Queer Identities, Normative Databases: Challenges to Capturing Queerness On Wikidata

KATY WEATHINGTON, University of Colorado Boulder, USA
JED R. BRUBAKER, University of Colorado Boulder, USA

The collection, organization, and retrieval of data about queer individuals and their identities challenge the creators and curators of highly structured database systems. While prior research in archival studies and demographics has examined processes of collecting and storing queer identities, they do not examine the complexities of highly democratized platforms that lack top-down mandates that often structure archival schemas. To examine the representation of queer people on open-platform databases, we performed a trace ethnography and thematic analysis of Wikidata, an open collaboration, highly structured database. We specifically examined the creation of, changes to, discussions around, and impacts of properties that encode queer identities, such as *sexual orientation* and *sex or gender*. We found that changes often have unexpected impacts, that contributors struggled to determine vocabulary for queer identities which were accurate across the diverse cultural contexts of the Wikidata community, that the recording of queer identities could cause a stigmatizing effect for LGBTQ+ individuals, with further concerns of spreading rumors or outing closeted people, and that contributors proposing changes which would cause biased representations of queer people. Our analysis demonstrates inherent and unaddressed frictions when translating queer identities to the confines of a structured database. We conclude by discussing ways that the highly bottom-up, collaborative nature of platforms such as Wikidata, often seen as a major strength, can be vulnerable to individuals or small groups derailing and filibustering changes they disagree with on politically charged topics such as queer identities.

CCS Concepts: • **Human-centered computing** → *Wikis*; • **Social and professional topics** → **Gender**; **Sexual orientation**.

Additional Key Words and Phrases: LGBTQ+, gender, sexuality, identity, collaborative ontology, database, Wikidata

ACM Reference Format:
Katy Weathington and Jed R. Brubaker. 2023. Queer Identities, Normative Databases: Challenges to Capturing Queerness On Wikidata. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 84 (April 2023), 26 pages. https://doi.org/10.1145/3579517

1 INTRODUCTION

On December 1, 2020, actor and activist Elliot Page came out as trans. The announcement via Instagram [46] explained that Page would be using he/him and they/them pronouns and was greeted with millions of likes and hundreds of thousands of reposts and supportive replies. Page's transition also prompted numerous articles in popular press documenting the event [23, 31, 57, 60], as well as publishers to revisit and update older articles. Page had previously openly identified as a lesbian woman, and as a result, media companies and publications revisited past content to update Page's name and pronouns. For instance, Netflix changed the actor's name in the credits of their original content such as Umbrella Academy [70]. Announcements like these are occasions

Authors' addresses: Katy Weathington, University of Colorado Boulder, Boulder, CO, USA, katy.weathington@colorado.edu; Jed R. Brubaker, University of Colorado Boulder, Boulder, CO, USA, jed.brubaker@colorado.edu.

2573-0142/2023/4-ART84
https://doi.org/10.1145/3579517

for updating and reappraising collaboratively-maintained knowledge entries. Page's Wikipedia article was marked for update to reflect his identity only nine minutes after his original Instagram post, and a flurry of edits followed. However, Wikipedia's less-known sister site, Wikidata, was undergoing a similar burst of activity in response to Page's coming-out post.

In this paper, we turn to Wikidata, a peer-produced database and sister project to Wikipedia, to examine tensions in the datafication of gender and sexuality[1]. While Wikipedia is an encyclopedia using context-rich prose to contain information, Wikidata is a database using highly structured relational statements to encode and classify information. In contrast to other structured databases, often developed by a singular entity, Wikidata's open discourse around its development offers a window into the decision-making process behind biographical information, including information about queer individuals. Queer identities are recorded in a database in clunky, inaccurate, or self-contradictory ways.

To better understand the tensions between database infrastructures and data about queer identities, we performed a trace ethnography [21] of queer-related biographical properties in order to understand how the underlying database codifies and operationalizes queer identity attributes. We specifically traced the development of *sexual orientation* and *sex or gender*, examining community discussions, logs of changes over time, and interdependencies between these properties and other parts of the Wikidata platform. Our analysis made use of both close-examinations of properties and specific pages, as well as quantitative analyses to understand impacts across the platform.

Our analysis highlights challenges in representing queer identities on Wikidata. First, naming conventions and terminology presented both social and technical challenges. Socially, contributors from diverse backgrounds and with different views and understandings of queerness struggle to agree on shared terminology. Technically, we describe how evolving and non-standardized labeling practices for queer identities present a technical challenge for the veracity of data and the shape of the database as a whole. Contributors additionally struggled with how the recording of queer identities could result in the outing and stigmatizing of queer people, as well as concerns over unverified statements of identity. Finally, we found extensive evidence where solutions to the above were complicated by contributors pushing for changes that would result in biased and problematic representations of queer people.

Based on our findings, we discuss the limitations that the open collaborative nature of projects like Wikidata have when facing polarizing topics such as queer identities, which could be mitigated by the addition of some top-down interventions. Furthermore, we argue that there are unaddressed tensions between the standardized, universal, archival needs of database systems and the non-normative fluidity of queer identities.

2 RELATED WORK

2.1 Queerness and HCI Research

Previous HCI research on LGBTQ+ communities has primarily examined the needs and experiences of queer individuals and communities in a variety of digitally mediated contexts. The focus of this

[1]In this work, we use the term "queer" as an umbrella term for identities related to sexuality or gender which largely challenge traditional cultural norms [29]. We use the term "queer identities" to refer to concrete instances of such identities, also commonly referred to by some variation on the acronym LGBTQ [37]. In contrast, we use cishetero to refer to non-queer people, that is, cisgender and heterosexual people. Furthermore, we use 'gender' to refer to an individual's culturally situated, yet deeply felt identity. We use 'sex' or 'biological sex' to refer to an assumed biological status based upon the interpretation of genitals observed at birth as is used in common parlance. We recognize that biological sex is an inaccurate and outdated concept that fails to account for the many possible variations between genes, primary, and secondary sex characteristics which are commonly associated with cisnormative (that is, the viewing of and privileging cisgender as the norm) concepts of sex.

scholarship is diverse. For example, Blackwell et al. examined how the context collapse associated with location-based queer hookup apps complicates safe self-expression [12]. Haimson et al. examined how social media platforms such as Facebook serve as the site of major coming out events, as well as the stress and support which accompany such events [27]. Scheuerman et al. examined the perceptions of digital interactions by trans and non-binary individuals, finding violations that go beyond merely reflecting largely societal patterns [58]. Finally, Hardy has developed a framework of "queer information literacy," describing the process by which varying levels of access to queer information and community create distinctly different LGBTQ+ experiences [30].

Often distinct from research on LGBTQ+ issues, HCI scholars have also used queer theory as an anti-normative critical lens. Light used queer theory as the basis for resisting political status quos through embracing practices of contrariness, such as forgetting, obscuring, cheating, and eluding, rather than designing for efficiency and effectiveness [41]. Queer theory (along with feminist theory) has also been used to critique binary conceptualizations of gender in HCI [15].

While the majority of research in HCI is focused on direct user experiences, much of these experiences are shaped by the underlying data infrastructure. As such, in this paper, we attend to how queer people are made the subject of data [9] and the tensions that can arise.

2.2 Knowledge Organization

Given the importance of data creation and organization in our study, we turn to literature from knowledge organization and archival studies to inform our investigation. Below we consider challenges in classifications and data ontology, and how queer archives have responded to these challenges in the context of LGBTQ+ experiences. Finally, we survey literature on participatory archives and databases.

2.2.1 Classification and Control. Representing queer identities presents a long-standing challenge for knowledge organization systems from libraries to archives. Notably, the classifications used to organize data in these systems rely on controlled vocabularies [17] which can encounter problems given the shifting and inconsistent labeling schema used to describe queer people.

Classification is an ever-present if often invisible part of daily human life, based on systems of classification with varying levels of structure and consistency. These classifications are dynamically changing, and shape how we view and interact with the world around us [13]. Many modern categorical definitions of queer identities were defined so that sexually diverse individuals could be defined, othered, and ultimately dominated [20].

Classification schemes used prior to the era of digital information retrieval, such as controlled vocabularies, cannot be used for any items which are not explicitly defined, making them unable to adapt and function efficiently with heterogeneous digital data [52]. To this end, data ontologies were developed. A data ontology refers to the structures and rules that describe what can exist within a database (both what currently exists and what may be added in the future) through the definition of entity classes, their behaviors, and types of connections between items [25]. An ontology's entities abstract away the specific verbiage found in controlled vocabularies to create generalized relationship frameworks which can be applied with more specific information to any set of conceptualizations within the ontological space [26]. When created with computation in mind, the consistent structure of ontologies results in data that is more computer-readable than traditional classification schema, let alone unstructured data [24].

2.2.2 Queer Archives. Concerns about the lack of inclusion of queer individuals in archives have prompted a number of queer-specific archives and archival projects [54]. Historically, these efforts find their roots in the queer liberation movement following the Stonewall Riots of 1969 and have increased in size and diversity as the queer community has grown [54].

While initially considering only homosexuals (which, at the time vaguely included most non-cishetero people), other identities such as trans people were later distinguished and have developed archives of their own [54]. Many of these archives, such as the Lesbian Herstory Archive (LHA), are highly independent projects, being created, managed, maintained, and funded largely by the slice of the queer community they preserve [16]. However, even when queer people are included in archives, scholars have described two specific challenges that queerness presents.

Time itself is a challenge for archival projects around queerness. Over the half-century history of queer archives, a new contention has arisen. Not only have commonly used labels for queer identities changed, so too have the identities they label [66]. While it is tempting to simply use modern labels and their current classifications for past individuals, forcing modern labels and classifications onto historical people may contradict the highly contextualized experience of gender, sexuality, and the self in which these historical peoples were situated [66]. For example, the modern term transgender is often applied to historical people who, at the time, identified as transvestites or transsexuals. In doing so, they would erase certain nuances, such as that transsexuals pursued surgical transition, and that transvestites often lived as their assigned gender in daily life.

In addition, many trans individuals specifically do not want to preserve information about life pre-transition, conflicting with archival standards of completeness [54].

Beyond temporality, queer archivists have discussed the ways that material archives disembody experiences. While not unique to the LGBTQ+ community, the trans experience, for example, is highly embodied and loses significant context with the disembodiment of information inherent in archiving [54]. Indeed, Cvetkovich argues that an adequate archive of queer experience must capture the intense emotional states of queer love, rage, activism, and pride, which get lost in traditional material archives [16]. Finally, Rawson has argued that more than the rest of the queer community, trans people have historically lacked access to archives, and thus have been left out of the rhetorical power structures created by such archives [53, 55]. While projects like Wikidata may offer greater access, in our findings we demonstrate that this potential, if it exists, has yet to have been realized.

2.2.3 Community and Participatory Archives. In some cases, the classification systems used to organize data are produced from the bottom-up rather than from the top-down. Such participatory archives and databases develop as a result of community discussion and debate. HCI research into collaborative knowledge organization covers two main areas: folksonomies and collaborative ontologies.

Community-developed information organization and classification schemes exist in several forms. While rarely thought of as such, tagging systems such as those found on popular social media sites are one of the most common instances of folksonomies and community classification [48]. Such tagging systems allow users to develop a system of information organization from the ground up [65].

Wikidata falls into the second area of research: collaborative classification schemes and ontologies. Wikipedia's contributor-defined hierarchical page categories have been used by computer scientists as the basis for classification systems, primarily employed for information retrieval [64]. However, these categories lack any formal definition of category relationships, making them difficult to translate into a rich relational database or knowledge graph [68]. While Wikidata may provide a more structured ontology, Piscopo et al. found Wikidata's ontology to have inconsistent breadth and depth for different subject matter as a result of varying levels of topic interest amongst the volunteer community [51].

2.3 Queer Data

As queer identities have become more socially accepted and more queer people are out, there has been an increasing effort amongst census takers of all stripes to collect demographics about this previously overlooked community [10]. However, traditional demographic collection methods are ill-equipped to collect data about queer identities, failing to account for identities changing over time or an individual having multiple, coexisting identities, and with some labels meaning different things to different individuals [8, 56]. Such missing or inaccurate demographic data hampers algorithmic fairness initiatives [8]. Often, sex and gender are recorded in a single survey question with a larger focus on sex (biological or legal) than gender, which both reduces response rates of and excludes gender minorities. The common remedy is to add more response categories beyond merely male/female/other [42, 62]. Some researchers recommend splitting a singular sex/gender question into two separate questions: one asking for sex assigned at birth and another asking about current gender identity and inferring trans status from any discrepancies between the two [10, 62]. However, both of those solutions still treat gender identity data as categorical rather than along a spectrum, limiting the ability to self-report gender and doing little to understand complicated gender expression in real life [11, 42]. Furthermore, intersectional approaches to recording multiple marginalized identities (*e.g.*, treating the experience of being a queer woman as more than just the sum of being queer and being a woman) is more accurate and productive than the conventional additive approach; however, this creates challenges for traditional data collection and analysis methods and is uncommon [14].

3 WIKIDATA BACKGROUND

Though its sister site, Wikipedia, is well known, fewer people are aware of and familiar with Wikidata. In addition to discussing the literature about Wikidata, this section will provide an overview of the purpose, function, and structure of Wikidata.

3.1 Knowledge Graphs

The majority of scientific literature about Wikidata is limited to its capabilities as a knowledge graph. A knowledge graph is a type of database where information is stored as a network of entities defined not only by their own attributes but also by the connections with other entities and overall placement in the scheme [18]. In contrast, Wikipedia research is far more diverse and expansive. Past ethnographic research has studied community conflict resolution on Wikipedia, finding that ideal consensus making is rarely achieved, and debates often become a battle of attrition to wear down opposition [32]. However, such community dynamics on Wikidata are less studied than those on Wikipedia, and it is unknown to what degree, if any, these findings may apply to Wikidata. Despite a high degree of multilingualism amongst Wikidata users, English is the default language, most commonly spoken to some degree, with English Wikidata having the most completed entries, and the majority of discussion posts are in English [35, 36]. Despite a large user base, almost all manual edits (95%) are made by very few users (2%), giving it a core-periphery structure found in many open production communities [45].

There is a long history of queer Wikimedians (that is, someone who contributes to any of the many projects under the Wikimedia umbrella) making major contributions to various Wikimedia projects, including but not limited to Wikipedia and Wikidata, and proudly participating with the community [69]. Furthermore, the Wikimedia Foundation created initiatives specifically to promote queer content and participation; WikiProject LGBT is one such project made up of volunteers dedicated to promoting the inclusion and treatment of LGBT+ content on Wikidata, and members

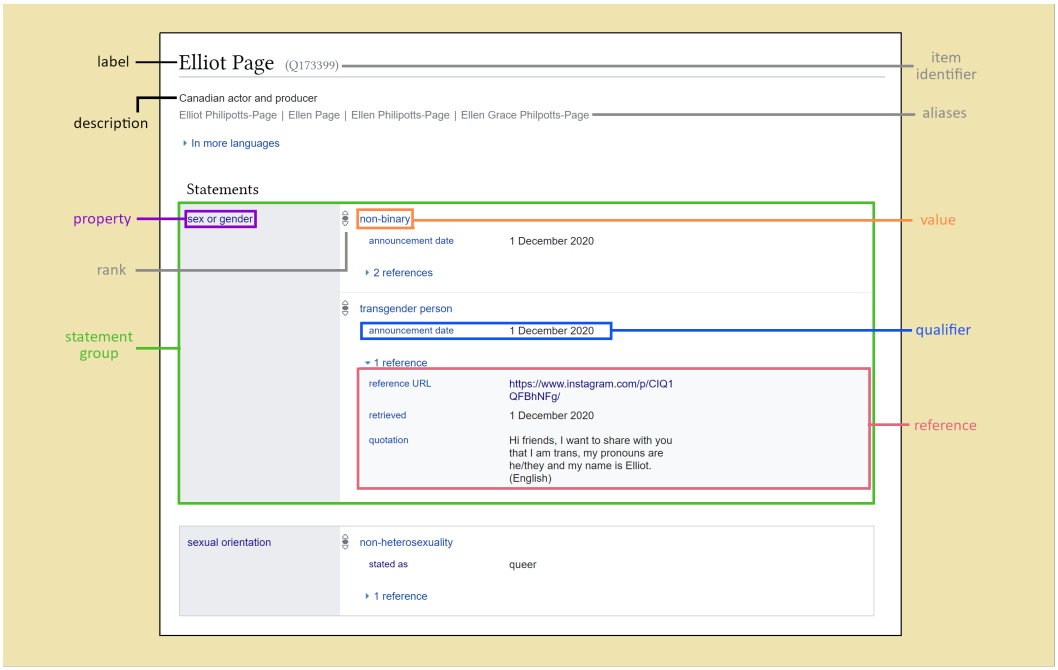

Fig. 1. An example of a Wikidata item page with annotations.

often engage in discussions related to the LGBTQ+ community as well as refer to norms and practices preferred by this group [6].

3.2 What Wikidata is

Wikidata is an open, collaborative, highly computable universal database containing information on any notable topic[2]. It describes itself as a "free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world" [4]. Since Wikidata was founded in 2012, it has grown to over 90 million content pages with 1.38 billion total page edits [3]. Wikidata has 4.8 million registered users, with more than 26,000 monthly active users editing an average of 481,000 pages every day in 2021 [3, 7].

As previously mentioned, Wikidata is best characterized as a knowledge graph [49, 50]. Wikidata leans heavily into this structure, with almost all information about a given entity being gleaned from its connections (via properties) to other entities on Wikidata. Besides an entry's label, aliases, and occasionally a brief prose description of the item's notability, the entries themselves contain no information.

The complex, web-like structure of Wikidata is highly computable and designed to be queried. By following this web of properties pointing to values, a computer can efficiently access information in response to human questions. Because of this structure, Wikidata is deeply imbricated within larger knowledge-intensive socio-technical infrastructures like machine translation and voice assistants far beyond Wikipedia or Wikimedia projects [5, 61]. When asking Alexa a simple question, for

[2]Wikidata defines notable topics as anything for which there already exists an entry on another Wikimedia site, provides a structural need, or is a clearly identifiable conceptual or material entity. Wikidata's full notability criteria can be found at https://www.wikidata.org/wiki/Wikidata:Notability

instance, where was Elliot Page born, Alexa first identifies the parts of the question (Page, place of birth), then goes to Page's Wikidata entry, looks for a place of birth claim, and returns whatever the value of this claim is, in this case, Halifax. The end-user would receive an answer along the lines of "Elliot Page was born in Halifax."

3.3 How Wikidata Works

Wikidata is a structured database where information is organized as a collection of connections between query-able documents that Wikidata refers to as "items." Each item is allocated a unique identifier consisting of a "Q" and a positive integer, known as a "QID."[3] Items can represent topics, concepts, or objects such as *Elliot Page* (Q173399), a profession like *actor* (Q33999), a film like *Whip It* (Q303213), and gender identities such as *non-binary* (Q48270). Though there is no technical distinction, items can be conceptually divided into instances, singular concrete cases, and classes, broader categories utilized in imparting information about instances [45]. For instance, *Canada* (Q16) is a specific instance of the more general class *country* (Q6256).

These items are connected through "properties," similarly denoted with P value designations (*e.g.*, P91 for *sexual orientation*). On a given item page, a property will point to another item page, in this context called a value (*e.g.*, the P91 properties might point to the Q43200 value for *bisexuality*). The property will give context about how or why a specific item relates to another item. Together, the property and the value it points to are called a "claim" or "statement." When there are multiple claims for the same property on the same item, this is called a "statement group." As a concrete example, to indicate that Elliot Page is non-binary, Wikidata says "Q173399 (*Elliot Page*) has a Q48270 (*non-binary*) P21 (*sex or gender*)" claim. P21 is describing the nature of the relationship between Elliot Page and non-binary, specifying that he has a non-binary gender.

Properties are structured with "constraints" that limit how a given property and its statements can be used. Constraints are implemented as property-value claims similar to what we see on item pages. The most relevant constraints to our study of P21 and P91 are "one-of" constraints, "type" constraints limiting claims to people (real or fictional), and a "citation needed" requirement for P91. One-of constraints limit allowed values, and by extension identities, to a pre-determined list. Type constraints indicate which category of values a claim can be made on (in this case, limited to people). Finally, citation needed constraints (as found on P91) indicate that any claim made with this property requires an external source, meaning any claims with this constraint must be able to point to some public record or statement (e.g., the subject tweeting a coming out post).

3.4 Working on Wikidata

The different types of pages have restrictions on who can create them. The least restricted are item pages, as anyone can create a new item even without a Wikidata account, though an IP address will be displayed publicly if not logged in. For creating a new item with a new QID, there is a link on the sidebar of the website leading to a special page containing instructions that remind the user to comply with notability guidelines and check if the item already exists, as well as a form that allows the user to choose a language, and enter a label, description, and aliases. Other than the language, none of these fields has to be filled out to submit the form.

The capability to create new properties is far more restricted than for items. Only the 62 admins and 40 additional property creators have access to this link [3]. Of the nearly 5 million users, only 102 are allowed to make properties. So how can any of the other users request a new property? They must post a proposal for a property for the community to debate, creating a public record explaining both proposed schema and use cases, to which other contributors reply with positive or

[3]The "Q" in QID was chosen by Wikidata founder Denny Vrandečić as it is the first letter of his wife's name, Qamarniso.

negative responses. If a suitable consensus is reached, left to the sole discretion of the property creator, they then fill out a form similar to how items are created. Once the form is submitted, the property creator is then expected to add relevant constraints from the property proposal discussion and update the proposal to indicate acceptance.

Barring an edit lock, where an administrator or user with extra privileges restricts who can edit a page in response to heightened vandalism or excess edit warring, anyone can edit a page, whether it be an item or property. On any given page, there are many "Edit" buttons, one for each claim made and an additional one for the description, label, and aliases. Unlike on Wikipedia, where editing largely consists of editing or writing additional prose, edits here are highly structured, with specific fields for each allowed piece of information. There are also "Add Value" and "Add Statement" buttons which once again give you a highly structured set of boxes to fill out.

Discussions take place under a separate tab, and anyone can add a new topic or reply to other topics. Discussion threads are basic forum posts. The topics themselves can be asking for advice about handling a specific case, asking clarifying questions about item or property structures, or as is most relevant to our work, calling into question the entire meaning and purpose of the entry.

4 OUR APPROACH

Our analysis can be divided into two types: trace ethnography and quantitative analysis. To support our analysis, our dataset consisted of all documentation on Wikidata related to sexual orientation and gender properties, including the entities themselves (P21: *sex or gender*, P91: *sexual orientation*); their revision history; the proposal for the creation of P91; all discussion posts related to these properties, both within the properties discussion pages and on other pages where debates occurred; as well as the proposal process for gender, a property that failed to be adopted. Furthermore, we quantified and analyzed the usage of these properties on hundreds of pages of queer individuals and thousands of non-queer individuals, with a closer inspection of the usage of deadnames as aliases or other data points and the effect of *sexual orientation* and *sex or gender* on dozens of pages.

4.1 Trace Ethnography

To better understand the evolving nature of sexuality and gender on Wikidata, we performed a trace ethnography [22] of P21 (*sex or gender*) and P91 (*sexual orientation*) as well as relevant discussion pages, proposals, and other traces described above. Our trace ethnography focused on the content of and changes over time to these properties, as well as the community discussion posts around these properties. We not only gathered discussion posts on changes to these properties, but also discussion posts about the creation and calls for the deletion of P91. We performed an inductive thematic analysis on the issues which arose in these discussion posts, performing multiple analytical passes. To bolster our understanding of changes over time, we tracked the edit history of P21 to follow changes within its schema throughout its existence, especially focusing on changes to the balance between its role in modeling gender *and* sex. The majority of discussion posts were collected during the initial analysis in the Fall of 2020, though new discussion posts were incorporated as they arose throughout the study and writing period. The specific case of Elliot Page was followed as it developed since his coming out on December 1st, 2020. We found four main themes in our analysis: "discrepancies between expected use and actual use," "difficulties with queer-related vocabulary," "concerns about data subject privacy and outting," and "neutrality and bias in community discussions."

4.2 Quantitative Analyses

We also used the MediaWiki API to retrieve digital trace data about the revision history[4] of the P21 (*sex or gender*) and P91 (*sexual orientation*) properties and the Q173399 (*Elliot Page*) item as well as their corresponding talk/discussion pages. Using a custom Python script, we then computed the cumulative revision activity (number of changes) over time and generated time series visualizations to help contextualize the discussion of qualitative results, and highlight how discussions and revisions tend to occur in relatively short, intense bursts.

Furthermore, to understand the how the actual use of these properties manifested, we used Wikidata's native query language, SPARQL, to obtain demographic information, especially values for P21 and P91 where present, for all human entities with a date of birth between January 1, 1950, and December 31, 2019. We focused on people who were born between the years 1950 and 2019 in order to capture the time period in which the majority of living people on Wikidata were born, avoid historical figures, and to focus on people who lived during the modern LGBTQ+ movement. We then visualized the prevalence of different values for P21 and P91 in this dataset.

5 FINDINGS

Our analysis highlights tensions between data schemas, the communities that develop and maintain them, and the lives these schemas are intended to represent. We detail these tensions across our findings. Many of these issues can be seen on Elliot Page's Wikidata entry (Q173399) following his coming out, and as such we will use him as a recurring example throughout our findings.

Prior to coming out on December 1st, 2020, Page's entry had a *sexual orientation* (P91) value of *homosexuality* (Q6636) and a *sex or gender* (P21) value of *female* (Q6581072) — values that reflected Page's previous stated sexual orientation and assumed cisgender identity. Yet following his announcement of his gender identity, these previous values became wrong or unsubstantiated in an instant. So how was this case of a queer person, with a deeply held yet not entirely explicitly defined gender identity and sexual orientation, handled by the database structure, use policy, and community practices within Wikidata?

The community quickly updated Page's entry. Eighteen minutes after his initial Instagram post, Wikidata contributors had updated his name and *sex or gender* to *transgender male* (Q2449503). Within the hour, editors would also add completely new *personal pronouns* (P6553) claims for Page, making separate claims for both *he* (L485) and *they* (L371)[5].

Despite quick initial progress, Page's entry was in flux with over 500 edits in the following week (in contrast, there were 4 edits during the preceding November). Although many edits were the result of propagating changes across different language versions, a notable portion was the result of edit wars [1] around Page's P21 and P91 statements. An unsurprising amount of these edits came from vandals who made changes that either reverted claims or labels to Page's pre-coming out state or to an intentionally ridiculous value. However, the majority of these edits were the result of a community collectively working to represent Page's identity in a respectful way within the schema. The changes resulted from disagreements on both what Page's identity actually was and with what words to describe it. Contributors held conflicting positions on how to represent Page's identity, resulting in several of Page's properties vacillating between different claims.

Below we present four themes that emerged in our analysis: discrepancies between expected use and actual use, difficulties with queer-related vocabulary, concerns about data subject privacy and outting, and neutrality and bias in community discussions. We start with a technical focus, examining the impacts of schema, and progress through findings that increasingly highlight social

[4] https://www.mediawiki.org/wiki/API:Revisions

[5] Words themselves are stored as lexemes on Wikidata, denoted with an L and positive integer

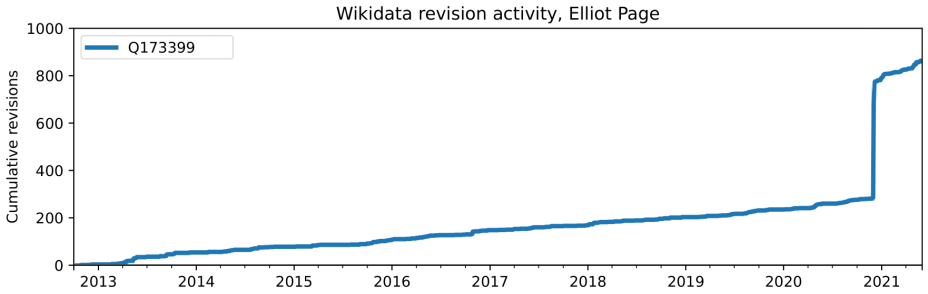

Fig. 2. Cumulative revision activity for Elliot Page's Wikidata entry.

tensions. Throughout we return to the case of Page to ground our findings and to demonstrate how the sociotechnical tensions we identified play out across the platform as a whole.

5.1 Discrepancies Between Expected and Actual Use

Wikidata is designed to store structured data about items and topics as diverse as the articles on Wikipedia. However, we identified numerous issues when representing queer people that were the result of discrepancies between how the Wikidata community imagined and designed the platform to be used versus actual observed use.

5.1.1 Failure to Reflect Gender Changes in Sexual Orientation. After his coming out post, contributors started changing claims on Page's entry. While some had already gone to editing, with many overwriting each other, a discussion thread was started to debate the specifics of Page's gender. The central question was whether to use one or more of transgender male, transgender person, or non-binary. Furthermore, contributors debated how to treat the existing *female* P21 claim, variously suggesting removing it entirely, deprecating the claim, or adding an end date qualifier but otherwise leaving the claim as is. Deprecation would leave the prior value up while indicating that the previous value is now known to be wrong but must be maintained for the integrity of the database. On the other hand, using an end date qualifier would imply that Page had indeed been a female up to the date of his coming out post and then became a man.

Page's announcement also presented a challenge for sexual orientation. Wikidata policy requires that any assignment of sexuality (i.e., a P91 claim) to a living person be accompanied by a citation for an explicit statement from the subject about their sexuality. Common citations point to interviews, public statements, autobiographical work, and, as was the case with Elliot Page, social media posts. The community intended to prevent 'outing' individuals and to make sure that any identity values align with how the individual thinks of their identity. The community created this requirement when proposing P91, hoping to avoid speculation or rumors becoming part of the data. Technically, the citation requirement is embedded into the structure of the system by giving *sexual orientation* a "citation needed constraint," which makes any statements for P91 invalid unless accompanied by a statement of sexuality from the subject (if living) or widespread consensus by historians (if dead).

While Page previously had a *sexual orientation* of *homosexuality* on Wikidata, implying lesbian, the accuracy of this label was now called into question. Some contributors changed the P91 claim to *heterosexuality*, presumably based on previously stated woman-loving-woman identities and the newly revealed masculine gender identity. However, nowhere in his coming-out announcement did Page ever call himself heterosexual, nor mention any other specific sexuality, which severely limited possible claims that would satisfy Wikidata's citation requirements for P91 claims. Despite

no discussions, much less consensus, contributors repeatedly altered Page's *sexual orientation*, leaving it in a state of near-constant flux. During the first 4 days, changes to *sexual orientation* did not last long. Changes were quickly reverted to the former value or replaced with a new one. Eventually, Page's P91 value stabilized to *non-heterosexuality* (Q339014) *stated as "queer"*. The *stated as*qualifier refers to an ambiguous use of the term queer in Page's coming out post[6], which in context could have equally been used in reference to the non-binary aspect of his identity or general membership in the broader LGBTQ+ community. However, stretching the interpretation of "queer" to a statement of sexuality allowed for a P91 claim which satisfied the source requirement.

Page's case demonstrates overlooked dependencies between properties in Wikidata. Specifically, the technical implementation of *sex or gender* and *sexual orientation* claims do not account for possible inter-dependencies of these properties.

Sexual orientation, at least in Western society, is typically bound up with gender. Being attracted to women is only considered straight if you are a man. Calling a man who is attracted to women a lesbian would seem to contradict their male gender identity. Yet on Wikidata, there are no structural ties or constraints between P21 and P91. On an ontological level, they are independent. Any relationship, therefore, is purely the result of contributors having a pre-existing contextual understanding of gender and sexuality which they apply on a per-person basis.

Citation requirements for *sexual orientation* (P91) create an additional technical challenge to contributors reacting to *sex or gender* (P21) changes. The strict citation requirement on P91 requires that claims include citations to public accounts in which the subject self-identifies with said sexual orientation. The only flexibility given to editors is in using synonymous terms for identities expressed in the source (for instance, someone who used the term lesbian may be recorded with a *sexual orientation* of *homosexuality*). There is no leniency to change P91 values to reflect changes with P21 in the current state of these properties. To change *sexual orientation* to coincide with *sex or gender* requires the subject to disclose their sexuality in addition to their gender transition. We see a clear case here wherein the relationship between two concepts, intertwined in the real world, lacks any analog in Wikidata. Any time a contributor changes an item's *sex or gender*, there is the possibility for continued use of a now-defunct *sexual orientation*, either from being overlooked or formalized requirements making an update outright impossible.

5.1.2 P91 as a Queer-Only Property. We also found discrepancies in how singular properties were used in practice. Consider *sexual orientation* (P91).

As mentioned previously, Wikidata policy requires that any assignment of sexuality (i.e., a P91 claim) to a living person be accompanied by a citation for an explicit statement from the subject about their sexuality. There is, however, an unintended consequence of the citation requirement that calls into question the accuracy and utility of sexual orientation data: Straight cisgender people do not "come out" and therefore infrequently produce statements that would meet the citation requirement. Suitable citation material remains essentially exclusive to queer individuals. As a result, P91 statements exist overwhelmingly for entries about queer people, making heterosexuality an invisible default for society, but also largely absent from Wikidata. One would rightly expect to see P91 claims that reflect the sexual orientations of the broader population. A recent Gallup survey found that 86.7% of the population identifies as heterosexual, while 5.6% identified as LGBT, and only 7.6% of respondents did not answer the question at all [33]. However, sexual orientation is largely absent from Wikidata, with less than 0.1% of our dataset having a P91 claim at all. Of those with a claim, fewer than 3% were heterosexual.

While the P91 citation requirement is intended to protect people, we see that it is actually singling out individuals with one or more queer identities. Having to specifically state your sexuality is

[6] "I love that I am trans. And I love that I am queer"

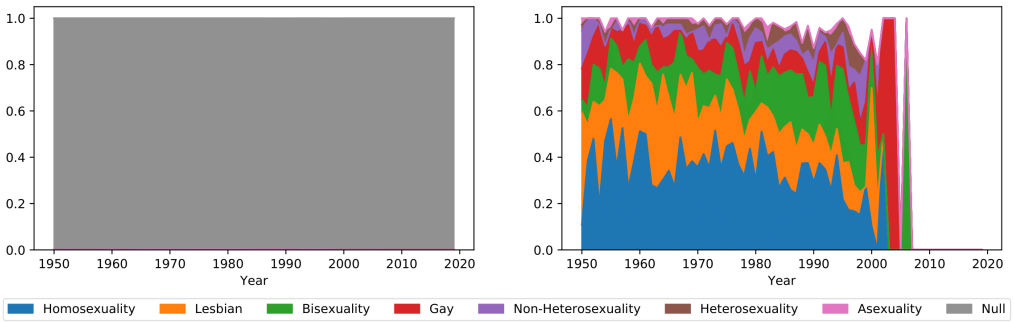

Fig. 3. Timeline of Wikidata biographies by birth year showing proportions of P91 claims by value. All biographies (including a majority with no P91 claims) are shown on the left, demonstrating that the vast majority of biographies do not include sexuality. The right panel excludes biographies with no P91 claims, demonstrating that the vast majority of P91 claims are for entries about queer people.

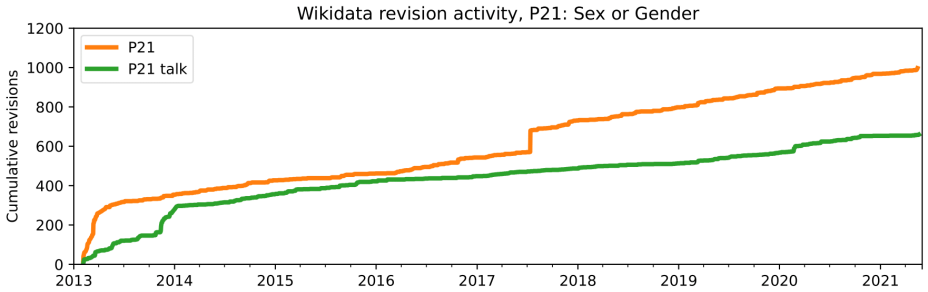

Fig. 4. Cumulative revision activity for the P21 property and property talk pages.

a burden almost exclusively for queer individuals. Thus, heterosexual values in P91 claims are disproportionately uncommon. Because (generally) only queer individuals state their sexuality, they are the only people who the citation requirement allows to have a *sexual orientation* statement. Thus there are two unintended outcomes of the citation constraint: it singles out sexual minorities while also resulting in non-representative data.

While different in both scale and effect, both issues described in this section highlight ways that technical structures and policies have had unintended and unexpected consequences on Wikidata. While the broad repercussions of such changes may have been guessed (*e.g.* users predicting P91 would stigmatize queer people), the actual mechanics and form of unintended outcomes were unforeseen and not prepared for. Whether wrestling with the systemic marginalization of introducing sexuality to Wikidata's ontology, or blindly navigating the nuances of the complicated relationship between gender and sexuality on any given entry, the lack of foresight left contributors with neither the flexibility nor guidance to rectify issues if and when they arose. Each new issue would need to be discussed and handled after the fact, forcing contributors to act reactively rather than proactively.

P91 is one example of how the Wikidata community struggles to account for how certain societal norms and assumptions around identities will manifest in the data, let alone how to address such

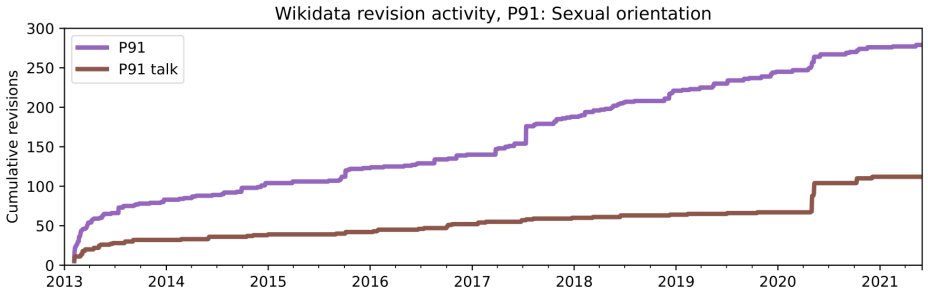

Fig. 5. Cumulative revision activity for the P91 property and property talk pages.

instances once they do occur. Understanding what norms are and how they behave in a community as diverse and global as Wikidata is a challenge in and of itself. Predicting precisely how these norms will play out in the database itself is incredibly challenging, so contributors cannot reliably know how the system they built will manifest in actual use. However, without significant forward-thinking to predict likely outcomes and develop plans for adapting to whatever may happen, whatever issues may arise cannot be addressed with the immediacy necessary to prevent negative consequences.

5.2 Troublesome Vocabulary

During our study, we encountered many cases of vocabulary struggling to address the complexities of queer identities as well as cisnormative values being projected onto a queer individual through the use of certain language. Issues our analysis identified include inconsistent delineations between sex, gender, and several variations on binary category labels to state sex and/or gender, with different contextual meanings for the same words.

With regards to his sexuality, Page's entry currently has a *sexual orientation* of 'non-heterosexuality stated as queer' referencing Page's coming out post wherein he states "I love that I am trans. And I love that I am queer" [46]. While it is possible to interpret 'queer' as a sexuality here, the context more likely implies that Page was referring to a non-binary aspect of his gender identity rather than his sexuality. Interpreting the term queer to be in reference to sexuality is effective in satisfying P91's sourcing requirement, and creates a functional resolution to redefining Page's sexual orientation.

There are significant variations in the terminology used to describe someone's gender or sexuality. As Zolo, a reoccurring figure in discussions around P21, stated:

> Even the meaning of the word ["Gender"] is not entirely clear. Wikipedia defines it as a 'person's private sense, and subjective experience, of their own gender' (and if so how do we know?)

Complicating the struggle around terminology on Wikidata is the sensitivity around talking about queer identities. Less knowledgeable individuals may speak with the most broadly acceptable language to avoid causing offense, and more knowledgeable individuals may argue over the smallest nuances of what is the most accurate, socially understood, and inclusive language for describing someone. These language battles reflect part of an ever-evolving cultural understanding of gender and sexuality.

Furthermore, there is a balance issue with regards to the standardization needs of a highly computable database wherein there is a need for relatively few categories that encompass the

majority of cases which inherently reduces the capability for nuance in the language used to label items. Perhaps the best example of the struggles around vocabulary can be seen in the debates around the definition of P21, including its label and associated values. The Wikidata community struggled with what P21 should represent, as well as what words to use, all because the diverse Wikidata contributors had inconsistent conceptualizations of gender, sex, and the combination of the two.

When first created on February 4th, 2013, P21 was labeled *gender* and allowed male, female, and intersex as values. Less than 3 hours later, one user, jdforrester, started a new discussion thread to point out the inconsistency between having the property label refer to gender but allowing values that refer to biological sex. They suggested that P21 should either be *gender* (and use masculine/feminine/neuter) or *sex* (and use male/female/intersex). What followed was a community debate over whether P21 should represent gender or represent sex, and what vocabulary would best model either of these. While the community acknowledged there were issues that needed to be addressed, what changes to make were not obvious, and the community struggled to agree upon any one model. Another user, Stryn, suggested merely changing male and female to man and woman instead of using what they felt was dehumanizing masculine and feminine terminology. Rather than directly address the merits of either proposal, the discussion thread focused on the applicability of various transliterations of man, woman, male, and female that non-English languages use when talking about gendered people. Eventually, jdforrester, citing support from an unspecified IRC chat, changed the label to sex. From property creation to posting their initial dispute, to implementing the change to sex, just under 24 hours passed.

Fast turnaround on anything involving P21 would turn out to be the exception, not the rule. A few weeks later, another user, without any discussion, changed the recommended values of P21 from male and female to male organism and female organism. This change, made with no community consensus, persisted for a month without comment until Zolo, advocating for the privileging of sex over gender in P21, authored a discussion post proposing to change P21's values back to male, female, and intersex for human subjects, which was quickly approved.

Over the next several months, there were multiple discussions arguing whether sex and gender should be split into two properties, or if the P21 label should be changed to gender. Consistent throughout were arguments over what the difference (if any) between sex and gender is as well as claims they could not combine them because the average editor would not know the difference. Following the failure of a proposal to create gender as its own property, contributors ignored the few remaining active dissenters to P21's label *sex or gender* and continued to allow both sex and gender values. In doing so, they ultimately avoided explicitly deciding between a biological sex-based model or a gender identity-based model, and thus accepted an unwieldy property rather than reconcile the nuances of different conceptualizations of sex and gender.

Here we have highlighted the tensions which arose in trying to give a diverse range of conceptualizations of gender a singular formal ontology with a singular vocabulary. Even within a singular cultural context, there is no monolithic understanding of gender when queer identities are taken into account. With no singular definition, no universal corpus or dictionary, it is unclear how to develop a coherent ontology to model gender. In order to scale, databases need such standardization, and need to use a concise vocabulary, that is sorely lacking for queerness. Queerness, and queer identities, are not scalable.

5.3 Privacy, Outing, and Singling Out

Thus far we have focused on the importance of including identity information to ensure the representation of queer people on Wikidata. However, the continued presence of some information presents issues of their own, primarily for trans individuals who may prefer to move away from a

pre-transition identity. In this section, we turn to issues that data can present for LGBTQ+ people, with a specific focus on where these cause friction with the archival needs for projects such as Wikidata.

Soon after his transition, Page's name on his entry presented issues. While the community updated his name on the overall entry label, they added in his deadname[7] as an alias for the entry. On Wikidata, aliases exist when an entry has multiple commonly used names. Aliases facilitate redirection, helping people and systems arrive at the correct data. Important here, aliases for an item can be queried and are also prominently displayed beneath the item description on the Wikidata entry (see Figure 1). From an archival perspective and for the purposes of a universal database, it is perfectly sensible, and indeed common practice, to record and use various aliases of any notable person. Even pre-transition, Page's entry had his last name both as given (Philpotts-Page) and as he performs (Page).

However, the use of aliases presents some specific challenges for trans individuals: aliases for trans people often include their deadname. In our examination of entries for trans people, deadnames were used as aliases for every trans person who was notable pre-transition. Having one's deadname not only publicly recorded but prominently accessible can be a hurtful and frustrating experience, reminding one of an identity they often would prefer to shed [19, 34, 44]. Deadnaming also can be invalidating of a person's trans identity and, by extension, the validity of trans people's identities as a whole. Even though many trans people would prefer their deadnames never appear in any circumstances, Wikidata's archival role prioritizes data completeness. Rather than treating deadnames as factually incorrect, and completely excising them from the database through deprecation or deletion, Wikidata's community practices treat past identities as outdated information, and still a part of the overall knowledge graph.

But even removing deadnames from aliases will not fully eliminate the possibility of deadnaming. We often found a subject's deadname present in another statement, often left as an outdated value of given name (P735) or as the value for birth name (P1477), a property that is used whenever the name someone is known by differs significantly from their birth name.

Values for given name can become either outdated, deprecated, or deleted when a new value is defined. Becoming outdated merely means adding an "end date" qualifier to the name and making it appear as a secondary instance of given name wherever it appears. While deprecation allows the deadname to remain viewable on the entry's Wikidata page, it will not be returned to 'end users' such as Wikipedia infoboxes via SPARQL or other queries. For trans people, their deadnames generally become outdated rather than deprecated, allowing for easy query access. Even the wholesale deletion of a deadname from an item does not completely remove it from Wikidata. Every past version of a page is still accessible in the document history tab and can be queried by some third-party extensions of Wikidata's SPARQL query service.

Our analysis found that the inclusion of deadnames is often the result of a perceived archival need rather than page redirection or simply version history. As evidenced by Chaz Bono, who publicly transitioned several years before Wikidata was even founded yet was notable as a celebrity child pre-transition, the presence of aliases is not merely an artifact of an individual transitioning after their entry was made and needing redirects due to a changed label. When surveying the entries for prominent public trans figures, we found that those who had a Wikidata page pre-transition routinely had a given name claim, with an end date qualifier, which still included their deadname. In contrast, trans public figures who came to prominence after they transitioned typically did

[7] "Deadnaming" is a colloquial term in the trans community to describe the often upsetting continued usage of a "deadname" used pre-transition.

not have their deadnames as either aliases or outdated given name claims, though some still had deadnames recorded as a birth name claim.

Despite social and platform preferences for maintaining information like aliases, in our analysis of contribution discussions, we found concerns around whether queer identities and related information should be shown at all. These conversations predominantly occurred in relation to sexuality. For example, in the property proposal discussion for P91, prior to the formal creation of the property itself (and the discussion of a deletion proposal that immediately followed), members of the community objected to the idea of recording and displaying sexuality. Several contributors characterized recording sexual orientation as "trivia" or "gossip," with user Giftzwerg 88, (a recurring opponent to the existence of the *sexual orientation* property), stating "Wikidata is not yellow press, the same with length, diameter, cup size, how often..." Another concern expressed was that people might add P91 claims based solely on rumors, and feel pressured to do so by the existence of this property. In the discussion of whether or not to delete P91 after it was created, Kolja21, who commented only once in the creation proposal, but was particularly active in the deletion proposal, initiated the proposal to delete *sexual orientation* by rhetorically demanding that if sexual orientation was to be kept as a property, penis size must also be created, specifically noting a potential use-case for porn stars and former presidents. Both are private pieces of information, implied Kolja21, which did not belong on Wikidata. While most community members called this argument ridiculous, one user replied suggesting that maybe Kolja21 was right: Perhaps penis size should be added as it could be notable in niche cases such as Jonah Falcon[8] (Q1321341).

Of course, databasing marginalized people has a complicated history [16], a history that Giftzwerg 88 referenced directly when describing their support for the deletion of the, at that time, very new P91:

> There are very few people marked as 'heterosexuality' so this property is mainly to mark homosexuals. Reminds me on something... a time when gay poeple [sic] had to wear a pink badges...

Expressing concern about the stigmatization of queer individuals, Giftzwerg 88 then linked to the Wikidata page for "Identification in Nazi camps" (Q169799). Similar concern over potential stigmatization cropped up across many comments, mirroring an almost identical conversation that took place on Wikipedia (complete with references to Nazis).

Concerns around stigmatization at the level of the schema (e.g., P91) were more prominent than concerns about stigmatizing data that fall into existing structures (e.g., aliases). This might suggest that the Wikidata community is willing to engage with issues when they are about the structure. However, low-level content is often overlooked. Alternatively, it may suggest concerns about stigmatization in potential changes, but a lack of concern for stigmatization that already exists on the platform. Furthermore, the utility of deadnames in aliases shows how some stigmatizing, harmful practices persist due to the privileging of functionality over sensitivity.

5.4 Neutrality and Bias in Community Discussions

Going into this study, we expected technological limitations and overheads of Wikidata to be a recurring rationale for confusing and unwieldy approaches to queer identities. While technical constraints certainly exist, we saw very few references to technological limitations in community discussions about P91 and even fewer for P21. Rather, we found that arguments that minimized and questioned the validity of queer identities were the more significant hindrance to improvements in the representation of queer people.

[8]Falcon gained widespread recognition in 1999 when he claimed to have the world's longest penis. These claims are indeed recorded as a length (P2043) statement

Because Wikidata operates via consensus, we saw that it was difficult for scientifically supported but not popularly embraced facts to gain a clear consensus. Perhaps the most prominent case involved recurrent discussions about splitting sex and gender identity into two distinct properties. Splitting gender and sex presents legitimate technical and usability questions – which should be prioritized, and could sex be used to populate gender if no other value was provided? However, debates instead focused on whether gender identity was "legitimate," slowing and stalling several proposed reworkings of P21.

Even when proposed changes had strong cases for improving performance, posts containing biologically essentialist (that is, privileging sex assigned at birth and dismissing any notion of gender as a social construct separate from sex) rhetoric prevented a clear consensus from forming, maintaining the status quo. For instance, when a new "gender identity" property was formally proposed, the arguments that plagued the discussion section of P21 arose here. Contributors questioned both if you can define and collect biological sex as well as if gender identity is even a 'real' thing. Giftzwerg 88 even claimed that gender identity is inherently non-factual, saying:

> Everyhing [sic] else is speculation. We also can not take care if a woman feels to be male or vice versa and the preference of reference. We must stick to hard facts.

While contributors on the P21 discussion page had largely expressed support for separating sex and gender in some way, this strong support was absent in the gender identity proposal thread. Instead, disagreements about small specifics, such as whether to keep P21 as sex and make the new property gender (or vice versa), and comments like the one quoted above prevented the community from ever achieving consensus. An admin closed the thread merely with: "**Not done** Consensus not reached." and despite continued and repeated suggestions for a separate property appearing in P21 discussions, no such change was ever made.

The examples presented thus far are neither the only nor most egregious in their cisheteronormative content. For instance, in response to an early proposal to change P21's label from sex to gender, Zolo bluntly stated:

> If by 'gender' you mean the sex someone identifies herself with, that seems rather problematic: really how do we know? Until recently it was not really common to publicly say "deep in my vagina, I feel I am a boy". If you just mean that gender should be used for biological sex, I guess that makes sense ...

Very recently, a user new to the *sex or gender* debate opened a new discussion thread, arguing to split sex and gender specifically to minimize gender, claiming:

> the related male/female/man/woman terms are also all horribly conflated, vague, and unclear. Which again seems like it ruins the entire point of the dataset. If "male" can refer to someone who is biologically female, how is that useful? Ultimately it seems the intended usage is for gender identity, and then assuming that unless otherwise stated, a person's sex is gender identity

They went on to clarify their reasoning as to why gender needs to be different from sex:

> it seems that a fictitious concept of gender identity is coined and applied to everyone (despite it not being a thing) and then people using this as the justification for how GID and Gender Dysphoria result, when this is untrue. There is then further attempts to conflate Transgender/GID/Dysphoria with Transsexualism, and Gender Identity with Sex.

Comments containing similar viewpoints crop up in the majority of discussion threads on P21. Likewise, arguments which minimized the impact of queer identities in a subject's life were common when *sexual orientation* was first proposed and implemented as a property. For instance, users

referring to a subject's sexuality as "just trivia" not important enough to record. Because the negative emotional impacts of anti-LGBTQ+ actions extend beyond the original victim [47], the constant presence of antagonistic views increases the emotional drain for queer individuals when participating in discussions about their own identities. In a system where debates go until apparent consensus is reached, the cost of seeing upsetting posts causes extra attrition in long, drawn-out stalemates.

6 DISCUSSION

As we traced the evolution of *sex or gender* (P21) and *sexual orientation* (P91) over seven years, we found significant struggles amongst the community to come to a consensus about what aspects of queer identities should be modeled in the database as well as how. First, we found that there were often discrepancies between what the community intended with a certain change and what ended up happening. While many discussion posts attempted to forecast the effects a change would have, some outcomes were inevitably missed, or secondary effects of a change were not considered. These oversights created a surprising amount of space for problems to unexpectedly arise, which could only be adequately addressed after the fact by the community in a patchwork way, which often failed to remedy the situation quickly.

Many contributors worried that recording queer identities, especially with regard to sexuality, would lead to significant privacy violations. Some believed that the existence of a *sexual orientation* property itself would lead to contributors aggressively outing or wrongly attributing identities to subjects based solely on rumor and speculation. Some contributors believed that, even when verified and factually correct, sexuality was no more than irrelevant gossip, while others worried that recording queerness would lead to othering and stigmatization.

After the concerns of whether or not we should record these identities, tensions arose in the specific verbiage, labels, and classification schemes that should be used. We found that the variety in vocabulary describing and the broad range of conceptualizations of certain aspects of queer identities, especially around gender, led to many proposed ways to model queerness, even amongst good-faith, highly knowledgeable, and LGBTQ+ allies in the edits. There is no standardized queer vocabulary, and whether a certain phrasing or categorization imparts enough relevant information and is sensitive to queer identities is highly contextually dependent.

We also found that, despite the best efforts of many well-meaning community members, some changes which were proposed and/or implemented inscribed negative biases against queer identities and individuals into the database. Discussion posts supporting such changes oftentimes resembled common talking points employed by explicitly anti-queer or anti-trans groups. When changes were proposed which would progress the representation of queer people, similar posts with similar arguments would often be made by a minority of individuals yet successfully hamper or even halt the opposed change.

6.1 Tensions Between Normative Databases and Queer Lives

Databases inherently traffic in standardized, analyzable data. Information is collected as a series of entries, each of which contains relevant attributes. However, when these attributes are more fluid than a list of responses, data becomes less computable, and databases quickly become unwieldy. When data is not captured in a standardized form with clear ways to measure or identify differences, the scale of the database is ultimately limited. In the pursuit of scaleable data schemas, database creators rely upon classification and standardization practices that reduce complex concepts into an essentialized representation, which assumes that any class of thing can be reliably described with the same set of predefined characteristics. Sure, a value may be lacking here or there, theoretically from measurement error or data loss, and the original observation is still being suitably described.

Such an essentialist grounding (that is, the belief that all things have a set of essential characteristics which defines what that object is) may work well when recording the number of chairs with a certain color in a warehouse, but is woefully unprepared for something whose essential characteristics cannot be clearly and consistently defined. Socially constructed and fluid concepts may especially lack a consistent set of essential characteristics.

Certainly, transgender identities, existing outside the typical cisnormative binary, challenge common institutional biological essentialism. In a society where gender is assigned to an infant based on the appearance of its genitals and recorded as simply male or female, assumptions based on a simplistic understanding of human biology dominate discourse surrounding what experiences of gender are valid or acceptable. Consider a woman who was born with a penis and no desire to change that along with their gender presentation; an intersex person, whose sex can not be clearly defined based on genitals or chromosomes; or a non-binary person whose gender correlates to neither of the commonly defined 'biological sexes' — each challenges the entire foundation of a sex-based view. Gender, within the realm of the queer, resists categorization, and consists of no more essential characteristics than how an individual feels (and even that much is debatable). One might argue that sexuality can be easily, though not perfectly, defined based on how one feels attraction towards another (romantic, sexual, or otherwise). In comparison, defining gender, which is merely one facet of a culturally situated feeling of the self, is a Sisyphean task. Contexts are not static, rather they are in constant flux. By the time one fully understands how gender behaves in a specific context, that context itself has already changed, and understanding of gender in the original time and place cannot blindly be extended to another time and place. There is an inherent disconnect between a database, trying to record essential characteristics as data points, and a nebulous idea of identity without any essential characteristics. As non-binary identities especially challenge cisnormative data creation practices, these challenges are amplified for cases where an individual's gender does not fall neatly into a man/woman binary, leading to extra tensions and discomfort for non-binary data subjects [59, 63]. Consider a young queer person who does not yet fully understand their gender identity. Perhaps they feel largely aligned with one gender, are comfortable with some elements from their assigned gender, and are not comfortable with a non-binary label. Regardless of what an outside observer may want to call this (bigender, probably) the identity and label must ultimately be understood and defined by the individual first. When that individual does not or is incapable of fully understanding their gender, it becomes impossible for an outside observer to neatly categorize their identity into some set of pre-labeled boxes.

One might imagine that tensions we have observed could be cleared up by simply asking a subject how they identify within the framework provided by Wikidata (or whichever system one is working with). However, doing so assumes that data subjects can be contacted and that they would be willing to respond in a public way (since a private response via e-mail or direct messages would not suffice as a public statement for Wikidata's reference verification requirements). Even if such a policy or mechanism was instituted, it would still leave problems for any data subjects who are no longer in the public eye and cannot be contacted, will not publicly disclose their identity, have died, or simply do not respond.

And what if Elliot Page did respond but indicated that none of the options fit? Would the community deliberate around the adoption of a new label? Although researchers have developed data ontologies specifically focusing on gender, sex, and sexual orientation [40], queer critiques argue that true objectivity is never possible, and no perfect, final version can be achieved [17]. Ruberg and Ruelos have previously discussed how queer identities challenge demographic norms [56]. We have seen similar anti-normative challenges with queer identities in Wikidata, with the additional difficulty that, while demographics are designed for the specific purpose they are being gathered for, Wikidata is supposed to be a universal database. Rather than collecting data for a very well-defined

purpose with a known scope, a universal database aims to be accurate in all cases, which greatly complicates the task of modeling gender and sexuality. While we as scholars would advocate for as inclusive of a database as possible, it is not hard to imagine how soliciting feedback from data subjects would result in overheads and debates that the Wikidata community is not able to manage.

Past identities, particularly for trans data subjects, introduce additional tensions. Given the desire of trans individuals to minimize past identities, we agree with prior work about the need for digital forgetting (as defined by [28]) in databases and archives. By identifying a major opposition to digital forgetting in archival needs for completeness, we highlight a tension not present in the literature focused on digital forgetting on social networking sites [28] and personal information and communication systems [43]. With public databases, the data subject often lacks the same awareness and control over their presentation that is often assumed on social networking sites, leaving data subjects reliant on system-level design choices and policies to forget their past identities. Rather than creating a profile to represent oneself, the archivist is collecting information about subjects. Even though a pre-coming out identity may no longer be accurate, the database author has to decide if there is sufficient archival value in preserving past names, genders, or sexual orientations to merit preserving them, even when faced with the fact that subjects may not want this. We have repeatedly seen that, whether implicitly or explicitly, contributors often privilege the completeness of the archive over the queer desire to forget.

The queer conceptualization of gender, importantly labeled with a word used to describe the strange and unfamiliar, is inherently poorly suited, even in some cases incapable, of being transformed into data. We clearly see this in *Troublesome Vocabulary* and *Neutrality and Bias in Community Discussions*, where the community struggled to identify the essential characteristics of gender. Community members were influenced by their worldview and often found themselves unable to develop a vocabulary robust enough to adequately accommodate various essential characteristics of gender under their different and often contrasting views. Developing such a vocabulary may be impossible, especially when the main polarizing difference amongst discussants was beliefs about the role of sex assigned at birth. One might hold out hope for a robust and inclusive set of terms that would be respectful of trans identities (prioritizing their personally felt gender) and indicate biological sex (either through a direct statement of sex assigned at birth or trans status), while still remaining relatively succinct so as to be scalable and usable by cisnormative contributors. However, even with such a verbiage panacea, many issues would remain. Anyone experiencing gender beyond the most common notions would be lumped into some poorly defined 'other' category, as we have seen non-binary be used at this time. Many trans people would still feel uncomfortable with their sex assigned at birth still being considered a notable part of their identity, and indicating either biological notions of sex or trans identity would still inherently other trans individuals. Gender anarchists would still argue that no such labels should exist at all, regardless of how well they may work, and that any such labels should be, at most, totally irrelevant. We see gender's resistance to being categorized, quantified, or even defined unless it is oversimplified to fit in unrealistically small boxes.

The tensions between normative databases and queer lives, especially in the anti-normative queering of categories that cannot scale, are relevant to any and all scalable databases which include data about queer individuals and their identities. Koopman argues that the forms and schemas of information creation and storage exert informational power on data subjects, and binds individuals to conceptualizations of identities which are calcified in the data formatting in order to function within our information-driven society [38]. We see that the items, categories, and properties that are codified in a database structure exert informational power on subjects, which is especially noticeable for queer people, and that queer individuals and communities can become bound to the structures and schema of databases they interact with (sometimes unknowingly or unwillingly).

What a data field allows, or requires, becomes reality. Through the exertion of informational power, a subject's complex and fluid gender identity is reduced to and bound by the simplistic categorization allowed by the database.

Here we find the distinction between two different definitions of the term "queer" instructive. Queer, as a reclaimed term, is often used to describe LGBTQ+ people. When thinking about queer archives, we are then left with challenges about how to capture the ever-evolving terminology used to describe queer people and their lives in ways that are authentic, representative, and respectful. Yet queer has a second meaning as well, more aligned with traditional queer theory: that which resists or upends that which is normative. It is in this definition that we see a second tension: queering the database. What does it mean for the database, a bastion of normative standardization, to be queered, to resist its own normative impulses? The properties which give the database its value and function (e.g., scalability, searchability, relational schemas, etc.) are all reliant on the use of highly normative standards that can fasten the smallest of data points into the broadest of systems. If there is no longer the core assumption of a consistent organizational structure, the database loses all practical value. It is too big to be read by a human, and too inconsistent to be processed by a computer. We ask, then, what would be involved in capturing queerness in a normative database like Wikidata? Any data would have to be captured in a limiting way, normalized, cleaned, and standardized, all to fit into an acceptable box. For Wikidata, the answer appears to be a flexible box, one with some allowed variety, but a box nonetheless. The harnessing, normalizing, and standardizing of that data, in essence, may strip queer data subjects of their queerness.

6.2 The Challenges of Open Collaboration

One of Wikidata's biggest strengths is also its Achilles' heel: its large, collaborative, editing community. It is only through the generous donation of time and labor by these volunteers that allows Wikidata—and every Wikimedia project—to thrive, but these volunteers often attempt to contribute to topics they lack expertise in and can not always work collaboratively. While Wikidata aims to take a neutral, objective, and factual approach, the viewpoints which have come to dominate discourse and practice show neutrality and objectivity have not been achieved. The discussions that we uncovered in our study demonstrate that debates often devolve into bickering and tribalism, with various factions merely butting heads rather than debating how to achieve an optimal solution that satisfies all. When personal opinions are bigoted or biased there are no mechanisms to combat them except the labor of other volunteers, who may not always be able or willing to intervene. The community itself has a primary goal of creating an effective and accurate database. However, queer identities challenge the very idea of factual accuracy. Fluidity of queer identities across time and context, both for an individual queer person and for the community as a whole, makes the possibility of a truly accurate data point impossible. With no capability to make a standardized ontology that is universally accurate, contributors resort to solutions which prioritize functionality over centering the experiences and preferences of queer individuals, which can be seen clearly in the broad use of deadnames for page aliases. Throughout the years of debate over whether P21 should be sex and/or gender, most contributors prioritized functionality, even when it meant compromises against the best interests of queer subjects.

To say that popular opinion around queer identities is highly politicized and polarized would be an understatement. It is hard to imagine that anyone is truly neutral towards queer existence, and even with neutrality you implicitly side with oppressive power structures. Compromise becomes a slightly more palatable option, giving some conceits to get some improvement, however, everything conceded will carry some potential harm or injustice of the existent structural oppression, souring the gains that were achieved. Leaving volunteer communities to slog out unwinnable ideological or political battles around the handling of queer identities will lead to never-ending debate. In these

cases, top-down interventions could dictate how a sensitive topic will be treated and eliminate the need for debates between opposing viewpoints. However, such authoritarian action which ignores a significant portion of the vocal contributors stands opposed to the collaborative, consensus-based principle that is the corner of Wikimedia projects, and is no more capable of creating an accurate ontology, only a consistent one.

Wikimedia's consensus-based approach assumes that a majority of (ideally, all) actors are starting from a reasonable position and hearing out opposing arguments in good faith [2]. But can that assumption be extended to anonymous laypeople on the internet? As there is no explicit threshold for consensus given by Wikidata or its parent Wikimedia, it is hard to know just how many individuals refusing to hear opposing views, stubbornly defending an "eccentric" position, or otherwise acting in bad faith would be needed to entirely break the consensus model. We have seen in both the debates over sex vs gender for P21 and the debates over the need for a *sexual orientation* property how easily a minority of contributors can halt discussion progress and prevent changes against the current status quo with merely an occasional objection or outright dismissal of opposing arguments and rebuttals. When every voice must be heard, it becomes easy for the loud and dangerous voices to drown out everyone else.

But what happens should the community come to a problematic consensus, either having unexpected consequences or being founded upon views the community no longer wants to promote? What mechanisms are in place to correct such a case, or better yet, prevent such a case from happening in the first place? Administrators may lock or restrict a page, as we saw repeatedly in the case of Elliot Page, but locks only work to prevent singular spontaneous edits and do nothing to stop the community as a whole. Changing a consensus after the fact is both slow and difficult. When we saw this happen with the changing label of P21, the initial consensus to change the label from gender to sex was a simple proposal with relatively few comments. The next time the label was changed by consensus was 10 months later, based on a much more elaborate proposal, and consisted of far more discussion posts mostly from contributors who had not participated in the previous change. The label change itself was only approved after negotiating down from the proposed full overhaul of P21. The consensus-based model itself hampers the capability to respond effectively to a prior consensus, which could be mitigated once again by top-down dictations to bypass achieving consensus. Once again, such authoritarian interventions to outright reverse a previously achieved consensus would go against one of the core principles of Wikidata.

Kriplean et al. [39], in an analysis of community dynamics on Wikipedia, identified several "power plays" used by Wikipedians to gain or maintain control of an article. Our observations of dynamics within Wikidata were consistent with these, and we especially noticed the practices of appealing to the article scope as well as utilizing prior consensus driving dynamics. However, the changes being discussed on Wikipedia are, generally, limited to a singular article as far as scope is concerned. On Wikidata, changes made to structural properties or item classes have a much broader effect than a change to one Wikipedia article. Therefore, the dynamics at play in Wikipedia often have higher stakes involved when they occur on Wikidata. The shared value of a neutral point of view, which has been seen driving dispute resolutions on Wikipedia [67], was not present or was insufficient for resolving the more impactful disputes we found when it comes to structural aspects of a database.

The issues here should not be taken to imply that Wikidata (and other such collaborative knowledge organization projects) should discard their collaborative foundation altogether. Ultimately, open collaboration amongst volunteers is the only reason such a massive database project can be undertaken and accessed freely by any end-user. However, our findings have highlighted how the process can easily break down around contentious, polarizing topics such as queer identities. Top-down interventions could be applied to correct such cases. Wikidata as a platform could specifically

mandate how to treat certain contentious topics. Identification and bans could be levied against bad-faith actors filibustering. Administrators could have the power to review and explicitly reverse a harmful or dysfunctional consensus-approved change. However, regardless of the effectiveness of such interventions, they go against one of the core principles of Wikidata: open collaboration.

7 CONCLUSION

After a week of frantic edits, Elliot Page's Wikidata page reached some level of equilibrium. While there continues to be occasional vandalism, and some users still try to variously push for his entry to emphasize or remove his pre-transition identity,the values of *sex or gender* (P21) and *sexual orientation* (P91) have settled into non-binary and non-heterosexual. Yet, Elliot Page is merely one case where the volatility and difficulty of codifying queerness erupt. Without systemic solutions, the problems which arose for Elliot Page will occur again for data about other queer and trans people.

In our analysis, we demonstrated tensions between data schemas, the communities that develop and maintain them, and the lives these schemas are intended to represent. Our findings clearly illustrate challenges when representing LGBTQ+ people in data, as well as frictions these data can cause for archival projects. As seen in our trace ethnography, queer identities challenge database requirements of standardization and scalability. However, in the case of platforms like Wikidata, the social layer seen in contributor discussions adds an extra layer of challenge, with contributors split not only over what is most functional, but also over what viewpoints to reflect.

While our work was motivated around ensuring accurate representation of queer people on Wikidata, our findings also highlight broader tensions that databases like Wikidata encounter when trying to both represent the diversity of lived experience and standardize data in the way technology requires. It is likely the case that the open collaborative nature of projects like Wikidata will always face challenges when addressing polarizing topics – challenges that will require designers to consider the limitations of community-based decisions.

ACKNOWLEDGMENTS

Funded in part with support from the Gill Foundation. Thanks to Brian C. Keegan for his guidance with Wikidata.

REFERENCES

[1] [n.d.]. https://meta.wikimedia.org/wiki/Edit_war Accessed: 2021-08-29.
[2] [n.d.]. https://meta.wikimedia.org/wiki/Consensus Accessed: 2022-01-14.
[3] [n.d.]. Statistics - Wikidata. https://www.wikidata.org/wiki/Special:Statistics Accessed: 2021-05-20.
[4] [n.d.]. Wikidata:Introduction - Wikidata. https://www.wikidata.org/wiki/Wikidata:Introduction Accessed: 2021-06-01.
[5] [n.d.]. Wikidata:Tools/Visualize data - Wikidata. https://www.wikidata.org/wiki/Wikidata:Tools/Visualize_data Accessed: 2021-05-28.
[6] [n.d.]. Wikidata:WikiProject LGBT - Wikidata. https://www.wikidata.org/wiki/Wikidata:WikiProject_LGBT Accessed: 2021-05-31.
[7] [n.d.]. Wikimedia Statistics - Wikidata - Edited pages. https://stats.wikimedia.org/#/wikidata.org/content/edited-pages/normal|line|2-year|~total|daily Accessed: 2021-05-27.
[8] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness. *arXiv:2011.02282 [cs]* (Jan. 2021). http://arxiv.org/abs/2011.02282 arXiv: 2011.02282.
[9] Jeffrey Bardzell and Shaowen Bardzell. 2015. The User Reconfigured: On Subjectivities of Information. In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives* (Aarhus, Denmark) *(CA '15)*. Aarhus University Press, Aarhus N, 133–144. https://doi.org/10.7146/aahcc.v1i1.21298
[10] Greta R. Bauer, Jessica Braimoh, Ayden I. Scheim, and Christoffer Dharma. 2017. Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PLOS ONE* 12, 5 (May 2017), e0178043. https://doi.org/10.1371/journal.pone.0178043
[11] Greta R. Bauer, Jessica Braimoh, Ayden I. Scheim, and Christoffer Dharma. 2017. Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PLOS ONE* 12, 5 (May 2017), e0178043. https://doi.org/10.1371/journal.pone.0178043
[12] Courtney Blackwell, Jeremy Birnholtz, and Charles Abbott. 2015. Seeing and being seen: Co-situation and impression formation using Grindr, a location-aware gay dating app. *New Media & Society* 17, 7 (Aug 2015), 1117–1136. https://doi.org/10.1177/1461444814521595
[13] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press. Google-Books-ID: xHlP8WqzizYC.
[14] Lisa Bowleg. 2008. When Black + Lesbian + Woman ≠ Black Lesbian Woman: The Methodological Challenges of Qualitative and Quantitative Intersectionality Research. *Sex Roles* 59, 5-6 (Sept. 2008), 312–325. https://doi.org/10.1007/s11199-008-9400-z
[15] Samantha Breslin and Bimlesh Wadhwa. 2014. Exploring Nuanced Gender Perspectives within the HCI Community. In *Proceedings of the India HCI 2014 Conference on Human Computer Interaction* (New Delhi, India) *(IndiaHCI '14)*. Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/2676702.2676709
[16] Ann Cvetkovich. 2003. *An Archive of Feelings: Trauma, Sexuality, and Lesbian Public Cultures*. Duke University Press. https://doi.org/10.1215/9780822384434
[17] Emily Drabinski. 2013. Queering the Catalog: Queer Theory and the Politics of Correction. *The Library Quarterly* 83, 2 (2013), 94–111. https://doi.org/10.1086/669547 arXiv:https://doi.org/10.1086/669547
[18] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a Definition of Knowledge Graphs. In *SEMANTiCS*.
[19] Michael J. Faris. 2019. On Trans Dignity, Deadnaming, and Misgendering: What Queer Theory Rhetorics might Teach Us about Sensitivity, Pedagogy, and Rhetoricity. (March 2019). https://ttu-ir.tdl.org/handle/2346/84247
[20] Michel Foucault and Robert Hurley. 1988. *The history of sexuality*.
[21] R. Stuart Geiger and David Ribes. 2011. Trace Ethnography: Following Coordination through Documentary Practices. In *2011 44th Hawaii International Conference on System Sciences*. 1–10. https://doi.org/10.1109/HICSS.2011.455
[22] R Stuart Geiger and David Ribes. 2011. Trace ethnography: Following coordination through documentary practices. In *2011 44th Hawaii international conference on system sciences*. IEEE, 1–10.
[23] Sandra Gonzalez. [n.d.]. "Juno" star Elliot Page shares transgender identity. https://www.cnn.com/2020/12/01/entertainment/elliot-page-trnd/index.html
[24] Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 2 (1993), 199–220. https://doi.org/10.1006/knac.1993.1008
[25] Thomas R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* 43, 5 (1995), 907–928. https://doi.org/10.1006/ijhc.1995.1081
[26] Nicola Guarino. 1997. Understanding, building and using ontologies. *International Journal of Human-Computer Studies* 46, 2 (1997), 293–310. https://doi.org/10.1006/ijhc.1996.0091
[27] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. 2015. Disclosure, Stress, and Support During Gender Transition on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY,

USA, 1176–1190. https://doi.org/10.1145/2675133.2675152

[28] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. 2016. Digital Footprints and Changing Networks During Online Identity Transitions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2895–2907. https://doi.org/10.1145/2858036.2858136

[29] Corie Jo Hammers. 2016. Queer. In *The SAGE Encyclopedia of LGBTQ Studies*. SAGE Publications, Inc., Thousand Oaks,, 907–908. https://doi.org/10.4135/9781483371283

[30] Jean Hardy. 2021. Queer information literacies: social and technological circulation in the rural Midwestern United States. *Information, Communication & Society* 24, 1 (2021), 102–117. https://doi.org/10.1080/1369118X.2019.1635184 arXiv:https://doi.org/10.1080/1369118X.2019.1635184

[31] Ben Hunte. 2020. Elliot Page: Juno star announces he is transgender. *BBC News* (Dec 2020). https://www.bbc.com/news/world-us-canada-55147975

[32] Dariusz Jemielniak. 2014. *Common Knowledge?: An Ethnography of Wikipedia*. Stanford University Press.

[33] Jeffrey M. Jones. 2021. LGBT Identification Rises to 5.6% in Latest U.S. Estimate. https://news.gallup.com/poll/329708/lgbt-identification-rises-latest-estimate.aspx

[34] Stephanie Julia Kapusta. 2016. Misgendering and Its Moral Contestability. *Hypatia* 31, 3 (2016), 502–519. https://doi.org/10.1111/hypa.12259

[35] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration* (Galway, Ireland) *(OpenSym '17)*. Association for Computing Machinery, New York, NY, USA, Article 14, 5 pages. https://doi.org/10.1145/3125433.3125465

[36] Lucie-Aimée Kaffee and Elena Simperl. 2018. Analysis of Editors' Languages in Wikidata. In *Proceedings of the 14th International Symposium on Open Collaboration* (Paris, France) *(OpenSym '18)*. Association for Computing Machinery, New York, NY, USA, Article 21, 5 pages. https://doi.org/10.1145/3233391.3233965

[37] Jonathan Kemp. 2009. Queer Past, Queer Present, Queer Future. *Graduate Journal of Social Science* 6, 1 (2009), 3–23. http://gjss.org/content/queer-past-queer-present-queer-future

[38] Colin Koopman. 2019. *How We Became Our Data: A Genealogy of the Informational Person*. University of Chicago Press. https://doi.org/doi:10.7208/9780226626611

[39] Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, Consensus, Coercion, Control: Cs*w or How Policy Mediates Mass Participation. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '07)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/1316624.1316648

[40] Clair A Kronk and Judith W Dexheimer. 2020. Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow. *Journal of the American Medical Informatics Association* 27, 7 (06 2020), 1110–1115. https://doi.org/10.1093/jamia/ocaa061 arXiv:https://academic.oup.com/jamia/article-pdf/27/7/1110/34153094/ocaa061.pdf

[41] Ann Light. 2011. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with Computers* 23, 5 (2011), 430–438. https://doi.org/10.1016/j.intcom.2011.02.002 Feminism and HCI: New Perspectives.

[42] Devon Magliozzi, Aliya Saperstein, and Laurel Westbrook. 2016. Scaling Up: Representing Gender Diversity in Survey Research. *Socius: Sociological Research for a Dynamic World* 2 (Jan. 2016), 237802311666435. https://doi.org/10.1177/2378023116664352

[43] Viktor Mayer-Schönberger. 2011. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton University Press. https://doi.org/doi:10.1515/9781400838455

[44] Kevin A. McLemore. 2015. Experiences with Misgendering: Identity Misclassification of Transgender Spectrum Individuals. *Self and Identity* 14, 1 (Jan 2015), 51–74. https://doi.org/10.1080/15298868.2014.950691

[45] Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. Peer-Production System or Collaborative Ontology Engineering Effort: What is Wikidata?. In *Proceedings of the 11th International Symposium on Open Collaboration* (San Francisco, California) *(OpenSym '15)*. Association for Computing Machinery, New York, NY, USA, Article 20, 10 pages. https://doi.org/10.1145/2788993.2789836

[46] Elliot Page. 2020. Instagram photo by @elliotpage. https://www.instagram.com/p/CIQ1QFBhNFg/ Accessed: 2021-05-26.

[47] Jenny L. Paterson, Rupert Brown, and Mark A. Walters. 2019. Feeling for and as a group member: Understanding LGBT victimization via group-based empathy and intergroup emotions. *British Journal of Social Psychology* 58, 1 (2019), 211–224. https://doi.org/10.1111/bjso.12269 arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjso.12269

[48] Isabella Peters and Wolfgang G. Stock. 2007. Folksonomy and information retrieval. *Proceedings of the American Society for Information Science and Technology* 44, 1 (2007), 1–28. https://doi.org/10.1002/meet.1450440226 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.1450440226

[49] Alessandro Piscopo, Chris Phethean, and Elena Simperl. 2017. What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *Social Informatics (Lecture Notes in Computer Science)*, Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer International Publishing, 305–322. https://doi.org/10.1007/978-3-319-67217-5_19

[50] Alessandro Piscopo, Christopher Phethean, and Elena Simperl. 2017. Wikidatians are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base. https://doi.org/10.24251/HICSS.2017.527

[51] Alessandro Piscopo and Elena Simperl. 2018. Who Models the World? Collaborative Ontology Creation and User Roles in Wikidata. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 141 (nov 2018), 18 pages. https://doi.org/10.1145/3274410

[52] Jian Qin and Stephen Paling. 2001. Converting a Controlled Vocabulary into an Ontology: the Case of GEM. *Information Research: An International Electronic Journal* 6, 2 (Jan 2001). https://surface.syr.edu/istpub/38

[53] K.J. Rawson. 2010. Accessing Transgender // Desiring Queer(er?) Archival Logics. *Archivaria* 68 (Jan. 2010), 123–140. https://archivaria.ca/index.php/archivaria/article/view/13234

[54] KJ Rawson. 2014. Archive. *Transgender Studies Quarterly* 1, 1-2 (2014), 24–26.

[55] K. J. Rawson. 2018. The Rhetorical Power of Archival Description: Classifying Images of Gender Transgression. *Rhetoric Society Quarterly* 48, 4 (2018), 327–351. https://doi.org/10.1080/02773945.2017.1347951 arXiv:https://doi.org/10.1080/02773945.2017.1347951

[56] Bonnie Ruberg and Spencer Ruelos. 2020. *Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics.* https://journals.sagepub.com/doi/full/10.1177/2053951720933286

[57] Maya Salam. 2020. Elliot Page, Oscar-Nominated 'Juno' Star, Announces He Is Transgender. (2020). https://www.nytimes.com/2020/12/01/movies/elliot-page-transgender-juno.html

[58] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 155 (nov 2018), 27 pages. https://doi.org/10.1145/3274424

[59] Morgan Klaus Scheuerman, Aaron Jiang, Katta Spiel, and Jed R. Brubaker. 2021. Revisiting Gendered Web Forms: An Evaluation of Gender Inputs with (Non-)Binary People. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 400, 18 pages. https://doi.org/10.1145/3411764.3445742

[60] Catherine Shoard. 2020. Elliot Page: star of Juno and X-Men announces he is transgender. http://www.theguardian.com/film/2020/dec/01/elliot-page-star-of-juno-x-men-announces-he-is-transgender

[61] Tom Simonte. 2019. Inside the Alexa-Friendly World of Wikidata. *Wired* (Feb. 2019). https://www.wired.com/story/inside-the-alexa-friendly-world-of-wikidata/

[62] Teri Slade, Douglas P. Gross, Laura Niwa, Ashley B. McKillop, and Christine Guptill. 2020. Sex and gender demographic questions: improving methodological quality, inclusivity, and ethical administration. *International Journal of Social Research Methodology* 0, 0 (Sept. 2020), 1–12. https://doi.org/10.1080/13645579.2020.1819518

[63] Katta Spiel. 2021. "Why Are They All Obsessed with Gender?" — (Non)Binary Navigations through Technological Infrastructures. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 478–494. https://doi.org/10.1145/3461778.3462033

[64] Jesús Tramullas, Piedad Garrido-Picazo, and Ana I. Sánchez-Casabón. 2018. Use of Wikipedia Categories on Information Retrieval Research: A Brief Review. In *Proceedings of the 5th Spanish Conference on Information Retrieval* (Zaragoza, Spain) *(CERI '18)*. ACM, New York, NY, USA, Article 17, 4 pages. https://doi.org/10.1145/3230599.3230617

[65] J. Trant. 2009. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10, 1 (Jan 2009). https://journals.tdl.org/jodi/index.php/jodi/article/view/269

[66] David Valentine. 2007. *Imagining Transgender: An Ethnography of a Category.* Duke University Press. https://doi.org/10.1215/9780822390213

[67] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying Cooperation and Conflict between Authors with History Flow Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 575–582. https://doi.org/10.1145/985692.985765

[68] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. 2006. Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland) *(WWW '06)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/1135777.1135863

[69] Rachel Wexelbaum. 2019. Edit Loud, Edit Proud: LGBTIQ+ Wikimedians and Global Information Activism. *Wikipedia @ 20* (June 2019). https://wikipedia20.pubpub.org/pub/rdtfo5v8/release/3

[70] SBS Staff Writers. 2020. Netflix amends Elliot Page's name on all past credits. https://www.sbs.com.au/topics/pride/fast-lane/article/2020/12/04/netflix-amends-elliot-pages-name-all-past-credits

Received January 2022; revised July 2022; accepted November 2022