTa-En Corpus

Wikidata Gender Talks – English (*WiGeTa-En*) is a specialized corpus of Wikidata discussions related to gender. The corpus was collected semi-automatically, first using scraping techniques to collect all the discussions containing relevant keywords as an automatic filter, and subsequently through a manual annotation of the relevant discussions for the corpus.

In particular, we analyzed a total of 2,511 Wikidata discussion pages. These pages were automatically extracted using the Wikidata API by searching for a list of 79 relevant keywords, which included gender identities such as "non-binary" or "woman"; sex-related terms such as "male", "female", "AMAB" (assigned male at birth), "AFAB" (assigned female at birth); relevant entity IDs such as Q1052281 (transgender female) and Q1097630 (intersex); general terms that may refer to gender-diverse people such as "LGBT", "LGBTQ", "LGBTQIA+", etc.

The extracted pages contained a total of 232,688 discussions. In most cases, however, only a few of the many discussions found on each page were actually about gender. We thus wrote a parser to automatically split the discussions and load them into a database, then filtered these discussions to identify the relevant ones. Through this process, 226,225 discussions were automatically removed because they contained no relevant keywords, 2,569 were excluded because they were not real discussions,[72] and 2,065 were removed because they were not in English. The remaining 1,829 discussions were checked manually by four human annotators through a purpose-built web interface and were classified as follows: 604 relevant, 1,225 not relevant.[73] We classified as relevant all discussions that have gender and its representation as their main topic. We also included those discussions that indirectly

72. These were, for example, many requests for deletion of Wikidata entities that were created by a single user and did not contain any proper discussion.

73. In case of doubt, the annotators followed a consensus-based approach.

refer to the topic of gender identity[74] (with some exceptions; e.g., when they are primarily about non-human gender). The relevant discussions were cleaned from punctuation and non-useful or redundant metadata, then stored in JSON and plain text format.

The discussions collected in *WiGeTa-En* ranged from October 25, 2012, the starting date of the Wikidata project, to September 18, 2021, and they were saved into a database along with their metadata. The corpus contains several metadata related to each discussion: a randomly-assigned *id*, the *start_date* and *end_date* of each discussion, the *users* involved in the discussion, the *wikidata_location* in which the discussion was stored or could be found and, finally, *discussion_title* and *text_discussion*. For further details about the corpus, see Metilli and Paolini (2021).[75]

WiGeTa-En was automatically compiled on Sketch Engine[76] counting 471,890 tokens, 325,956 words, 9786 sentences and 9 documents.

Inside WiGeTa-En: Topic Modeling with LDA

The topic analysis of *WiGeTa-En* was carried out using the unsupervised *Latent Dirichlet Allocation* technique.[77] LDA, which aims to automatically identify and describe latent topics within a collection of text documents, is one of the most frequently used bag-of-words (BOW) probabilistic models for topic modeling.[78] Topics should be understood as a summary of the meaningful

74. For example, there are some discussions in which the main topic is personal names, but there are some references to misconceptions about gender-neutral names.

75. Daniele Metilli and Chiara Paolini, "Non-Binary Gender Identities in Wikidata," WikidataCon, October 30, 2021, https://pretalx.com/wdcon21/talk/7TRCWD/.

76. Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel, "The Sketch Engine: Ten Years On," *Lexicography* 1, no. 1 (2014): 7–36.

77. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022.

78. David M. Blei, "Probabilistic topic models," *Communications of the ACM* 55, no. 4 (2012):77–84.

contents of a collection of documents in which each topic is formed by the most frequent words that characterize that specific content. Latency is an intrinsic characteristic of topics—they do not emerge explicitly but are considered to be hidden, inferred content variables.

The ultimate goal of LDA is to reconstruct a compelling, coherent story from textual data in order to shape and substantiate hypotheses. For an in-depth review of this technique, see Maier et al.[79] The technique has already been applied to discussion analysis[80] to discover relationships between topics, as well as to identify their trends over time and gain insight into target communities. LDA is not able to identify what the discussions are about (beyond a simple set of terms) but, rather, requires interpretation through contextual data.[81]

In our study, LDA allows us to identify clusters of discussions that center around specific topics. We expect the main areas of discussion, identified in our qualitative analysis of the timeline of gender modeling in Wikidata (see section, *A Timeline of Wikidata Gender Modeling*), to be represented in the quantitative data. However, it is also possible for other clusters to emerge, perhaps related to topics that we have not yet considered. Furthermore, LDA allows us to track the emergence of topics over time.

We applied LDA as follows: first, we represented the topics for the entire *WiGeTa-En* corpus; then, for each year represented in the corpus, we created a subset of discussions and performed the topic analysis in order to extrapolate a coherent development of the narrative regarding gender. For the sake of brevity, we will

79. Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, et al., "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology," *Communication Methods and Measures* 12, nos. 2-3 (2018): 93–118.

80. Anton Barua, Stephen W Thomas, and Ahmed E Hassan, "What Are Developers Talking about? An Analysis of Topics and Trends in Stack Overflow," *Empirical Software Engineering* 19, no. 3 (2014): 619–54.

81. As such, LDA results may be biased and should be subjected to scrutiny.

not elaborate on the implementation details of the algorithm, which will be made available in the project's GitHub repository.[82]

Results

In this section, we report the results of the topic modeling based on LDA. Fig. 12 shows the results of the topic modeling for the whole corpus, i.e., all discussions from 2012 to 2021. The figure contains nine charts, each reporting the ten most coherent and frequent words that characterize a specific topic, ranked by their weight; that is, scores generated dynamically based on the weighted distribution of words to reduce the influence of high frequency words and improve the role of keywords. The charts are ordered by the score of the most salient term in the topic, but each chart should be considered independently.

In the figure, we can see that our corpus of discussions clusters around the following topics:[83]

1. *Gender identities*, featuring terms such as "gender", "sex", "transgender", "male", "female", "intersex", and "identity". This is the main cluster of discussion about the topic of gender identity and its relation to sex. It should be noted that the term "non-binary" does not appear with high frequency in this chart,[84] nor does it appear in any of the following charts.
2. *Personal names*, featuring terms such as "names", "female", "male", "gender", "unisex". This cluster of discussions is related to the assignment of gender based on personal names (see the section, *The Rise of Bots*).
3. *Grammatical gender*, containing terms such as "male", "female", "label", "masculine", "feminine", and "occupation". This relates to gendered labels; i.e., entity labels

82. https://github.com/Daniele/non-binary-matters.

83. It should be noted that the names of the topics have been assigned by us, and reflect our own interpretation of the topic modeling results.

84. It is actually in position 24; thus, not displayed in the figure.

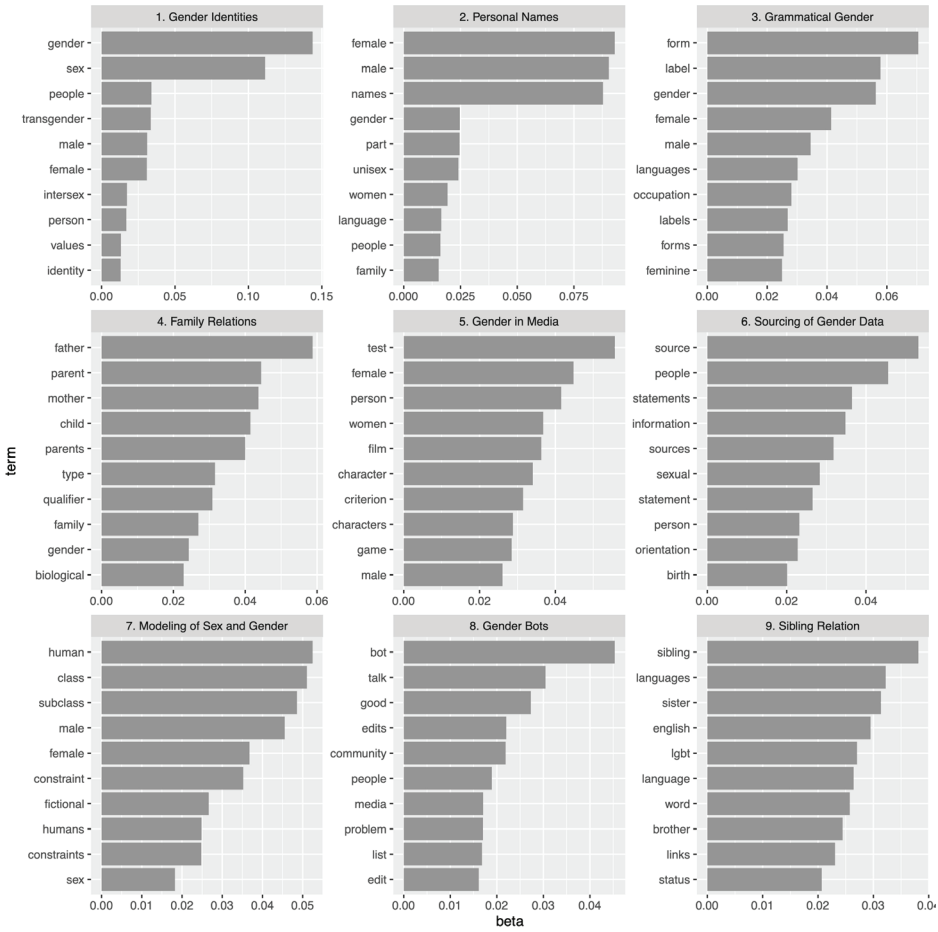

Figure 12: Topic modeling for the years 2012–2021

that have different forms according to the person's gender. We plan to study this topic in more detail in the future.

4. *Family relations*, featuring terms such as "father", "parent", "mother". This cluster of discussions is related to the existence of gendered properties that are not well suited for representing non-binary people (see the section, *Gendered Properties for Family Relations*).

5. *Gender in media*, a cluster of discussions about criteria for gender inclusion in media, such as the Bechdel test.[85] These are not particularly relevant to gender modeling, but it would be interesting to see if these discussions have also evaluated gender diversity.
6. *Sourcing of gender data*, this cluster contains terms such as "source", "statement", "reference", and the discussions featured in it are about the way gender data is sourced. It is interesting to note that these discussions often mentioned sexual orientation as a point of reference (unlike gender, data on sexual orientation has been considered much more carefully by the Wikidata community, and it has not been added indiscriminately to every biographical entity).
7. *Modeling of sex and gender*; this cluster of discussions relates to the modeling of sex and gender, including their taxonomies, and to the application of constraints to the P21 property (see the section, *Ontological Representation of Gender*).
8. *Gender bots*, featuring terms such as "bots" and "edit", but also "problem". These discussions are related to the use of bots to add gender data to the knowledge base and the problems it caused (see the section, *The Rise of Bots*).
9. *Sibling relation*, a specific cluster of discussions about one of the most contentious issues in the modeling of family relations on Wikidata—i.e., whether the *sibling* property should be gendered or not (see the section *Gendered Properties for Family Relations*).

Beside the overall topic modeling, we also performed a year-based topic modeling to assess the topics of discussions for each year. This fine-grained analysis is necessary to understand

85. Alison Bechdel, *The Essential Dykes to Watch Out For* (New York: Houghton Mifflin, 2008).

the shades and implications of the main topics discovered in the overall analysis—namely, the sub-topics tackled and related to the main discussions. For the sake of brevity here, we will focus just on the topics that refer to the lemma *nonbinary*. As shown in fig. 13, the LDA analysis shows that this term appears with a high ranking in only two topic clusters, one in 2018 and one in 2019. However, the frequency is still low when compared to the other terms featured in the cluster.

In 2018, the term *nonbinary* appears in the topic cluster *Gender identities*, while in 2019 it appears in *Grammatical gender*, but in both cases, in a low position compared to other terms. As a point of comparison, we also show that the emergence of a *Trans identities* topic for the year 2019 reflects a significantly increased interest and debate on the representation of trans people's identities.

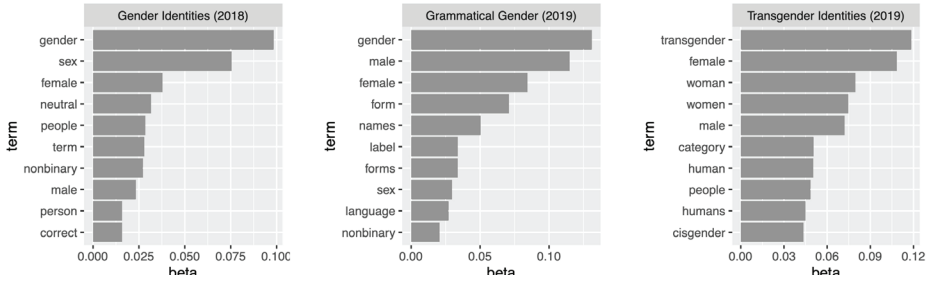

Figure 13: Partial results of topic modeling for the years 2018 (left) and 2019 (right)

Discussion

The overall topic modeling on the corpus (Fig. 12) sheds light on how gender identities are generally discussed by the users who make up the Wikidata community. In most discussion clusters, including those about personal names (topic 2), labeling (topic

3), gender in media (topic 5), and modeling of sex and gender (topic 7), we observe a prevalence of binary terms and a corresponding lack of terms referring to diverse gender identities.

The main cluster of discussions about gender identity (topic 1) features the terms "transgender" and "intersex", but not "non-binary" or any other term referring specifically to a non-binary identity. However, in discussions about family relations (topics 4 and 9), we see prominent usage of gender-neutral terms such as *parent* and *sibling*, in addition to the corresponding binary terms, reflecting the extensive discussions about the topic reported in the section *Gendered Properties for Family Relations*.

Surprisingly, no mention of non-binary identities was found in the overall topic modeling. The low scores of terms referring to non-binary identities show that these identities have been marginalized in the discussions collected in the corpus, providing further support for the findings presented in sections 4 and 5 related to the underrepresentation of non-binary gender identities in Wikidata. In addition, while some of the discussion clusters are related to topics that were already under our scrutiny, there are a few (3, 5, 6) that warrant further analysis. We intend to publish a more complete topic modeling study, with a more in-depth analysis, in a future publication.

The year-based topic modeling (Fig. 13) provides additional insights regarding whether and when discussions on non-binarity have started to be significant and frequent in the Wikidata community. Strikingly, non-binary identities are rarely discussed by the Wikidata users—the lemma *nonbinary* appears only twice in the top positions of the clusters throughout the years and in both cases shows low frequency compared to the other terms. This result suggests that the narrative around non-binary identities has not yet reached the attention it deserves among Wikidata users.

It is interesting to compare this result to the cluster of discussions related to transgender identities that emerges in 2019 (Fig. 13, right), where the term "transgender" is far more central

in the discussions, likely due to the tireless work of LGBTQIA+ movements in favor of trans rights being reported by the media, and also to the cases of famous people coming out as transgender.

Conclusions and Future Work

In this chapter, we have reported a preliminary quantitative analysis of non-binary gender identities in the Wikidata knowledge base. This work has been performed as a first step towards the realization of our project, Wikidata Gender Diversity (WiGeDi), aimed at investigating the issue of gender diversity in Wikidata.

Non-binary gender identities are significantly marginalized in society, and this societal discrimination is often reflected in the way these identities are represented in knowledge bases. The data that are contained in a knowledge base are subject to implicit biases that reflect how the data are sourced, modeled, and published. When these biases are not addressed, they can amplify the discrimination of marginalized communities in society.

Our work aims to contribute to the growing field of data ethics by offering a quantitative exploration of the representation of non-binary identities in a large knowledge base, giving fresh insights to gender studies scholars interested in more qualitative approaches to research.

We have investigated non-binary identities from three different—and complementary—perspectives: first, we have looked at the Wikidata ontology model to understand how it currently represents gender identities and how it has evolved to the representation of non-binary identities. Then, we have reported detailed statistics about the current extent of non-binary representation in the knowledge base, also looking at the distribution of non-binary identities according to several factors (time, country, language, occupation). Finally, we have performed a Latent Dirichlet Allocation topic modeling analysis on the Wikidata community discussions collected in the *WiGeTa-En* corpus.

Taken together, these results suggest that the Wikidata knowledge base is still not fully inclusive of non-binary identities. While some important steps towards recognition of these identities have been made during the years, important issues are yet unresolved. First, the Wikidata gender model is still imperfect and in need of further improvements. Moreover, the representation of non-binary people in the knowledge base is still low when compared to the prevalence of these identities in society, and highly skewed towards the Global North and contemporary times. Finally, the topic modeling analysis suggests that non-binary identities are still significantly marginalized in discussions about gender on Wikidata.

The study that we have presented in this chapter is just a first step in our Wikidata Gender Diversity project. Considerably more work needs to be done to computationally and statistically study gender diversity in Wikidata in a more complete way, analyzing the evolution of the knowledge base over time and the role of the community in shaping the current (and future) modeling of gender. It is also highly likely that significant changes will take place in the future to reflect the evolving views of the community and of society as a whole.

As future work, we plan to extend our linguistic analysis (e.g., by including other community languages), publish a complete timeline about gender modeling in Wikidata, and widen our field of study to include other marginalized identities. We hope that our project will help bring awareness about gender-diverse identities in the Wikidata community and beyond.

Acknowledgements

The authors would like to thank Marta Fioravanti and Beatrice Melis for their invaluable help and support in the design of the project, the annotation of the corpus, and their thoughtful suggestions about this chapter; Elisa Metilli for her crucial proofreading;

Michael Mandiberg for sharing research ideas and providing input on our work; and last but not least, the editors and reviewers for their invaluable advice and recommendations, which allowed us to significantly improve the chapter.