Diversitat i representació de biografies a les portades de Wikipedia

Anàlisis interseccional dels articles de persones de l'edició en anglès

Aisa Serra Gil

Tutors: Miquel Centelles i Núria Ferran

Presentat el 17 de juny de 2024

Curs 2023-2024

Màster en Humanitats Digitals

Universitat de Barcelona

Índex

RESUM

En l'era digital contemporània, Wikipedia s'ha consolidat com una de les principals fonts de coneixement i informació global. Tot i la seva missió de proporcionar accés lliure al coneixement, la plataforma presenta bretxes significatives en la representació de persones, especialment pel que fa al gènere, l'ètnia i altres aspectes identitaris. Aquest treball investiga aquestes bretxes de contingut a l'edició en anglès de Wikipedia, centrant-se en les biografies destacades a les portades. Mitjançant una perspectiva interseccional, s'avalua com les identitats de gènere, ètnia, orientació sexual i altres factors influeixen en la visibilitat i representació de les persones a la plataforma. L'anàlisi quantitativa revela una subrepresentació de biografies femenines i no-binàries, així com una predominança de persones del Nord Global en contraposició a altres regions, entre d'altres resultats. Aquesta investigació pretén no només identificar aquests desequilibris, sinó també proposar un esquema de metadades per millorar la qualitat i precisió de les biografies, fomentant una Wikipedia més inclusiva. Els resultats ofereixen una base empírica per a recomanacions que promoguin la diversitat i equitat en la representació de persones a la plataforma, contribuint a una cultura del coneixement més justa i accessible per a tothom.

PARAULES CLAU: Wikipedia, portada, bretxa de contingut, gènere, identitat, interseccionalitat, biografies, inclusió.

ABSTRACT

In the contemporary digital age, Wikipedia has established itself as one of the main sources of global knowledge and information. Despite its mission to provide free access to knowledge, the platform presents significant gaps in the representation of people, especially in terms of gender, ethnicity and other aspects of identity. This paper investigates these content gaps in the English edition of Wikipedia, focusing on the biographies featured on the covers. Using an intersectional perspective, it assesses how gender identities, ethnicity, sexual orientation and other factors influence people's visibility and representation on the platform. The quantitative analysis reveals an underrepresentation of female and non-binary biographies, as well as a predominance of people from the Global North as opposed to other regions, among other results. This research aims not only to identify these imbalances, but also to propose a metadata scheme to improve the quality and accuracy of biographies, encouraging a more inclusive Wikipedia. The results provide an empirical basis for recommendations that promote diversity and equity in the representation of people on the platform, contributing to a more just and accessible knowledge culture for all.

KEYWORDS: Wikipedia, Main page, content gap, gender, identity, intersectionality, biographies, inclusion.

Introducció

En l'era digital contemporània, Wikipedia ha emergit com una de les principals fonts de coneixement i informació a escala global (Beytía i Wagner, 2022; Fan i Gardent, 2022; Sefidari, 2022). Amb el seu accés lliure i la seva estructura col·laborativa, la plataforma ofereix una vasta gamma de continguts sobre temes diversos, essent una eina d'ús quotidià per a una gran quantitat de persones d'arreu del món que busquen informació accessible i veraç. No obstant això, malgrat la seva missió de proporcionar «free access to that knowledge, and be a start in the effort to bring about a world in which all knowledge is freely available to everyone» (Wikimedia:Prime objective), Wikipedia no està lliure de bretxes i desigualtats significatives en la representació de persones, especialment en relació al gènere, però també amb factors com l'etnicitat, l'origen geogràfic i altres aspectes identitaris.

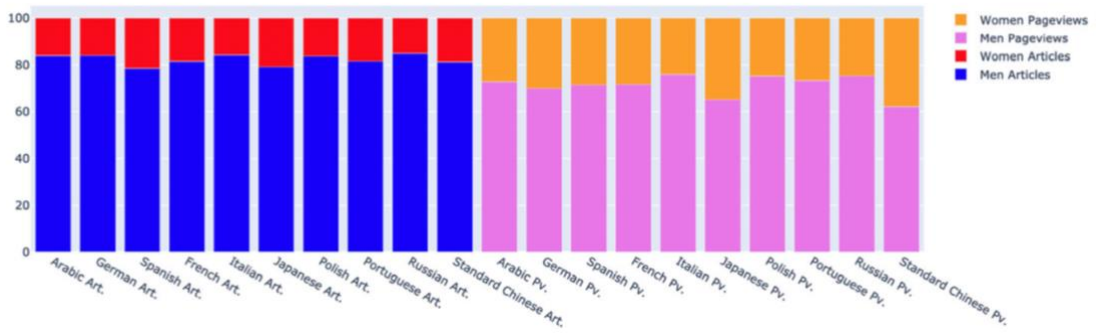

Figura 1. Bretxa de gènere tant en el nombre d'articles com en les pàgines visitades de les 10 edicions lingüístiques principals de Wikipedia en termes de nombre de persones editores (Miquel-Ribé i Laniado, 2021)

Aquest treball explora i analitza les bretxes de contingut a l'edició en anglès de Wikipedia, amb un enfocament específic en les biografies presentades a les seves portades. Mitjançant una perspectiva interseccional, aquest estudi avalua com les identitats d'una persona impacten en la representació i visibilitat de les persones i la diversitat d'identitats en aquesta plataforma digital. A través de la recopilació i l'anàlisi sistemàtica de dades, es busca identificar els aspectes identitaris que tenen una representació escassa o presenten una manca de diversitat, contribuint a una comprensió més profunda de com els biaixos estructurals es manifesten en els espais digitals. La importància d'aquesta investigació radica en la capacitat de Wikipedia per influir en la percepció pública i en el reconeixement de les persones a nivell global. Com a plataforma de referència per a la informació, Wikipedia juga un paper clau en la construcció de narratives i en la determinació de qui és considerat "notable" i, per tant, mereixedor de ser destacat en el món digital. Entendre com les diverses identitats són representades a les portades de Wikipedia és crucial per avaluar si la plataforma està complint amb el seu objectiu de reflectir de manera justa la diversitat humana i promoure una inclusió equitativa de totes les persones.

Les investigacions anteriors han identificat diversos biaixos en la representació de gènere a Wikipedia, destacant la notable disparitat entre biografies d'homes i de dones o persones no-binàries en la seva portada. No obstant això, és necessari ampliar aquesta anàlisi per incloure altres dimensions identitàries i comprendre com aquests aspectes interaccionen per influir en la visibilitat i reconeixement de les persones a la plataforma.

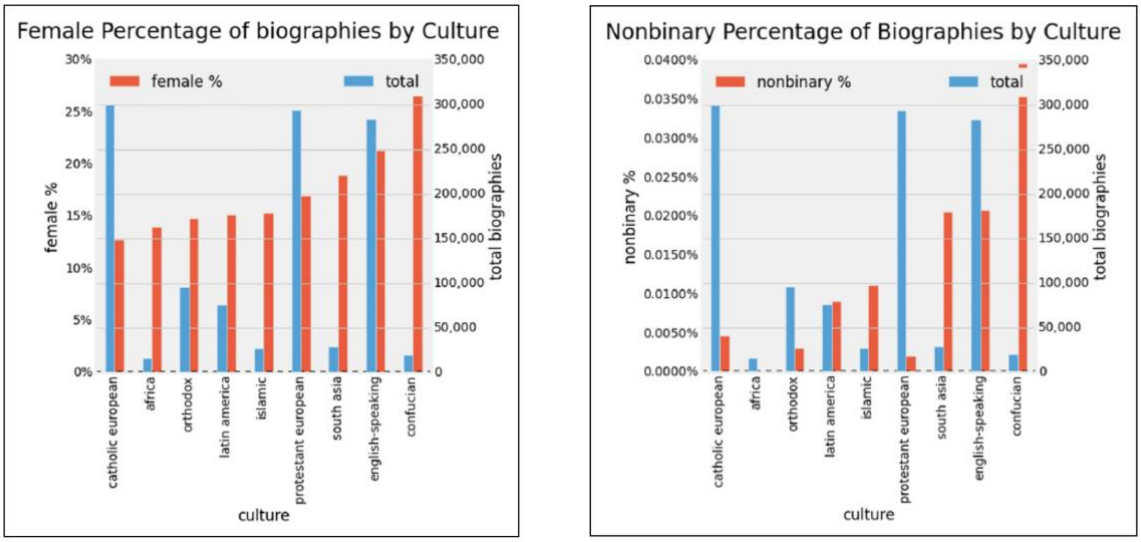

Figura 2. Proporció de biografies sobre dones (esquerra) i de persones no-binàries (dreta) per cultura a Wikipedia (Konieczny i Klein, 2018)

L'abordatge interseccional d'aquest estudi permetrà no només identificar les bretxes de contingut a Wikipedia, sinó també explorar les oportunitats per millorar la representació i fomentar una plataforma més inclusiva. Mitjançant l'ús de tècniques d'anàlisi quantitativa a través de l'*scraping* de dades, aquest treball aspira quantificar la visibilitat atorgada a diferents grups de persones en funció de les seves característiques interseccionals. Això proporcionarà una base empírica per a la formulació de recomanacions i estratègies per promoure una major diversitat i equitat en la representació de persones a Wikipedia, i permetrà proposar un esquema de metadades i propietats mínimes necessàries (o *entity schema*) per formular correctament una biografia en aquesta enciclopèdia en línia.

Amb aquest projecte, esperem obtenir una comprensió més profunda del pes i la diversitat de les biografies a les portades de Wikipedia en la seva edició en anglès. Mitjançant l'anàlisi detallada de les biografies destacades, els resultats podran revelar patrons i tendències en la representació de persones, identificant possibles biaixos i desequilibris en funció de factors com el gènere, l'ètnia, l'origen geogràfic, la cultura, la llengua i altres aspectes. Aquesta informació no només serà valuosa per a la comunitat de Wikipedia i les persones que hi editen, sinó també per a altres investigadors i plataformes en línia que busquen promoure una major diversitat i inclusió en els seus continguts digitals.

En última instància, aquest projecte podria contribuir a fomentar una representació més equitativa i diversa de la societat a Wikipedia. Identificar i abordar les bretxes de representació pot ajudar a crear un espai de coneixement més just i inclusiu, reflectint millor la diversitat humana. A més, la proposta d'un esquema de metadades per millorar la qualitat i la precisió de les biografies a Wikipedia és necessària per facilitar un accés més equitatiu i complet a la informació sobre persones destacades de tot el món i totes les èpoques, contribuint a una cultura del coneixement més inclusiva i accessible per a tothom.

Objectius

Aquest treball es proposa com a objectiu general (OG) analitzar la presència o absència i la diversitat, des del punt de vista interseccional, de les biografies a les portades de la Wikipedia de l'edició en anglès. Aquesta anàlisi permetrà avaluar la representativitat i diversitat de les persones destacades i proporcionarà una visió crítica sobre com es reflecteixen les dinàmiques socials i culturals en una plataforma de coneixement global com Wikipedia.

Per assolir aquest objectiu general, el projecte es desglossa en diversos objectius específics que guiaran l'anàlisi detallada de les portades de l'edició en anglès de Wikipedia entre el 2013 i el 2023:

Oe1. Comparar la presència de biografies respecte d'altres tipologies d'articles a les seccions "From today's featured article", "Did you know..." i "On this day" de les portades.

Oe2. Aplicar una perspectiva crítica per analitzar com factors com el gènere, l'ètnia, l'orientació sexual o la llengua d'origen poden influir en la representació de les persones a les portades.

Oe3. Identificar possibles desequilibris i biaixos en la representació de la diversitat a partir de quines persones apareixen a les portades.

Oe4. Proposar un esquema de metadades i propietats mínimes necessàries (*entity schema*) per descriure correctament les persones a Wikipedia, assegurant-ne la diversitat.

Per abordar aquests objectius, el projecte es planteja diverses preguntes d'investigació que orientaran l'anàlisi i ajudaran a evidenciar la dinàmica de representació a les portades de Wikipedia:

Pi1. Quin pes tenen les biografies respecte els altres tipus d'article que apareixen a les portades de Wikipedia?

Pi2. Com influeixen en la presència o representativitat característiques com la llengua, la nacionalitat, l'ètnia, l'època, el gènere o la sexualitat en la representació i la inclusió de les biografies a les portades?

Pi3. Quina és la diversitat de l'origen i la llengua de les persones representades a les portades de Wikipedia? Hi predominen les persones d'origen o llengua anglesa o hi ha més diversitat?

Pi4. Quines dades són necessàries per descriure adequadament una persona en una biografia de Wikipedia?

Metodologia

A continuació, s'especifica la metodologia utilitzada per a dur a terme l'estudi. En primer lloc, presentem una taula-resum de les metodologies combinades amb els objectius que aborden, els apartats on es reflecteix cada fase i les eines o tècniques utilitzades. En segon lloc, es desglossaran els passos fets per a plantejar el marc conceptual de l'anàlisi i obtenir les dades d'estudi. Seguidament, s'especificaran i justificaran les propietats seleccionades per a l'*scraping* de dades i es detallaran les fases d'anàlisi dels resultats. Per acabar, comentarem breument la metodologia emprada per formular la proposta d'esquema de metadades.

Per aconseguir els objectius, proposem adoptar un enfocament qualitatiu i quantitatiu basat en tres fases metodològiques: una revisió de la literatura *(scoping review*), un *scraping* de dades i l'anàlisi i visualització de les dades obtingudes mitjançant gràfiques.

Objectiu	**Metodologia**	**Apartat**	**Tècniques i eines**
Oe1	*Scoping review* Anàlisis de dades mitjançant *scraping*	*Scoping review* Exposició dels resultats	Scopus, WoS OpenRefine, Excel
Oe2	Anàlisi dels resultats des d'una perspectiva interseccional	Exposició dels resultats	*Scoping Review* Taxonomia sobre interseccionalitat que utilitza Rodó-Zárate (2021)
Oe3	Aproximació quantitativa amb *scraping* Reconciliació d'URI i enriquiment amb propietats específiques de Wikidata	Exposició dels resultats	OpenRefine, Wikidata
Oe4	Revisió de la literatura Anàlisi de les propietats de Wikidata	Proposta d'esquema de metadades	*Scoping Review* Wikidata

Taula 1. Esquema metodològic

En primer lloc, realitzarem una revisió de la literatura acadèmica en relació amb la temàtica del projecte, el que em permetrà, per una banda, determinar l'estat de la qüestió i saber quines línies de treball semblants al projecte s'han explorat, amb la finalitat de poder construir el

projecte en base al treball anterior. Per altra banda, també ens permetrà determinar el marc teòric a partir del qual desenvolupar la proposta d'esquema de metadades (o *entity schema*). Per fer la revisió sistematitzada de la literatura, ens basem en el mètode SALSA proposat per Grant i Booth (2009).

En segon lloc, farem una aproximació quantitativa a les portades de Wikipedia mitjançant un *scraping* de les dades de les seccions "From today's featured article", "Did you know..." i "On this day". Per fer-ho, farem servir l'aplicació OpenRefine per reconciliar les URI de cada secció de les portades. En primer lloc, necessitarem extreure el contingut de l'històric de Main Pages d'en.Wikipedia, l'edició de Wikipedia en anglès; i a continuació, farem la reconciliació amb identificadors únics i un enriquiment amb propietats de Wikidata per obtenir els valors de cada element per analitzar.

Aquestes dades aporten informació sobre les persones, com el gènere, l'ètnia, l'orientació sexual, la llengua d'origen i altres factors rellevants des d'una perspectiva interseccional. L'elecció d'aquestes propietats es basa en la seva rellevància per entendre les desigualtats en la representació a Wikipedia i parteixen principalment de la teoria interseccional exposada per Maria Rodó-Zárate a *Interseccionalitat. Desigualtats, llocs i emocions* (2021), a més d'altres fonts.

Segons Rodó-Zárate (2021, pp. 43-44), «[l]'elaboració de llistes d'eixos és un tema qüestionable que porta sempre a la visibilització d'uns aspectes i la invisibilització d'uns altres i que a més pot reforçar la separabilitat en comptes de focalitzar-se en la seva interrelació. Tot i això, crec que és important poder-los anomenar per poder distingir entre diferents formes de diferenciació social i també per fer visibles els que rarament s'acostumen a considerar. A continuació se'n presenten alguns amb l'objectiu de mostrar eixos que es podrien tenir en compte en anàlisis interseccionals, entenent la llista com a inacabada, incompleta i en permanent redefinició». Tenint aquesta nota en consideració, exposem i justifiquem les propietats seleccionades:

Nom	Codi	Definició[1]	Justificació
Place of birth	P19	Most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character	L'origen fa referència al lloc de naixement d'una persona i es relaciona amb processos de migració. En el context europeu, el sistema de control migratori es fonamenta sobre el colonialisme, que dona drets diferents a les persones segons el seu origen. (Rodó-Zárate, p. 45)
Place of death	P20	Most specific known (e.g. city instead of	L'origen fa referència al lloc de naixement d'una persona i es relaciona

[1] Definicions extretes de la llista de propietats de Wikidata (https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all)

		country, or hospital instead of city) death location of a person, animal or fictional character	amb processos de migració. En el context europeu, el sistema de control migratori es fonamenta sobre el colonialisme, que dona drets diferents a les persones segons el seu origen. (Rodó-Zárate, p. 45) A partir d'aquesta propietat podré saber si la persona va migrar, la qual cosa entra en el marc de la interseccionalitat.
Sex or gender	P21	Sex or gender identity of human or animal. For human: male, female, non-binary, intersex, transgender female, transgender male, agender, etc. For animal: male organism, female organism.	El sistema sexe–gènere té diferents dimensions que s'emmarquen en les normativitats derivades del sistema cisheteropatriarcal. Aquestes diferents dimensions inclouen el sexe —classificació de les corporalitats segons les categories mascle» i «femella» i la patologització dels cossos que no encaixen en el binomi—, la identitat de gènere —en què es pressuposa una correspondència entre sexes i identitat de gènere i es patologitzen i discriminen les identitats trans o no binàries—, la posició de gènere —basada en l'atribució d'una superioritat als homes i al masculí enfront de les dones i el femení, cosa que implica tota mena de desigualtats i violències—, l'expressió de gènere —en què s'associen formes d'expressió de rols masculins i femenins a les persones identificades com a homes i dones, respectivament, i s'estigmatitzen les que no encaixen en aquesta associació— i l'orientació sexual (…). (Rodó-Zárate, p. 44)
Father	P22	Male parent of the subject	Articles about women have less dissemination in multiple languages and their multilingual coverage is sometimes associated with the fact that depicted women had a relevant relationship with a notable man. (Beytía I Wagner, 2022, p. 17)

			Contemporary research on fatherhood describes fatherhood as a multifaceted, dynamic social, and cultural construction, deeply affected by class, race, and gender inequalities. (Perez-Vaisvidovsky, 2019)
Mother	P25	Female parent of the subject	Contemporary research on fatherhood describes fatherhood as a multifaceted, dynamic social, and cultural construction, deeply affected by class, race, and gender inequalities. (Perez-Vaisvidovsky, 2019)
Spouse	P26	The subject has the object as their spouse (husband, wife, partner, etc.)	The selection of notable women in Wikipedia –as opposed to the selection of men– seems to be correlated with the fact that they are married to someone also notable (Graells-Garrido et al., 2015). The latter could be a sign that female notability is sometimes subordinated to male notability processes. (Beytía I Wagner, 2022, p. 14) Articles about women have less dissemination in multiple languages and their multilingual coverage is sometimes associated with the fact that depicted women had a relevant relationship with a notable man. (Beytía I Wagner, 2022, p. 17) Finally, the comparison between Transnational and Western women offers three additional insights. First, the only event about private life which is salient for one of the two groups is 'married'. This indicates that private life events of women –in general– are always presented in relation to their conjugal status. (Stranisci et al., 2023, p. 8) 1960. Promulgación de la Ley de derechos políticos, profesionales y laborales de la mujer y del niño, que requiere a la mujer casada autorización del marido para trabajar. (*Genealogías*

			feministas en el arte español: 1960-2010, 2013, p. 11) El padre y el marido, figuras de autoridad incontestable, se erigían así como la hipóstasis del estado-patrón debiendo despertar en las mujeres de la familia admiración y respeto absolutos. (Tejeda, 2013, p. 202)
Country of citizenship	P27	The object is a country that recognizes the subject as its citizen	L'origen fa referència al lloc de naixement d'una persona i es relaciona amb processos de migració. En el context europeu, el sistema de control migratori es fonamenta sobre el colonialisme, que dona drets diferents a les persones segons el seu origen. (Rodó-Zárate, p. 45)
Continent	P30	Continent of which the subject is a part	L'origen fa referència al lloc de naixement d'una persona i es relaciona amb processos de migració. En el context europeu, el sistema de control migratori es fonamenta sobre el colonialisme, que dona drets diferents a les persones segons el seu origen. (Rodó-Zárate, p. 45) Several studies showed that geographical factors influence the topical distribution of content in Wikipedia language editions [13–20]. The geography gap means that some areas of the world are poorly represented in Wikipedia, Wikidata and the other sister projects [21]. (Miquel-Ribé I Laniado, 2021, p. 3)
Sexual orientation	P91	The sexual orientation of the person relative to their declared gender — use ONLY IF they have stated it themselves, unambiguously, or it has been widely agreed upon by	El sistema sexe–gènere té diferents dimensions que s'emmarquen en les normativitats derivades del sistema cisheteropatriarcal. Aquestes diferents dimensions inclouen el sexe (…), la identitat de gènere (…), la posició de gènere (…), l'expressió de gènere (…) i l'orientació sexual —en què s'estableix l'heterosexualitat com a norma i totes les

		historians after their death	altres formes de desig com a desviació. (Rodó-Zárate, p. 44)
Native language	P103	Language or languages a person has learned from early childhood	La llengua o idioma vehicular seria un altre eix vinculat a l'ús de la llengua primària o llengua vehicular, que pot suposar una font d'exclusió, en la mesura que pot no coincidir amb l'idioma oficial d'un territori o pot patir minorització per part d'una altra llengua. (Rodó-Zárate, p. 48)
Occupation	P106	Occupation of a person	One study found that women are, on average, slightly more notable than men in English Wikipedia, even controlling for occupation and year of birth (Wagner et al., 2016). (Beytía i Wagner, 2022, p. 14-15) Las ocupaciones a menudo están sujetas a estereotipos culturales y sociales. Algunos grupos pueden estar estigmatizados o marginados en ciertas profesiones, lo que afecta su visibilidad y representación en plataformas como Wikipedia. (Kalleberg, 2018, según cita Bejarano, 2023, p. 24)
Religion or worldview	P140	Religion of a person, organization or religious building, or associated with this subject	La diversitat ètnica, cultural i religiosa fa referència als costums, a la tradició i a la cultura d'una persona, i té implicacions també en relació amb les creences, les pràctiques religioses i l'imaginari col·lectiu. La violència, la desigualtat, la discriminació i l'aplicació d'estereotips negatius per raó d'origen, etnicitat, racialització o identitat religiosa tenen a veure amb el colonialisme, l'imperialisme, l'eurocentrisme i la supremacia blanca. (Rodó-Zárate, p. 46)
Award received	P166	Award or recognition received by a person, organization or creative work	[O]nly very prominent women are included in Wikipedia, while men have fewer access barriers, and therefore average a lower level of notability in

			different languages (Wagner et al., 2016). (Beytía i Wagner, 2022, p. 15)
Ethnic group	P172	Subject's ethnicity (consensus is that a VERY high standard of proof is needed for this field to be used. In general this means 1) the subject claims it themselves, or 2) it is widely agreed on by scholars, or 3) is fictional and portrayed as such)	La diversitat ètnica, cultural i religiosa fa referència als costums, a la tradició i a la cultura d'una persona, i té implicacions també en relació amb les creences, les pràctiques religioses i l'imaginari col·lectiu. La violència, la desigualtat, la discriminació i l'aplicació d'estereotips negatius per raó d'origen, etnicitat, racialització o identitat religiosa tenen a veure amb el colonialisme, l'imperialisme, l'eurocentrisme i la supremacia blanca. (Rodó-Zárate, p. 46)
Date of birth	P569	Date on which the subject was born	La societat també s'organitza en funció de l'edat de les persones, i atorga determinats drets, rols i expectatives a les persones segons la seva edat. L'edatisme és el conjunt de creences, normes i valors sobre diferents grups d'edat en els quals es basa la discriminació. En el nostre context, la infància, la joventut i la vellesa són els grups socials que pateixen discriminació per raó d'edat. Tot i que el terme *edatisme* s'acostuma a utilitzar per referir-se a la discriminació a les persones grans –també se'n diu gerontofòbia–, és important tenir present que els infants i les persones joves també en pateixen, ja sigui per qüestions com la negació de certs drets (a vot o a prendre decisions sobre el propi cos), o, en la joventut, la criminalització, l'estigmatització o la legitimació de la precarietat laboral, per exemple. (Rodó-Zárate, p. 47) A partir d'aquesta propietat i la següent, *date of death*, podré calcular l'edat en què la persona va morir o en què va fer

			una aportació important a la societat i que va fer que aquesta persona destaqués.
Date of death	P570	Date on which the subject died	La societat també s'organitza en funció de l'edat de les persones, i atorga determinats drets, rols i expectatives a les persones segons la seva edat. L'edatisme és el conjunt de creences, normes i valors sobre diferents grups d'edat en els quals es basa la discriminació. En el nostre context, la infància, la joventut i la vellesa són els grups socials que pateixen discriminació per raó d'edat. Tot i que el terme *edatisme* s'acostuma a utilitzar per referir-se a la discriminació a les persones grans –també se'n diu gerontofòbia–, és important tenir present que els infants i les persones joves també en pateixen, ja sigui per qüestions com la negació de certs drets (a vot o a prendre decisions sobre el propi cos), o, en la joventut, la criminalització, l'estigmatització o la legitimació de la precarietat laboral, per exemple. (Rodó-Zárate, p. 47) A partir d'aquesta propietat i l'anterior, *date of birth*, podré calcular l'edat en què la persona va morir o en què va fer una aportació important a la societat i que va fer que aquesta persona destaqués.
Languages spoken, written or signed	P1412	Language(s) that a person or a people speaks, writes or signs, including the native language(s)	La llengua o idioma vehicular seria un altre eix vinculat a l'ús de la llengua primària o llengua vehicular, que pot suposar una font d'exclusió, en la mesura que pot no coincidir amb l'idioma oficial d'un territori o pot patir minorització per part d'una altra llengua. (Rodó-Zárate, p. 48)
Time period	P2348	Time period (historic period or era, sports season, theatre season,	Concerning the results, Klein and Konieczny (2015) find that the strongest correlations are with individuals born

		legislative period etc.) in which the subject occurred	around 1910, indicating that Wikipedia's representation may more accurately reflect current rather than historical gender statuses. (Centelles I Ferran-Ferrer, 2024, p. 6) L'intersectionnalité analyse comment intergisent des types spécifiques, construits historiquement, de distributions inégales du pouvoir et/ou de normativités contraignantes, fondées sur des catégorisations socio-culturelles construites discursivement, institutionnellement et/ou structurellement telles que le genre, l'ethnicité, la race, la classe, la sexualité, l'âge ou la génération, le hándicap, la nationalité, la langue maternelle, etc, de sorte à produire différents types d'inégalités sociales. (Kóczé, 2011:134; citada por Expósito, 2012, p. 204)
Personal pronoun	P6553	Personal pronoun(s) this person goes by	El sistema sexe–gènere té diferents dimensions que s'emmarquen en les normativitats derivades del sistema cisheteropatriarcal. Aquestes diferents dimensions inclouen el sexe —classificació de les corporalitats segons les categories mascle» i «femella» i la patologització dels cossos que no encaixen en el binomi—, la identitat de gènere —en què es pressuposa una correspondència entre sexes i identitat de gènere i es patologitzen i discriminen les identitats trans o no binàries—, la posició de gènere —basada en l'atribució d'una superioritat als homes i al masculí enfront de les dones i el femení, cosa que implica tota mena de desigualtats i violències—, l'expressió de gènere —en què s'associen formes d'expressió de rols masculins i femenins a les persones identificades com a homes i dones, respectivament, i

			s'estigmatitzen les que no encaixen en aquesta associació— i l'orientació sexual (…). (Rodó-Zárate, p. 44)

Taula 2. Definició de les propietats seleccionades per a l'estudi a partir de Wikidata

D'aquesta manera, després del procés de *scraping*, quan puguem analitzar totes aquestes propietats en conjunt, podrem saber quina tipologia d'article apareix més, quan apareixen biografies, quines persones apareixen i quines són les seves característiques, de manera que podré estudiar la diversitat representativa de biografies a Wikipedia i fins i tot, en futures línies de treball, es podria analitzar com ha evolucionat la representativitat a la plataforma al llarg dels anys.

Cal destacar que, per a recollir les dades per a l'anàlisi, hem hagut de parar especial atenció a les propietats "instance of" i "sex or gender", atès que s'indica el sexe o gènere d'entitats humanes i d'altres entitats: animals o éssers humans ficticis, però també de manifestos (Q4315013), de ciclons (Q116784559), títols nobiliaris (Q11905347), temporades esportives (Q116455203), gratacels (Q22329554), entre altres entitats —fet que suggerim revisar, a fi que no s'adjudiquin valors i propietats a elements on no corresponen. Així doncs, ha calgut filtrar bé l'anàlisi per assegurar que només estudiéssim les propietats d'éssers humans.

Per dur a terme l'anàlisi, ens proposem utilitzar mètodes estadístics i exposar les dades mitjançant gràfics, de manera que es pugui representar visualment la diversitat a les portades. Per a fer l'anàlisi de les dades, utilitzarem principalment comandes de Unix, que ens serviran per saber quins són els resultats de cada propietat i quantes repeticions hi ha de cada resultat. En primer lloc, per crear els documents a partir del quals treballarem cada propietat, utilitzarem la comanda pico *nom-del-fitxer*. A continuació, farem servir la comanda sort < *nom-del-fitxer* | uniq -c | sort per ordenar alfabèticament els resultats, eliminar-ne els duplicats, saber les ocurrències de cada resultat i ordenar-los de més petit a més gran, de manera que puguem saber fàcilment quins són els principals resultats de cada propietat.

Un cop recollides totes aquestes dades, utilitzarem Excel per crear els gràfics corresponents a cada propietat, de manera que puguem tenir una representació visual dels resultats. Es podran veure les dades a partir de les quals s'elaboren els gràfics a l'annex 4.

Pel que fa a l'esquema de metadades mínimes necessàries, ens plantegem seguir l'estructura proposada per Wikidata i partir de l'esquema per descriure instàncies de la classe humà (Q5) (Entity Schema E1074), i afegir-hi les propietats necessàries per a escriure adequadament una biografia.

Així mateix, també convé comentar les principals limitacions de l'estudi. Per una banda, cal destacar que no s'han pogut recuperar totes les portades, ja que el repositori de la Wikipedia no té un històric de cada portada, de manera que no sabem quins articles van aparèixer en

portada alguns dies. Per altra banda, l'estudi només ha analitzat els resultats de l'edició en anglès de la Wikipedia, i per a fer un estudi més exhaustiu seria útil fer una anàlisi comparativa amb altres edicions, com l'espanyola o alguna altra de les més populars i fructíferes. També seria interessant analitzar els biaixos de contingut d'edicions més petites i regionals de la plataforma, com per exemple l'edició en català. Per últim, tampoc hem pogut analitzar totes les categories interseccionals, ja que n'hi ha moltes i algunes no encaixen exactament amb les propietats de Wikidata, fet que dificulta poder analitzar-les.

Revisió de la literatura

Ara que ja hem vist els objectius i la metodologia utilitzada per a dur a terme aquest estudi, en aquesta secció es presentaran els resultats de la revisió de la literatura basada en l'anàlisi de 16 publicacions. L'anàlisi es divideix en dues parts: una descriptiva i una de contingut. La primera, examina les autories, disciplines acadèmiques, afiliacions i localitzacions acadèmiques, l'evolució cronològica i la perspectiva de gènere pel que fa a les autories de les publicacions. La part de contingut se centra en les manifestacions i quantificació de la bretxa de gènere i contingut a la Wikipedia, causes i factors que contribueixen a aquesta bretxa, recomanacions i projectes per millorar la situació, així com altres biaixos trobats a l'enciclopèdia digital i les pàgines de Wikipedia que s'analitzen en les publicacions.

La revisió bibliogràfica d'aquest treball s'ha dut a terme mitjançant una *scoping review*; és a dir, una revisió de la literatura acadèmica que té com a objectiu trobar de manera sistemàtica la literatura disponible sobre una àrea temàtica concreta. Per a fer-la, hem seguit el mètode SALSA proposat per Grant i Booth (2009).

Estructura SALSA

<table>
<tr><td rowspan="12">SEARCH</td><td colspan="2">Q1: TITLE-ABS-KEY (Wikipedia AND “literature”)</td></tr>
<tr><td colspan="2">Q2: TITLE-ABS-KEY (Wikipedia AND “author*”)</td></tr>
<tr><td colspan="2">Q3: TITLE-ABS-KEY (Wikipedia AND (“front page section*” OR “main page section*”))</td></tr>
<tr><td colspan="2">Q4: TITLE-ABS-KEY (Wikipedia AND “section*”)</td></tr>
<tr><td colspan="2">Q5: TITLE-ABS-KEY (Wikipedia AND (“front page” OR “main page”))</td></tr>
<tr><td colspan="2">Q6: TITLE-ABS-KEY (Wikipedia AND “author*” AND (“bias” OR “disparity” OR “diversity” OR “difference”)</td></tr>
<tr><td colspan="2">Q7: TITLE-ABS-KEY (Wikipedia AND “author*” AND “gender gap”)</td></tr>
<tr><td colspan="2">Q8: TITLE-ABS-KEY (Wikipedia AND “gender gap”)</td></tr>
<tr><td colspan="2">Q9: TITLE-ABS-KEY (Wikipedia AND “intersectional*”)</td></tr>
<tr><td colspan="2">Q10: TITLE-ABS-KEY (Wikipedia AND “biograph*”)</td></tr>
<tr><td colspan="2">Nº inicial de documents: 2843
- WoS: 1588
- Scopus: 1255</td></tr>
<tr><td rowspan="2">APPRAISAL</td><td colspan="2">Supressió de duplicats (N=1037)</td></tr>
<tr><td>Criteris d’inclusió:
- Període de publicació (2018-2024) (N=737)</td><td>Criteris d’exclusió:
- Fora del període de publicació (N=300)
- Altres idiomes (N=24)</td></tr>
</table>

<table>
<tr><td rowspan="3"></td><td>Idiomes (català, castellà, anglès o francès) (N=713)
Tipus de document (article, congrés, tesi) (N=609)
Correspondència amb l'àrea temàtica (gènere, biografies, llengua, ètnia) (N=15)
S'utilitza Wikipedia com a pàgina d'estudi (N=15)</td><td>Tipus de document (ressenya, entrevista, altres) (N=104)
No hi ha correspondència temàtica (N=594)
S'utilitza Wikipedia com un corpus d'estudi (N=0)</td></tr>
<tr><td colspan="2">Documents exclosos: 1022
Documents inclosos a través d'altres fonts: 1</td></tr>
<tr><td colspan="2">Nº total de documents per analitzar: 16</td></tr>
<tr><td>SYNTHESIS</td><td colspan="2">Síntesis narrativa basada en les anàlisis realitzades</td></tr>
<tr><td>ANALYSIS</td><td colspan="2">Categories d'anàlisi:
Referència completa del document
Resum
Metodologia
Àmbit d'estudi
Objecte d'estudi
Objectius
Preguntes d'investigació
Resultats i principals aportacions
Limitacions
Futures línies d'investigació
Perspectiva de gènere
Components d'anàlisi interseccional</td></tr>
</table>

Taula 3. Estructura SALSA (Search, Appraisal, Synthesis, Analysis) per a revisió sistematitzada de la literatura.

Resultats de la *Scoping Review*

En aquesta secció, s'analitzaran els resultats de la revisió de la literatura (o *scoping review*) de les 16 publicacions treballades, que inclouen 11 articles, 4 ponències de conferències i 1 tesi doctoral. L'anàlisi es divideix en dues parts: la primera és descriptiva, mentre que la segona se

centra en el contingut. Cada part ofereix diferents perspectives d'anàlisi, aportant una visió completa de la recerca revisada.

Anàlisi descriptiva

1. Autories

Hem analitzat 16 documents amb contribucions de 33 autories. La majoria de les persones autores (29) han contribuït en un sol document, mentre que 4 persones autores han treballat en múltiples documents. Concretament, 2 autories han contribuït a 2 documents cadascuna i 2 altres autories han contribuït a 3 documents cadascuna.

Pel que fa al gènere dels autors, hi ha 13 autores dones i 20 autors homes, i cap persona no binària. Les dones estan representades de manera equitativa com a primeres autores (8), mentre que els homes apareixen més freqüentment com a segons (7), tercers (6), cinquens (1) i sisens (1) autors. La composició dels grups varia amb 5 grups formats només per homes, 1 grup només per dones i 6 grups mixtos. Cal destacar que no hi ha cap article escrit únicament per un home, mentre que 4 articles estan escrits per una sola dona.

2. Disciplines acadèmiques de les persones autores

En examinar les disciplines acadèmiques de les persones autores, la Informàtica (Computer Science) ha resultat ser el camp amb més representació, amb 16 persones autores. A continuació, es troba la Ciència de la Informació (Information Science) amb 13 autories, les Ciències Socials (Social Science) amb 9, les Humanitats (Humanities) amb 8, i les Ciències de l'Educació (Educational Science) amb 5 autories. Addicionalment, la disciplina Mitjans de Comunicació (Media and Communications) està representada amb 2 persones autores, mentre que tant la Sociologia (Sociology) com la Psicologia (Psychology) tenen 1 persona autora cadascuna. Aquests resultats són interessants, ja que subratllen la naturalesa interdisciplinària de la recerca sobre Wikipedia i el biaix de contingut i de gènere.

3. Afiliacions acadèmiques i localitzacions de les persones autores

Les afiliacions de les persones autores a universitats o centres d'investigació es distribueixen per diversos països. Espanya lidera amb 8 autories, seguida d'Itàlia amb 6, Iran amb 4, Alemanya, els Estats Units i Polònia amb 3 cadascuna, l'Índia amb 3, França amb 2, i Corea del Sud amb 1 persona autora. En la majoria de publicacions, les persones autores estan afiliafdes al mateix país, però no sempre és així. El cas de Piotr Konieczny (Hanyang University, South Korea) i Maximilian Klein (University of Minnesota, USA), autors de "Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator", n'és un exemple. Aquesta diversitat geogràfica també és una dada interessant, ja que subratlla l'interès global i la col·laboració en aquest camp de recerca.

4. Anàlisi cronològica

Els documents abasten un període des del 2018 fins al 2024. Hi ha 1 article de cada un dels anys 2018, 2019 i 2020. L'any 2021 hi va haver un lleuger augment amb 3 articles publicats, seguit d'un increment significatiu el 2022, amb 6 articles. El 2023 es van publicar 3 dels articles analitzats i aquest any 2024 se n'ha publicat un altre. Aquesta tendència indica un creixent interès i producció en els darrers anys, la qual cosa pot estar relacionada amb l'objectiu del Wikimedia Movement de "alcanzar la equidad de conocimiento, o *knowledge equity*, para el año 2030" (Sefidari, 2022, p. 147; Miquel-Ribé i Laniado, 2020, p. 1).

5. Perspectiva de gènere

Pel que fa a la perspectiva de gènere, 8 articles analitzen el gènere des d'una perspectiva binària, i quatre articles adopten una perspectiva no-binària, reflectint una comprensió més àmplia del gènere. I en els 4 articles restants no s'estudia el biaix de gènere, sinó que s'analitza la diversitat general en diferents edicions lingüístiques o es presenten projectes i mètodes per avaluar la diversitat de contingut i quantificar-ne la bretxa. Aquests resultats suggereixen un enfocament divers a l'anàlisi de gènere i de contingut en la recerca.

Anàlisi de continguts

1. Manifestacions de la bretxa

L'evolució històrica de les enciclopèdies revela canvis significatius en la representació de les dones, especialment a partir de finals del segle XIX. Tot i aquestes millores, les desigualtats històriques en la visibilitat i la inclusió de dones en les biografies persisteixen. Aquestes desigualtats, resultat de patrons editorials i culturals que han dominat les narratives enciclopèdiques tradicionals, també són presents a Wikipedia (Konieczny i Klein, 2018).

La Wikipedia, des de la seva fundació, ha tingut un impacte profund en la democratització del coneixement. Com a plataforma col·laborativa i accessible, ha trencat amb els models tradicionals d'enciclopedisme, permetent una major participació de les persones usuàries en la creació i edició del contingut. No obstant això, aquesta democratització també ha exposat la plataforma a biaixos preexistents en la societat, que es reflecteixen en la bretxa de gènere i de contingut de l'enciclopèdia virtual (Meyer, 2022).

La bretxa de gènere i contingut a Wikipedia es manifesta a través d'una sèrie de dimensions que reflecteixen i perpetuen desigualtats sistemàtiques. La literatura acadèmica revisada testimonia l'àmplia reconeixença de la bretxa de gènere a Wikipedia, subratllant la urgència de superar els biaixos i les barreres per aconseguir una comunitat més inclusiva. La revisió de la literatura de Ferran-Ferrer et al. (2023), que abasta 97 publicacions d'entre el 2007 i el 2022 també ho demostra.

Els biaixos de gènere no només es tradueixen en una menor quantitat de cobertura de biografies de dones, sinó també en la qualitat i la notabilitat del contingut. Com hem comentat, les pàgines sobre dones tenen més probabilitats de ser qüestionades i esborrades si no compleixen estrictament amb els criteris de notabilitat establerts, contribuint a una representació disminuïda i més precària (Sefidari, 2022). A més, la jerarquia heteronormativa de la plataforma marca les biografies de dones amb un llenguatge que denota implícitament que les persones notables són homes, a menys que s'indiqui el contrari, perpetuant un biaix subtil però persistent (Tripodi, 2021).

A més, l'anàlisi de les biografies destacades en la Portada de l'edició en anglès de Wikipedia revela una clara segregació per gènere, raça i ocupació, demostrant que el biaix de contingut va més enllà del gènere (l'eix d'opressió més estudiat fins ara), sinó que també afecta altres eixos identitaris (Sefidari, 2022). Les biografies d'homes, majoritàriament de raça blanca i en àmbits com la política i l'esport, tenen una presència dominant, mentre que les de dones, més freqüents en àmbits artístics i literaris, estan significativament menys representades (Sefidari, 2022). L'anàlisi de més de 48789 biografies a Wikipedia duta a terme per Stranisci et al. (2023) també destaca biaixos significatius en la representació interseccional d'ètnia i gènere, contradient la pretensió d'objectivitat de la plataforma.

Un altre aspecte a tenir en compte és la selecció de contingut, que exhibeix patrons de centralitat i perifèria, on les biografies "centrals" –és a dir, d'interès generalitzat arreu del món–, amb una major difusió multilingüe i millors referències, tendeixen a ser dominades per figures masculines. Les biografies de dones, en canvi, ocupen sovint una posició més perifèrica en la xarxa de referència, reflectint una menor centralitat en la plataforma (Beytía i Wagner, 2022).

Les visualitzacions interactives desenvolupades per Miquel-Ribé i Laniado (2020, 2021) també revelen desequilibris substancials en temes i conceptes entre les diferents edicions lingüístiques de Wikipedia, identificant mancances significatives en categories com geografia, gènere i temes LGBT+. Lewoniewski et al. (2019) i Roy et al. (2021) també mostren com la qualitat i la popularitat dels articles varien significativament entre temes i idiomes: els articles sobre temes populars en una determinada llengua tendeixen a tenir una qualitat relativa més alta, i malgrat la presència de temes similars en diverses edicions, es troben diferències significatives en el detall i la cobertura dels subtemes.

2. Quantificació de la bretxa

La baixa representació de dones a la Wikipedia és un fenomen globalment reconegut i acceptat. Segons indiquen alguns estudis, només entre el 13,2% i el 22,5% de les biografies de l'enciclopèdia virtual són sobre dones, subratllant una representació modesta que repercuteix en la diversitat d'informació disponible (Ferran-Ferrer et al., 2023; Fan i Gardent, 2022). Maria Sefidari (2022), en l'anàlisi que fa de les biografies destacades de la Portada de Wikipedia en anglès durant l'any 2019, posa de manifest que el 80,2% de les biografies destacades eren d'homes, mentre només el 19,8% eren de dones. Aquesta segregació també es veu reflectida

per raça i ocupació, amb una major presència masculina en àmbits de poder i una representació femenina més marcada en àmbits artístics i literaris (Sefidari, 2022).

En les discussions sobre la notabilitat de les biografies, les de dones sovint reben un escrutini més intens i prolongat, i acostumen a ser nominades per supressió més ràpida o freqüentment que les d'homes (Tripodi, 2021). Tot i això, les persones autores semblen arribar al consens que no són suprimides més freqüentment que les d'homes, la qual cosa indica una major contestació inicial de la seva inclusió, amb una necessitat de defensa més àmplia per part dels seus partidaris (Martini, 2023; Beytía i Wagner, 2022). Així, la principal preocupació radica en la selecció inicial de contingut, on les dones poden estar subrepresentades (Beytía i Wagner, 2022).

A més, els articles sobre dones, tot i tenir una longitud igual o superior als dels homes, sovint inclouen més informació sobre el gènere i les relacions familiars, mentre que reben menys suport visual i cobertura quan les dones destaquen en àmbits com les ciències, les humanitats o la tecnologia (Beytía i Wagner, 2022).

La subrepresentació transnacional de les dones també és evident, amb una menor quantitat de biografies sobre dones transnacionals respecte als homes en comparació amb persones de raça blanca occidental (Stranisci et al., 2023). Quant a la diversitat cultural i lingüística, les biografies de dones varien significativament entre diferents cultures i edicions lingüístiques de Wikipedia, amb algunes regions mostrant proporcions més altes de biografies de dones que altres (Miquel-Ribé i Laniado, 2021; Konieczny i Klein, 2018).

Finalment, la falta de correspondència entre les edicions no angleses i la Wikipedia en anglès també contribueix a la disparitat en el contingut disponible, amb aproximadament el 60% dels articles en altres llengües mancant d'una versió equivalent en anglès (Roy et al., 2021).

3. Causes i factors que contribueixen a la bretxa de gènere i de contingut a la Wikipedia

La bretxa de gènere i de contingut a Wikipedia és àmpliament reconeguda a l'àmbit acadèmic (Ferran-Ferrer et al., 2023). Aquest fenomen complex deriva de diversos factors interrelacionats. Segons conclouen alguns estudis, es poden identificar diverses causes principals que contribueixen a aquesta desigualtat.

Alguns estudis atribueixen el problema de la bretxa de gènere a tres factors (Ferran-Ferrer et al., 2023): el "problema de les dones", que consisteix en atribuir el biaix representatiu a les característiques intrínseques de les dones, simplificant així la complexitat del problema; l'"efecte mirall", que planteja la bretxa com un reflex de les dinàmiques socials preexistents, de manera que treu responsabilitat a l'enciclopèdia, i el "problema sistèmic", que consisteix en atribuir la responsabilitat principal de la situació a la dominància masculina entre les persones editores.

Les polítiques i pràctiques de Wikipedia, com els criteris estrictes de notabilitat i la selecció de contingut, també contribueixen a la bretxa. Les biografies de dones sovint s'enfronten a un escrutini més rigorós i a un major risc de ser considerades "no notables" i ser susceptibles d'eliminació, malgrat complir amb els mateixos criteris que les biografies d'homes (Martini, 2023; Tripodi, 2021). Això afecta principalment al biaix de contingut sobre dones (també conegut com a *gender content gap*), on s'ha detectat una baixa cobertura temàtica de les biografies de dones. A més, la manca d'informació accessible sobre dones també contribueix a la persistència de la bretxa, ja que per crear-ne articles s'han d'aportar fonts que n'atestin la notabilitat, i no sempre són fàcils de trobar (Fan i Gardent, 2022; Meyer, 2022). A més, les directrius de notabilitat, que exigeixen fonts externes i neutralitat, suposen un repte especial per a les biografies de dones que no sempre reben la mateixa cobertura mediàtica que els homes (Ferran-Ferrer et al., 2022). Meyer (2022) analitza les polítiques de fiabilitat, de prohibició de la investigació original i de notabilitat per comprendre com poden afectar la creació de coneixement i com poden accentuar la bretxa de gènere i altres biaixos sistèmics, com la subestimació de les tradicions orals i indígenes.

Altres factors inclouen els patrons de selecció de contingut basats en la centralitat dels articles, la participació limitada i la diversitat entre les persones editores, que prioritzen les seves pròpies llengües i coneixements culturals (Beytía i Wagner, 2022; Roy et al., 2021). Les preferències de les persones editores i els biaixos culturals també influeixen en les diferències observades en la cobertura de subtemes entre les diferents edicions lingüístiques. N'és un exemple la cobertura de temes de l'edició en anglès en contrapunt amb les altres edicions lingüístiques: aproximadament el 60% dels articles en edicions no angleses manquen d'una versió en anglès corresponent, tot i que la Wikipedia anglesa sovint ofereix una cobertura més àmplia de temes (Roy et al., 2021). Això contribueix a una disparitat en el contingut i reflecteix la importància de les edicions locals en proporcionar informació detallada i específica sobre temes que poden estar més ajustats a les realitats culturals i locals, i posa de relleu la importància de compartir contingut entre les diferents edicions per poder mitigar les disparitats de cobertura en funció de l'edició lingüística.

L'ús del llenguatge de gènere o sexista en les biografies de dones és un altre factor que perpetua la bretxa de gènere a Wikipedia. Aquest llenguatge manté una jerarquia heteronormativa i es perpetua remarcant informació sobre el gènere o les relacions familiars de les dones més que les seves contribucions notables, a diferència de les biografies d'homes (Beytía i Wagner, 2022). Aquesta pràctica reforça estereotips de gènere i contribueix a la subrepresentació de les dones, afectant la percepció pública sobre la importància i el valor de les seves contribucions.

Un altre factor important que contribueix a la bretxa de gènere i contingut a Wikipedia és l'estructura de les categories. Les categories principals varien significativament entre les diferents versions lingüístiques de la Wikipedia, on cada idioma té la seva pròpia definició i jerarquia de categories. Aquesta variabilitat pot crear inconsistències i dificultats en la representació equitativa de gèneres i continguts, perpetuant els biaixos (Lewoniewski et al., 2019). Segons indiquen Centelles i Ferran-Ferrer (2024), les categories de Wikipedia sovint

estan dissenyades per facilitar el treball de les persones editores, però no sempre satisfan les necessitats d'informació de les persones usuàries. Les categories i les etiquetes utilitzades també poden ser dissenyades principalment per facilitar la feina de les persones editores i indexadores, en lloc de respondre directament a les necessitats i interessos de les persones usuàries que consulten la informació a la plataforma, la qual cosa pot contribuir a la bretxa de gènere i contingut, ja que les categories poden no reflectir adequadament la diversitat dels temes i les identitats representades.

A més, des d'una perspectiva interseccional, es critica la falta d'inclusió de les identitats de gènere com a criteri de categorització a Wikipedia. Aquesta manca d'inclusió porta a incoherències, com l'ús de categories de gènere femení que no reflecteixen la diversitat de gènere. La falta de reconeixement de diverses identitats de gènere en la categorització contribueix a la subrepresentació i al biaix de gènere a la plataforma (Centelles i Ferran-Ferrer, 2024). En contrast amb Wikipedia, Wikidata mostra un nivell més alt de sensibilitat cap a la diversitat de gènere, amb la inclusió de diverses categories de gènere a través de la propietat P21, segons indiquen les persones autores de l'estudi. Tot i això, recomanen fer una distinció clara entre sexe biològic i identitat de gènere per millorar l'etiquetatge, i apunten que aquesta major sensibilitat per la diversitat de gèneres a Wikidata podria servir com a model per a Wikipedia en la seva aproximació a la diversitat de gènere.

Centelles i Ferran-Ferrer (2024) també conclouen que els processos de presa de decisions a Wikipedia i Wikidata contribueixen a la bretxa de gènere i contingut. Mentre que Wikidata tendeix a enfocar-se en arguments tècnics i centrats en les dades, Wikipedia sovint es veu immersa en debats més socioculturals. Aquesta diferència en l'enfocament pot influir en com es resolen les qüestions de diversitat i representació a cada plataforma (Centelles i Ferran-Ferrer, 2024). I finalment, assenyalen que Wikidata també s'enfronta a reptes lingüístics, especialment en llengües amb diferenciació de gènere (com el català), atès que la sensibilitat cultural i lingüística és crucial per mantenir l'exactitud en les dades, i qualsevol falta en aquesta àrea pot contribuir a biaixos en la representació de gèneres i continguts.

4. Recomanacions per millorar la problemàtica

Les recomanacions per abordar la bretxa de gènere i de contingut a Wikipedia se centren en estratègies diverses a llarg termini. Sobretot es destaca la necessitat de crear espais segurs, programes de tutoria i promoció de la paritat de gènere entre les persones editores, així com la participació dels mitjans de comunicació per millorar la representació femenina (Ferran-Ferrer et al., 2023).

Per corregir els biaixos existents, algunes autories proposen anar més enllà de les propostes tradicionals de col·laborar amb comunitats dedicades a mitigar el biaix de grups marginalitzats. Segons diuen, és necessari un canvi estructural més profund que asseguri una representació equitativa i diversa en els articles destacats (Sefidari, 2022; Tripodi, 2021; Stranisci et al., 2023). Investigadors com Miquel-Ribé i Laniado (2020, 2021) i Beytía i Wagner (2022) han

elaborat models conceptuals o eines com taules de comparació i visualitzacions que mostren buits en conceptes no representats o compartits entre idiomes, ajudant així les persones editores a abordar les llacunes de diversitat de contingut.

Un element fonamental per augmentar la diversitat del contingut és la participació de la comunitat *wikipedista*. Les organitzacions comunitàries han fet esforços per abordar la subrepresentació de les dones a les biografies i han proposat estratègies per millorar la igualtat de gènere a la plataforma (Meyer, 2022). Col·laborar amb iniciatives com WikiWomen, WikiAfrica i grups de persones usuàries dedicats a mitigar el biaix de contingut pot millorar la representació de diversos grups i temes (Miquel-Ribé i Laniado, 2021). Promoure la creació de continguts inclusius i diversos, que abordin temes de gènere, geografia, temes LGBT+ i grups ètnics, també ajudarà a equilibrar la representació del coneixement global a Wikipedia, segons indiquen les mateixes persones autores.

També es recomana l'ús de perspectives multidimensionals, com el ciberfeminisme, per abordar les asimetries de contingut de gènere a Internet, de manera que es poden reconèixer les múltiples capes de desigualtat i com interactuen al llarg del temps (Sefidari, 2022; Beytía i Wagner, 2022). A més, s'aconsella adoptar un enfocament interseccional per millorar la representació de dones i persones no occidentals, combatre els biaixos de gènere i racials, i així millorar la representativitat i igualtat a la plataforma (Stranisci et al., 2023).

Adoptar una perspectiva de gènere en la creació de continguts i en la participació de les persones editores és essencial per assegurar que Wikipedia reflecteixi amb precisió la diversitat de la societat i elimini estereotips. Així mateix, revisar els criteris de notabilitat per a les biografies de dones, utilitzant fonts alternatives o noves proves per avaluar la seva importància, és crucial. També és important comprendre millor la retenció de les dones editores per identificar punts crítics d'abandonament i fomentar la seva participació continuada (Ferran-Ferrer et al., 2022).

Per millorar la representació de gènere, cal atendre les polítiques de notabilitat, incloure més fonts diverses i adoptar pràctiques que promoguin una representació més equitativa del coneixement i de les persones col·laboradores, segons apunten Konieczny i Klein (2018). A més, també convé escoltar les experiències personals de les persones editores, que coneixen els desafiaments en la creació de contingut sobre dones i poden proposar solucions per mitigar aquest biaix, com millorar la inclusió i la qualitat del contingut, i destacar la importància de la representació diversa a la plataforma –Meyer (2022) n'és un exemple. Així mateix, es destaca la importància de trobar fonts fiables, encara que sovint difícils d'accedir, i el paper crucial de les discussions comunitàries en la millora del coneixement compartit a Wikipedia (Meyer, 2022).

Finalment, es recomana una revisió integral del sistema d'organització de continguts de Wikipedia per incloure i millorar la representació de la diversitat de gènere, reconeixent el potencial de les categories d'identitat de gènere per millorar la cerca i la recuperació

d'informació per a les persones usuàries (Centelles i Ferran-Ferrer, 2024). És essencial una solució que combini millores tecnològiques amb consideracions culturals per abordar eficaçment els biaixos de gènere i millorar l'organització dels continguts en aquestes plataformes de coneixement (Centelles i Ferran-Ferrer, 2024).

5. Projectes i propostes per millorar la qüestió

Durant els últims anys, s'han desenvolupat eines com taules de comparació, suggeriments d'articles prioritaris i visualitzacions per ajudar les persones editores a identificar i abordar les llacunes de diversitat de contingut entre diferents edicions lingüístiques de Wikipedia (Lewoniewski et al., 2019; Miquel-Ribé i Laniado, 2020, 2021). Aquestes eines són útils per analitzar categories de diversitat com cultura, gènere i ubicació per classificar els articles existents i aplicar aprenentatge automàtic per suggerir articles prioritaris (Miquel-Ribé i Laniado, 2020). Utilitzar taulers interactius per mostrar desequilibris en temes i conceptes entre les edicions lingüístiques de Wikipedia ajuda les persones editores a identificar buits de contingut, especialment en categories com geografia, gènere, temes LGBT+ i grups ètnics (Miquel-Ribé i Laniado, 2021). Classificar els articles segons aquests temes facilita la identificació de mancances i la creació de continguts més diversos (Miquel-Ribé i Laniado, 2020, 2021).

Els projectes de Miquel-Ribé i Laniado (2020, 2021) tenen un impacte significatiu en els objectius estratègics de Wikimedia per a l'equitat del coneixement, atès que proporcionen un marc tècnic exhaustiu per avaluar desequilibris de contingut a les 309 edicions de la Wikipedia. A més, també introdueixen innovacions en l'àmbit de les Humanitats Digitals per fomentar la diversitat de continguts en la producció entre iguals.

Una altra aportació interessant en el camp és la generació de biografies secció per secció, que han millorat la qualitat de les biografies generades augmentant la precisió en la informació específica de cada secció (Fan i Gardent, 2022). La introducció de mòduls de recuperació i mecanismes de memòria ha millorat significativament la qualitat dels textos generats, però, de totes maneres, encara cal desenvolupar més aquests nous mètodes per a que siguin més inclusius, atès que, fins al moment, la generació automàtica de biografies s'ha utilitzat per pal·liar la bretxa de gènere, però encara no s'ha estudiat com aplicar-ho per reduir les altres bretxes de contingut (Fan i Gardent, 2022).

I un altre projecte dedicat a mitigar el biaix de contingut és WikiBio, desenvolupat per Stranisci et al. (2023), un nou conjunt de dades anotat per a la detecció d'esdeveniments biogràfics. Aquest projecte permet explorar i comparar les formes en què es narren i es representen les vides de les persones a Wikipedia, especialment en relació amb els eixos interseccionals d'ètnia i de gènere, i, per tant, analitzar el funcionament dels biaixos de contingut en relació amb aquests dos eixos d'opressió.

Per últim, el marc conceptual proposat per Fahimnia et al. (2022) representa una aportació interessant per avaluar la qualitat de la informació continguda als articles de la Wikipedia a través d'una estructura detallada i sistemàtica. Aquest marc conceptual, desenvolupat mitjançant una revisió exhaustiva de la literatura i la integració de dimensions i característiques rellevants –la credibilitat, la popularitat, la completesa, la coherència, la comprensibilitat i l'actualitat, entre altres factors–, ofereix una eina objectiva i estructurada per a les persones usuàries per determinar la fiabilitat i precisió de la informació de Wikipedia. La seva funcionalitat principal radica en la capacitat d'avaluar la qualitat de la informació en diversos nivells i des de diverses perspectives, permetent a les persones usuàries identificar contingut d'alta qualitat sense la necessitat de recórrer a especialistes externs. Així, aquest marc conceptual no només facilita la presa de decisions informades per les persones usuàries, sinó que també pot servir com a guia per a les persones editores i contribuents de la Wikipedia per millorar la qualitat del contingut que generen.

Així mateix, la participació de la comunitat de Wikipedia també és fonamental per augmentar la diversitat del contingut. Per això al llarg dels anys s'han creat comunitats i s'han impartit tallers i "Edit-a-thons" per promoure la creació de nous continguts. Col·laborar amb iniciatives com WikiWomen, WikiAfrica i grups de persones usuàries minoritzades ajuda a millorar la representació de diversos grups i temes (Miquel-Ribé i Laniado, 2021; Tripodi, 2021; Meyer, 2022; Martini, 2023; Ferran-Ferrer et al., 2023).

6. Impacte, conseqüències i implicacions de la bretxa

Els biaixos de contingut, com ara els de gènere, raça i país, són factors significatius en la qualitat de l'enciclopèdia virtual. Aquests biaixos distorsionen la representació històrica i perpetuen estereotips a través de les biografies de Wikipedia, una dinàmica que contribueix a eliminar els grups minoritzats de la història (Tripodi, 2021). La visibilitat a la Portada de Wikipedia amplifica els biaixos existents, influint en la percepció pública sobre els rols de gènere, raça i ocupació (Sefidari, 2022), i la representació parcial provoca tensions i conflictes quan es busca incloure contingut sobre persones que intersequen amb feminisme, comunitats LGBT+ i altres grups marginalitzats. Això sovint enfronta persones editores que mantenen estructures hegemòniques de classificació del coneixement amb aquelles que busquen una representació més inclusiva i diversa, desafiant la continuïtat dels biaixos estructurals (Sefidari, 2022).

A més, la subrepresentació de les dones a Wikipedia té implicacions que van més enllà de la mateixa plataforma. Aquesta falta de representació influeix en la percepció pública, donant lloc a un biaix de gènere en la manera com es veuen i valoren les contribucions femenines a la societat (Tripodi, 2021). Així mateix, aquesta bretxa té un impacte en l'ensenyament dels sistemes d'intel·ligència artificial (IA) i altres àmbits, ja que aquests sistemes sovint es basen en dades disponibles en plataformes com Wikipedia. Per tant, la subrepresentació de dones i altres grups marginats en aquests conjunts de dades pot perpetuar els biaixos en la informació que es distribueix a nivell global (Tripodi, 2021).

7. Altres biaixos (interseccionalitat)

L'anàlisi de les biografies destacades a la portada de Wikipedia revela una representació mòdica de gènere i raça. Durant 2019, només el 19,8% de les biografies destacades eren de dones, mentre que el 80,2% eren d'homes. A més, la majoria de les persones representades (86 de 96) eren de raça blanca, amb una presència molt limitada de persones de raça negra, de les Illes del Pacífic i d'altres races, segons determina Sefidari (2022). També s'observa una segregació per ocupacions, on els homes predominen en àmbits de poder com la política i els esports, mentre que les dones són més visibles en àmbits artístics i literaris (Sefidari, 2022).

Pel que fa a l'ètnia, l'estudi de Stranisci et al. (2023) mostra biaixos significatius de representació basats en l'ètnia i el gènere, evidenciant que les disparitats en els tipus i freqüències d'esdeveniments estan influïdes per aquests factors, segons indiquen. En particular, troben que les dones transnacionals estan notablement menys representades en comparació amb els homes transnacionals i occidentals. Segons conclouen, l'enfocament interseccional és crucial per identificar i abordar adequadament aquests biaixos, mostrant com el biaix de gènere esdevé més evident quan es comparen biografies dins del grup ètnic ampli.

Les biografies de persones LGBT+ també pateixen una discriminació similar, amb una major probabilitat de ser marcades com a no notables malgrat complir amb els criteris d'inclusió, indica Tripodi (2021). Aquesta discriminació reforça els biaixos existents i limita la representació d'aquests grups a la plataforma.

Finalment, Konieczny i Klein (2018) identifiquen variacions significatives en la representació de gènere a les biografies de Wikipedia segons la cultura i la llengua. Per exemple, cultures confucianistes i del sud d'Àsia mostren proporcions més altes de biografies de dones, indicant que els biaixos de gènere també estan influïts per contextos culturals i lingüístics específics.

Segons Sefidari (2022), «Wikipedia podría ser una extraordinaria oportunidad para superar estereotipos y sesgos, o bien un enorme riesgo a la hora de amplificarlos o transformarlos en otros sesgos» (p. 148). L'estudi dels biaixos de representació a Wikipedia necessita una perspectiva interseccional per tal de captar la complexitat de les desigualtats existents. Aquesta perspectiva ha de considerar una gran varietat de factors, incloent-hi gènere, ètnia, orientació sexual, i altres eixos d'identitat que influeixen en la manera com les persones es descriuen a si mateixes i són representades (Roy et al, 2021).

En alguns dels articles revisats s'aplica aquesta perspectiva interseccional i s'analitzen altres factors a part del gènere: cultura, ètnia o raça, país de procedència i sexualitat (o col·lectiu LGTB+ a l'engròs) són els més habituals (Sefidari, 2022; Miquel-Ribé i Laniado, 2020, 2021; Stranisci et al., 2023; Tripodi, 2021). Ara bé, només Stranisci et al. (2023) fan palesa la voluntat d'aplicar aquesta perspectiva en el seu estudi («[i]n this section we provide an analysis of writers' biographies on Wikipedia adopting intersectionality as a theoretical framework» [p.

12375]). Els altres, si bé analitzen dos o més eixos d'opressió, no manifesten la interseccionalitat com a marc teòric o metodològic a partir del qual articulen l'anàlisi.

De totes maneres, sí que podem trobar afirmacions en què les autories indiquen que convindria estudiar fins a quin punt els patrons interseccionals d'opressió es tenen en compte en com s'avalua la notabilitat d'un subjecte, la relació entre gènere i origen (Stranisci et al., 2023), o en què proposen un marc conceptual per analitzar la intersecció entre gènere, etnicitat, geografia o religió (Miquel-Ribé i Laniado, 2021).

Partint de la voluntat d'analitzar la intersecció d'eixos, els investigadors aborden les seves anàlisis mitjançant diferents metodologies. Per una banda, alguns estudis utilitzen dades quantitatives per analitzar les bretxes de gènere, raça i orientació sexual en les biografies de Wikipedia a partir de les dades diàries dels articles destacats de la plataforma (Sefidari, 2022; Miquel-Ribé i Laniado, 2020 i 2021; Tripodi, 2021). Això inclou la recopilació de dades sobre la representació de diferents grups i l'anàlisi estadística per identificar patrons i disparitats. Per altra banda, altres estudis adopten un enfocament qualitatiu, examinant casos específics de biografies per entendre millor com es manifesten els biaixos i com afecten la representació de grups marginats (Stranisci et al., 2023). Aquest enfocament pot incloure entrevistes amb persones editores de Wikipedia o l'anàlisi de les discussions i polítiques de la comunitat, per exemple (Meyer, 2022; Tripodi, 2021).

A més, alguns investigadors han desenvolupat eines digitals per estudiar la relació entre diferents factors. N'és un exemple el Wikipedia Diversity Observatory, desenvolupat per Miquel-Ribé i Laniado (2020, 2021). Aquest projecte estudia la relació d'eixos interseccionals a partir d'una base de dades que categoritza els articles de totes les edicions lingüístiques de Wikipedia segons diferents tipus de bretxes, i permet comparar una entitat específica (com una entitat geogràfica, de gènere o cultura) amb un grup d'edicions lingüístiques o un grup d'entitats dins d'una única edició lingüística de Wikipedia. D'aquesta manera, el projecte permet visualitzar les bretxes de contingut i com es relacionen i influeixen les unes a les altres, aportant més profunditat a l'anàlisi del biaix de contingut, a part de proporcionar recomanacions concretes per a les persones editores sobre com contribuir a revertir la situació.

8. Pàgines analitzades

En quant a les pàgines més estudiades de Wikipedia, en els articles analitzats trobem diverses pàgines d'interès. Sefidari (2022), analitza la Portada de l'edició anglesa de l'enciclopèdia. Per a fer-ho, se centra en els articles destacats que apareixen a la Portada al llarg del 2019, ja que «[el] Artículo Destacado del Día es el contenido más prominente para cualquier persona que entre a Wikipedia por la página principal, la más visitada de la enciclopedia» (Sefidari, 2022, p. 147). En canvi, Beytía i Wagner (2022) i Meyer (2022) analitzen les "Talk pages"; ara bé, cadascuna de les persones autores n'estudia un aspecte diferent. Els primers, se centren en la capacitat de traçar l'evolució dels articles i veure'n els comentaris que les persones editores fan sobre els articles, les discussions sobre canvis editorials o desacords entre persones

col·laboradores a partir d'una anàlisi qualitativa. La segona, en canvi, se centra en l'assetjament que reben les dones o les persones trans (i, possiblement, les d'altres col·lectius oprimits) en aquestes pàgines de discussions, destacant la dificultat que suposa a vegades haver de navegar per un entorn hostil i on se t'ataca per la teva identitat. Per a fer l'anàlisi, parteix de la seva pròpia experiència i la d'altres persones que han viscut situacions similars a la seva, que relata i comenta al llarg de l'estudi.

Per altra banda, Beytía i Wagner (2022), Martini (2023), Tripodi (2021) i Meyer (2022) analitzen els articles nominats per ser suprimits, coneguts amb les sigles AfD (Article for Deletion). Beytía i Wagner (2022) comenten que els estudis realitzats sobre la disparitat de gènere dels articles nominats a ser suprimits no han trobat desigualtat entre les biografies sobre dones que les d'home, però ells no duen a terme la seva pròpia anàlisi, sinó que es basen en estudis previs. És a dir, les persones autores plantegen que no hi ha biaix de gènere en els AfD. Martini (2023), que focalitza el seu estudi en les biografies nominades per supressió de l'edició en alemany de Wikipedia a través d'un doble enfocament qualitatiu i quantitatiu, considera que sí que hi ha biaix, que els articles sobre dones acostumen a nominar-se per supressió més freqüentment que els d'homes. Tanmateix, segons conclou, els articles sobre dones també es mantenen més sovint després de les discussions sobre la supressió, fet que indica que les dones realment notables es posen en qüestió il·legítimament. Tripodi (2021), que analitza les biografies de dones de l'edició en anglès de la Wikipedia, també pensa com Martini, i es proposa demostrar que les biografies sobre dones que compleixen els criteris d'inclusió de la Wikipedia es consideren no destacables i són nominades per supressió més freqüentment en comparació amb les biografies d'homes mitjançant un estudi estadístic. A més, l'autora afirma que «[r]esearch on deletions reveals that the most frequently used rationale for deleting an article was that it had "no indication of importance" (Geiger and Ford, 2011: 201; Lam and Read, 2009), and deletions due to a non-notability classification have increased over time (Lam and Read, 2009)» (Tripodi, 2021, p. 1690). Finalment, Meyer (2022) també analitza el fenomen dels articles nominats per ser suprimits, i explica que una categorització insuficient dels articles fa que les biografies sobre dones cridin menys l'atenció, igual que els articles poc enllaçats (és a dir, els que tenen pocs enllaços que dirigeixen a altres articles), augmenta les possibilitats de ser nomenats per supressió.

Exposició dels resultats

A continuació, en aquesta secció, analitzarem els resultats obtinguts del creuament de propietats de Wikidata amb els articles que han aparegut a la portada de l'edició en anglès de l'enciclopèdia. Per a fer-ho, combinarem la interpretació dels resultats obtinguts de cadascuna de les propietats especificades a la metodologia amb l'anàlisi dels resultats segons el gènere de les persones que apareixen a la portada. Així, podrem observar tant els resultats globals, sense distinció de gènere, com els resultats classificats a partir de la propietat P21 (sex or gender), que permetran detectar possibles biaixos en la representació de la diversitat d'identitats. Seguidament, presentarem una taula resum de les biografies analitzades i dels valors recollits de cada propietat.

En primer lloc, podem observar la diferència entre la quantitat de biografies respecte d'altres tipus d'articles a les seccions "From today's featured article", "Did you know..." i "On this day" de les portades de l'edició en anglès de Wikipedia. Del total de 99872 articles analitzats, només 22303 dels articles són sobre persones (P31:instance of:human); és a dir, només el 22,3% dels articles destacats que apareixen a la portada són biografies.

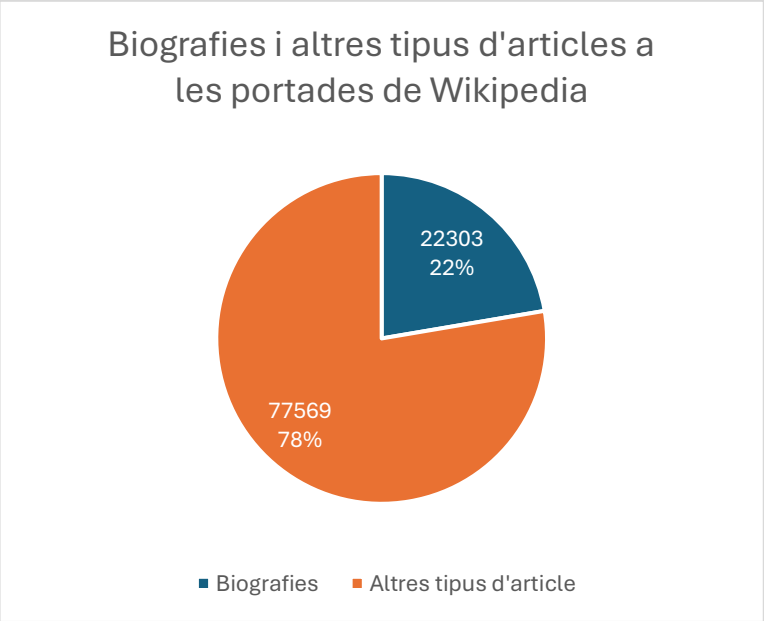

Figura 3. Diferència entre el resultat "ser humano" i les altres categories de la propietat P31 (instance of).

En segon lloc, i ara ja centrant-nos únicament en les biografies, podem analitzar la representativitat de gèneres a les portades, que treballarem a partir de la propietat P21:sex or gender. Per a fer-ho, analitzem les 19992 biografies en què s'especifica aquesta propietat, que empara 12 valors diferents.

«El terme *persona no-binària* és un paraigua ampli que aixopluga identitats molt diverses, amb la característica comuna que es desmarquen del binarisme de gènere. Aquest binarisme es basa

en la idea o concepció que només hi ha dues maneres de viure el gènere: o com a home o com a dona. Però som moltes les persones que quedem fora d'aquesta idea i que vivim la nostra identitat de gènere més enllà d'aquestes dues categories predeterminades.» (Moyano (dir.), 2023, p. 13). Aquest fragment testimonia una dissidència respecte la idea binària del gènere que s'ha tingut durant anys, una classificació que descriuen com a «categories predeterminades» i que no representa la diversitat d'identitats de gènere de la societat. Partint d'aquí, analitzarem els resultats generals de les portades (és a dir, sense fer distinció de gènere) i segons l'agrupació "home", "dona" i "dissidència" (en aquest cas, només es tindran en compte els articles sobre persones de les quals s'especifica el gènere a Wikidata). En aquesta última categoria, hi englobem *no-binarietat, home intersex, eunuc, gènere fluïd, travesti, dona trans* i *home trans*[2]. Pel que fa a la categoria "home", hi són inclosos els resultats *masculino* i *male* extrets de les portades; i en quant a la categoria "dona", s'hi engloben els resultats *femenino* i *female*. A més, hem agrupat algunes categories que sortien desdoblades en anglès i castellà ("trans woman"-"mujer transgénero") per a fer-ho més cohesionat, i hem traduït els noms al català.

Segons es mostra a la Figura 4, la majoria de biografies són sobre homes, seguides de les de dones. Si ho mirem en percentatges, el 71,2% de les biografies són sobre homes, el 28,5% són sobre dones i el 0,4% restant correspon a les altres categories analitzades (si bé s'hi inclouen homes o dones trans). Ara bé, cal tenir en compte que aquesta propietat (P21:sex or gender) combina tant el sexe com el gènere, i que es diferencia entre "dona" i "dona trans" (per exemple), creant una separació dintre de la mateixa identitat de gènere. Aquests resultats mostren un biaix de gènere molt gran entre les diverses categories, fet que evidencia la necessitat de treballar per crear i mantenir contingut sobre dones o no-homes, de manera que la diferència quantitativa es pugui anar reduint.

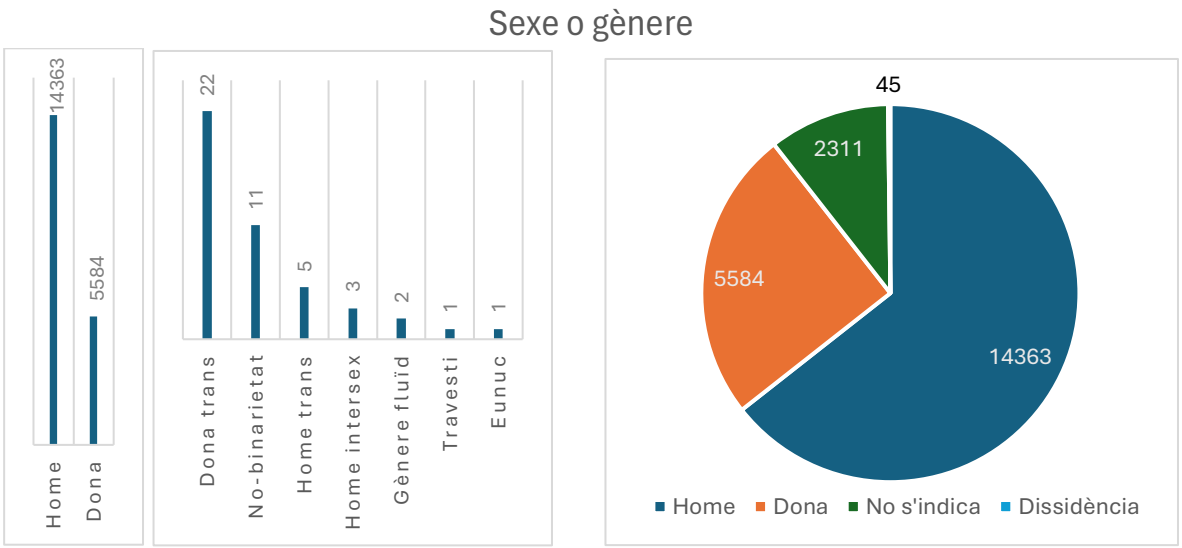

Figura 4. Resultats de la propietat P21 (sex or gender) de les biografies que apareixen a les portades.

[2] Si bé les dones i els homes trans són dones i homes, respectivament, els englobo a la categoria general "dissidència" com a marca de dissidència del gènere assignat al néixer.

Un cop analitzada la representació de les diverses identitats de gènere, volíem analitzar la categoria P6553 (personal pronoun) per veure si hi havia divergències. Sorprenentment, hem trobat que en cap article consta els pronoms que utilitza o utilitzava la persona. Així doncs, no hem pogut analitzar aquesta propietat.

Un altre aspecte que s'ha analitzat és l'orientació sexual de les persones, amb la voluntat de veure si es tendia a representar més un tipus d'orientació (i, per tant, no es donava visibilitat a una part de la societat per motius de sexualitat) o si es donava espai i visibilitat a la diversitat d'orientacions sexuals. La majoria de les biografies, segons hem pogut veure, no especifiquen l'orientació sexual de la persona —només 287 biografies ho fan—, fet que es pot deure a la recomanació «use ONLY IF they have stated it themselves, unambiguously, or it has been widely agreed upon by historians after their death» que s'especifica a la propietat de Wikidata. Així doncs, hem analitzat les biografies en què consta l'orientació sexual i hem vist que hi ha més presència de persones no-heterosexuals (266) que no pas d'heterosexuals (17). Concretament, hi ha 17 persones classificades com a "no-heterosexual", 65 lesbianes, 10 gays, 93 persones homosexuals (s'utilitza aquesta etiqueta a part de gay o lesbiana, de manera que no s'especifica el gènere d'aquestes persones), 79 bisexuals i 2 pansexuals. En canvi, només hi ha 17 persones categoritzades com a heterosexuals. I, a més d'aquestes categories, també trobem 3 persones amb la classificació "sense etiqueta" i 1 persona asexual.

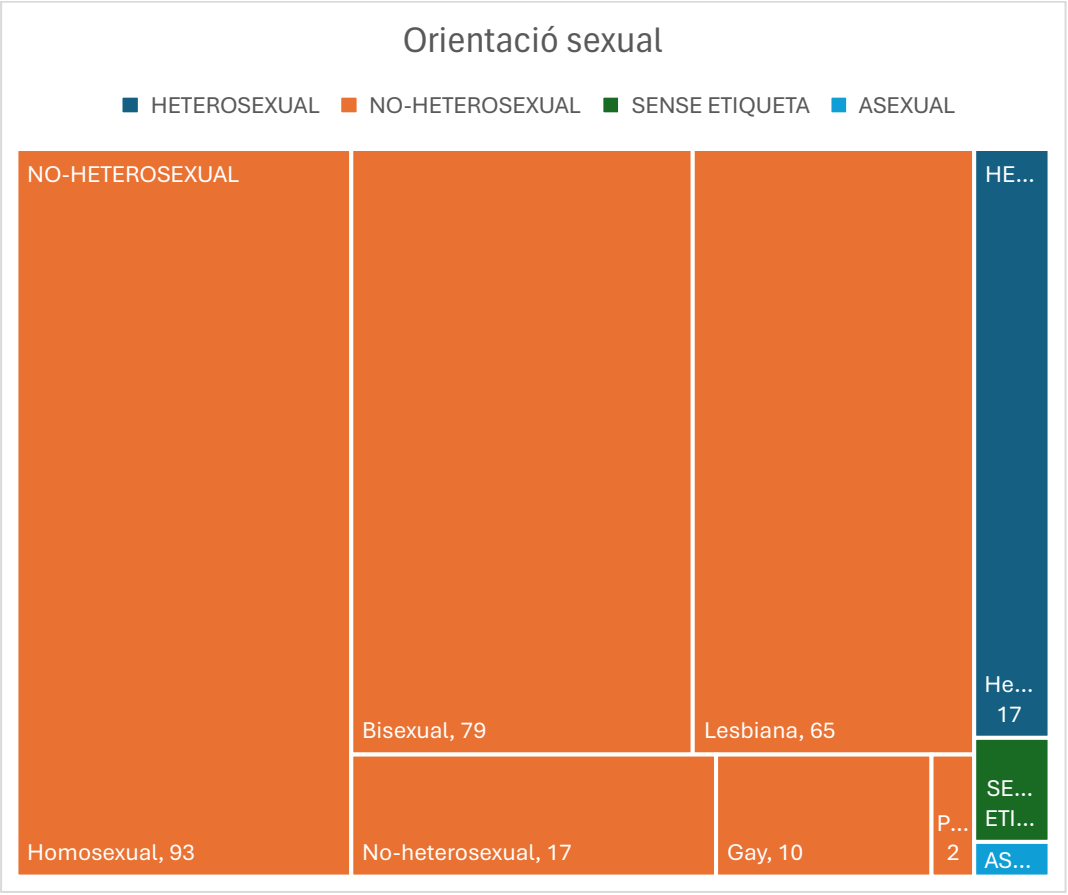

Figura 5. Resultats de la propietat P91 (sexual orientation) de les biografies que apareixen a les portades de l'edició anglesa.

El següent aspecte que hem analitzat és la representació d'ètnies a les portades, present en 1543 biografies. En total hi hem trobat representades 193 ètnies diferents, de les quals es mostren les quinze més comunes a la Figura 6. Segons hem pogut veure, hi ha 81 ètnies representades amb 1 única persona (41,9%), fet que potser es deu a la distinció que han fet les persones editores a l'hora d'atribuir les ètnies (és possible que es puguin unificar en menys ètnies en funció del corrent teòric que s'adopti a l'hora d'atribuir-les). Un fet impactant ha sigut veure que la categoria "caucàsic" només s'utilitza dues vegades, però en canvi "afroamericà" està atribuït a 287 persones, la qual cosa es deu al biaix que dona per fet que tothom és caucàsic i que, si no ho és, cal indicar-ho.

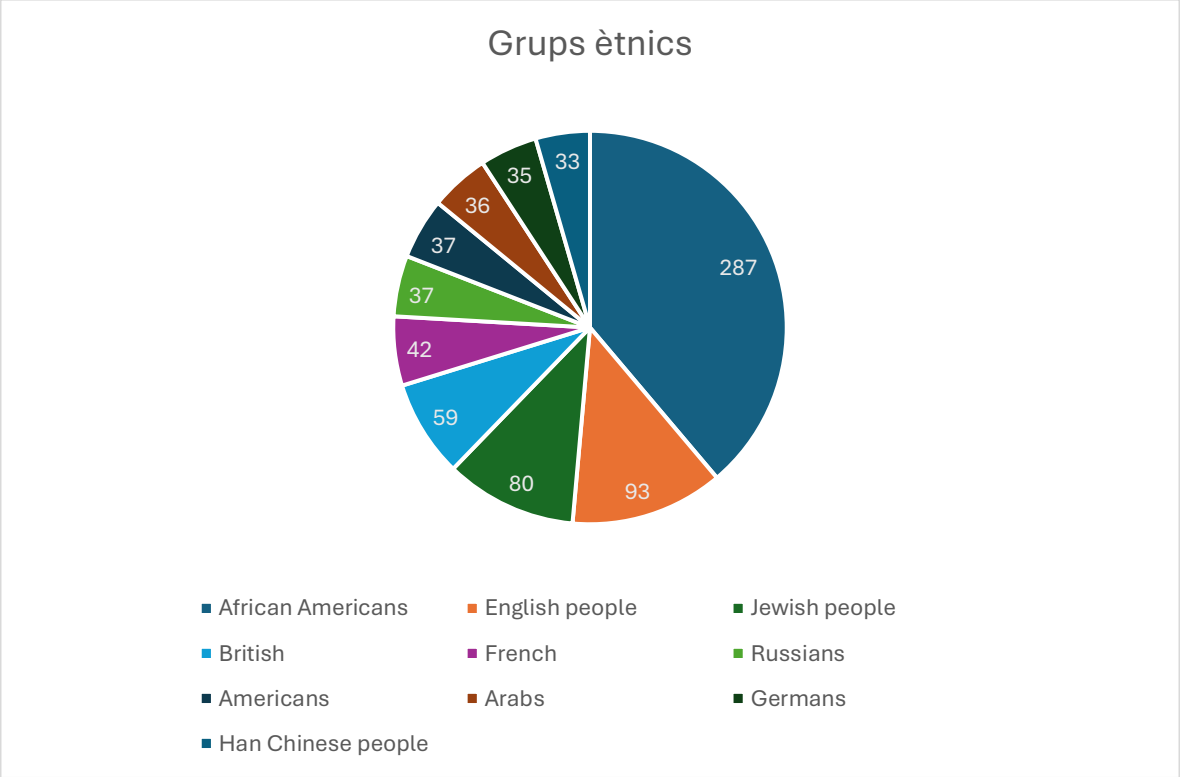

Figura 6. Deu ètnies més comunes a les biografies de l'edició en anglès de Wikipedia, segons els resultats de la propietat P172 (ethnic group).

Si analitzem la distribució d'ètnies en funció de les categories "dissidència", "home" i "dona", podem veure que en el cas de les persones de gènere dissident només s'indica el grup ètnic de 2 persones, de 310 homes i de 229 dones. L'ètnia més freqüent de les biografies sobre homes o sobre dones és *afroamericà*, que es correspon amb l'ètnia més comuna en general (sense distingir entre grups de gènere), i les altres ètnies més freqüents pertanyen a grups d'Europa i d'Amèrica del Nord. Convé destacar que, de les 1876 biografies analitzades inicialment, només n'hem pogut estudiar 1360 en funció del gènere, ja que desconeixem el gènere de les persones de les biografies restants i, per tant, no les hem pogut utilitzar per a fer aquesta anàlisi.

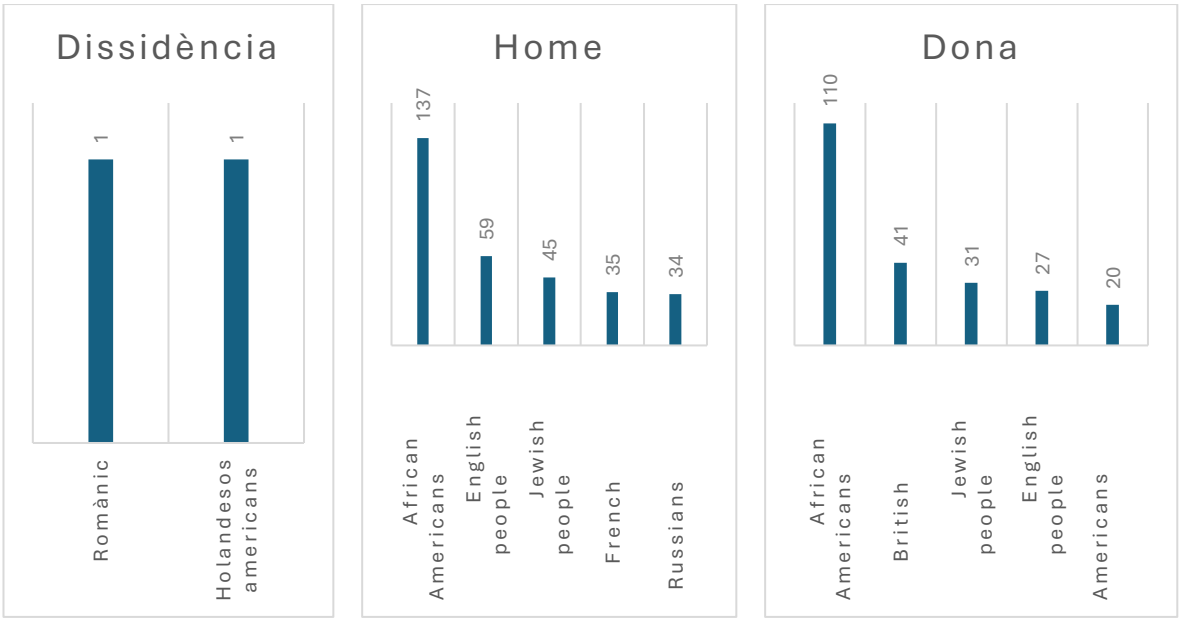

Figura 7. Ètnies més comunes en funció del gènere.

Una altra propietat d'interès ha sigut la categoria P140 (religion or worldview), que ha servit per analitzar la diversitat de religions o maneres de veure el món que tenen representació a Wikipedia. Els resultats han sigut molt diversos, ja que hi ha recollides fins a 158 religions o maneres de veure el món diferents en el total de 5485 biografies en les quals s'especifica aquesta dada. A la Figura 8 es mostren les deu religions o visions del món més comunes, i s'hi afegeixen les dades de les persones que es consideren irreligioses, espirituals però no religioses i agnòstics ateus (tot i que hi ha forces altres religions amb més presència), ja que considerem interessant veure la diferència entre la proporció de religions i la de visions no religioses del món.

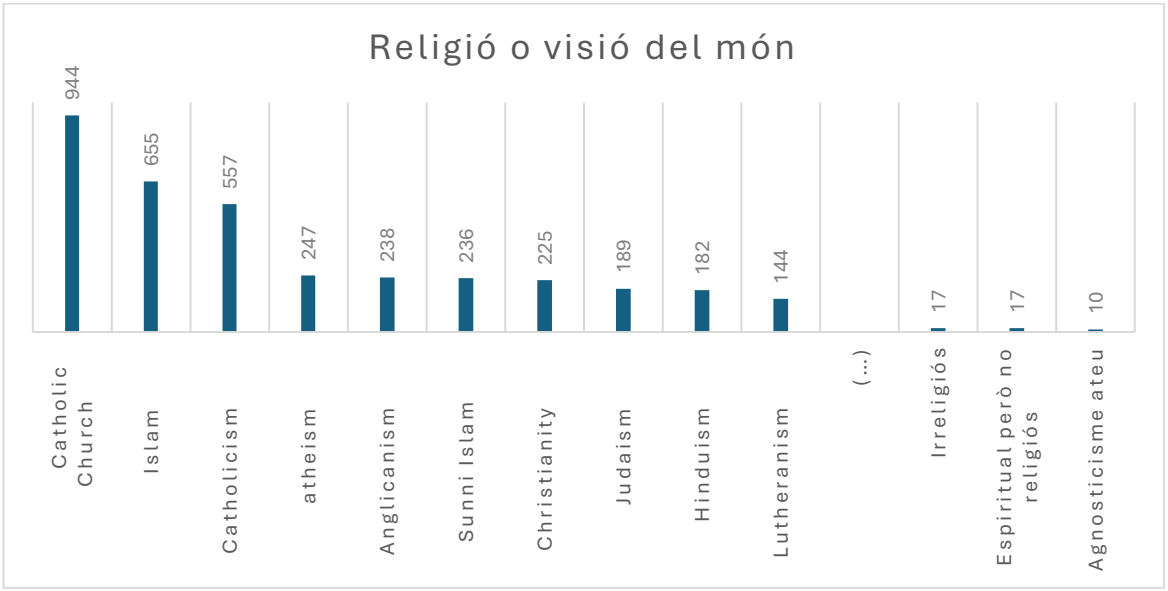

Figura 8. Deu religions més comunes a les biografies de l'edició en anglès de Wikipedia, juntament amb les dades de "irreligiós", "espiritual però no religiós" i "agnosticisme ateu".

Si considerem les religions o visions del món més habituals en funció del gènere (a partir de les 4935 biografies que contenen ambdues propietats), veiem que 6 persones de gènere dissident són atees, i 1 creu en l'Islam. En canvi, en el cas dels homes i les dones, les religions més comunes són el catolicisme i l'Islam. En el cas dels homes, si només analitzem les cinc religions més freqüents, veiem en primera posició l'església catòlica (706), seguida de l'Islam (526) i del catolicisme (388) (és el mateix que l'església catòlica, però com que apareix per separat, ho mantinc), l'ateisme (187) en quarta posició i el Sunni Islam (184) en cinquena (que no uneixo amb Islam igual que no uneixo el catolicisme i l'església catòlica). Pel que fa a les dones, l'església catòlica també ocupa la primera posició, amb 123 creients, la segueix el catolicisme amb 120 persones, l'anglicanisme ocupa la tercera posició amb 72 persones, i després segueixen l'Islam (52) i el judaisme (51). Segons mostren tots aquests resultats, no hi ha grans salts entre aquestes religions o maneres de veure el món, tot i que si uníssim alguns resultats (església catòlica i catolicisme, o diverses branques protestants com a protestantisme, per exemple), això podria canviar.

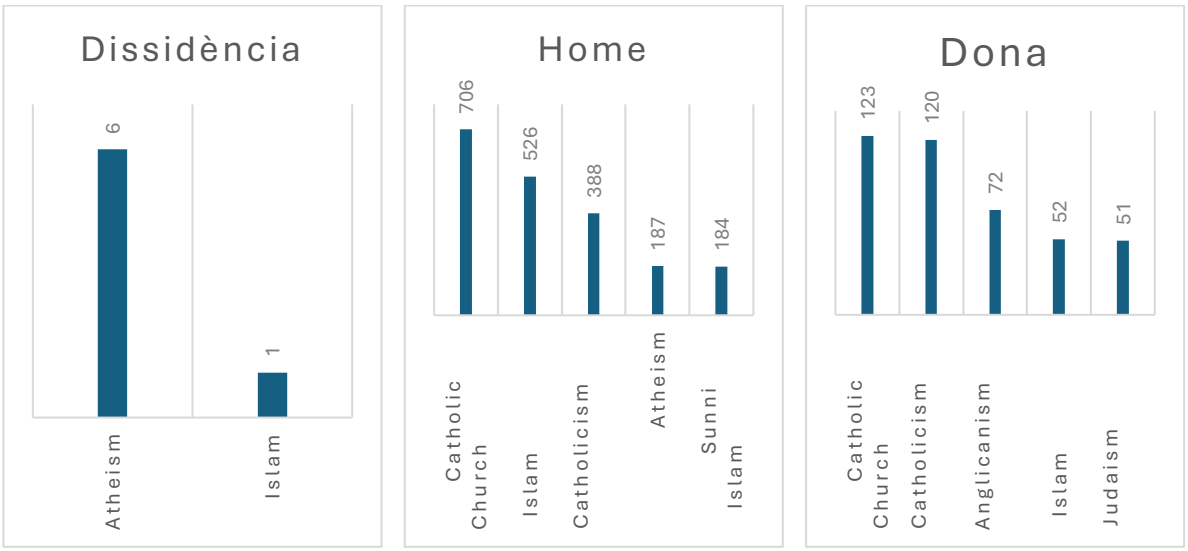

Figura 9. Religions més comunes a les biografies de l'edició en anglès de Wikipedia, en funció del gènere.

En següent terme, comentem conjuntament els resultats de la propietat "llengua materna" (P103:native language) i de "idiomes parlats, escrits o signats" (P1412: languages spoken, written or signed), de les quals he pogut analitzar 3436 i 13127 resultats, respectivament. La llengua materna més comuna és l'anglès, amb 1168 persones, seguida de l'àrab, amb 190 parlants, i de l'alemany, amb 189. Ara bé, aquesta distribució no es manté en els idiomes parlats, escrits o signats. L'anglès sí que es manté en primera posició, amb un total de 5758 persones que el parlen o escriuen –l'anglès signat és una altra llengua–, però la segona posició l'ocupa l'alemany, amb 1072 persones, i el francès és la tercera llengua més parlada o escrita, amb 992. A les figures 10 i 11 es poden veure les deu llengües maternes (Figura 10) i les parlades, escrites o signades (Figura 11) més freqüents.

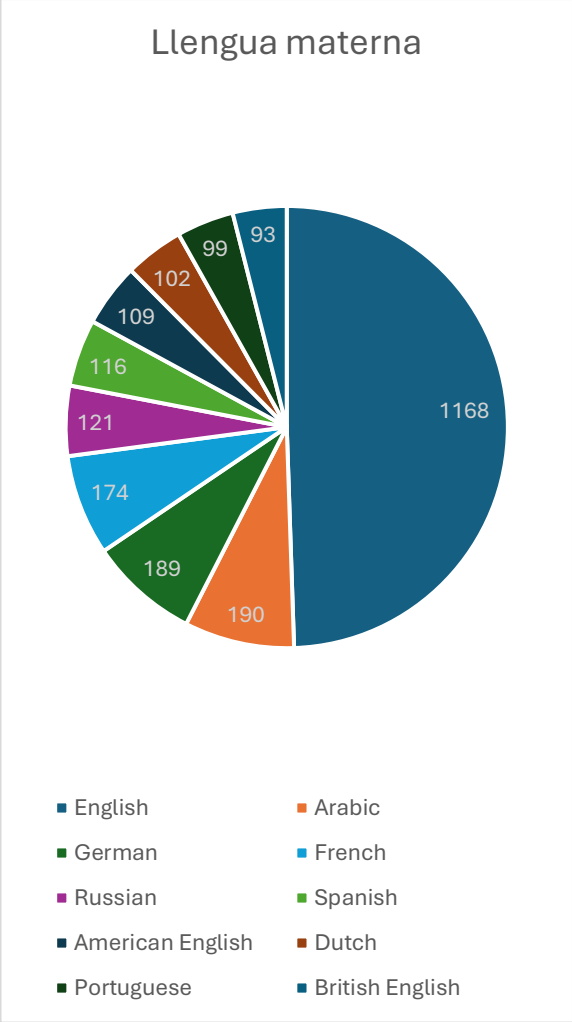

Figura 10. Deu llengües maternes més freqüents.

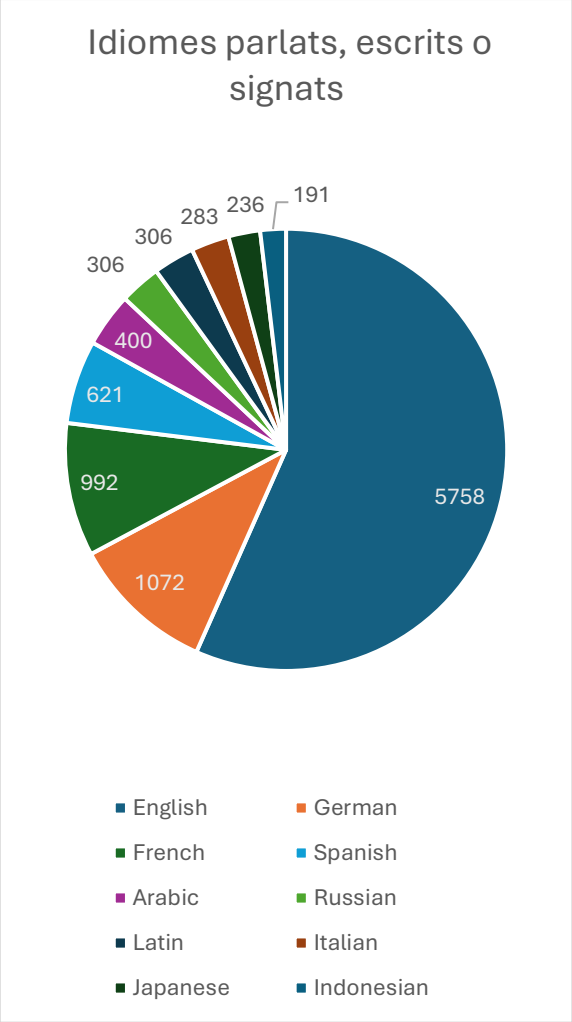

Figura 11. Deu llengües parlades, escrites o signades més freqüents.

En canvi, els resultats canvien si els analitzem en funció del gènere, que analitzo a partir de 3098 (llengua materna) i 11832 (idiomes escrits, signats o parlats) dades. Segons permeten veure les figures 12 i 13, la llengua materna i la llengua parlada o escrita més habitual és l'anglès, que apareix tant unificat com dividit en funció de si és britànic o americà. A continuació, l'altra llengua més habitual és l'alemany, que apareix en totes les categories excepte en la llengua materna de persones dissidents de gènere. Pel que fa a les llengües parlades, escrites o signades, l'espanyol apareix entre les més habituals dels tres grups (dissidència, homes i dones).

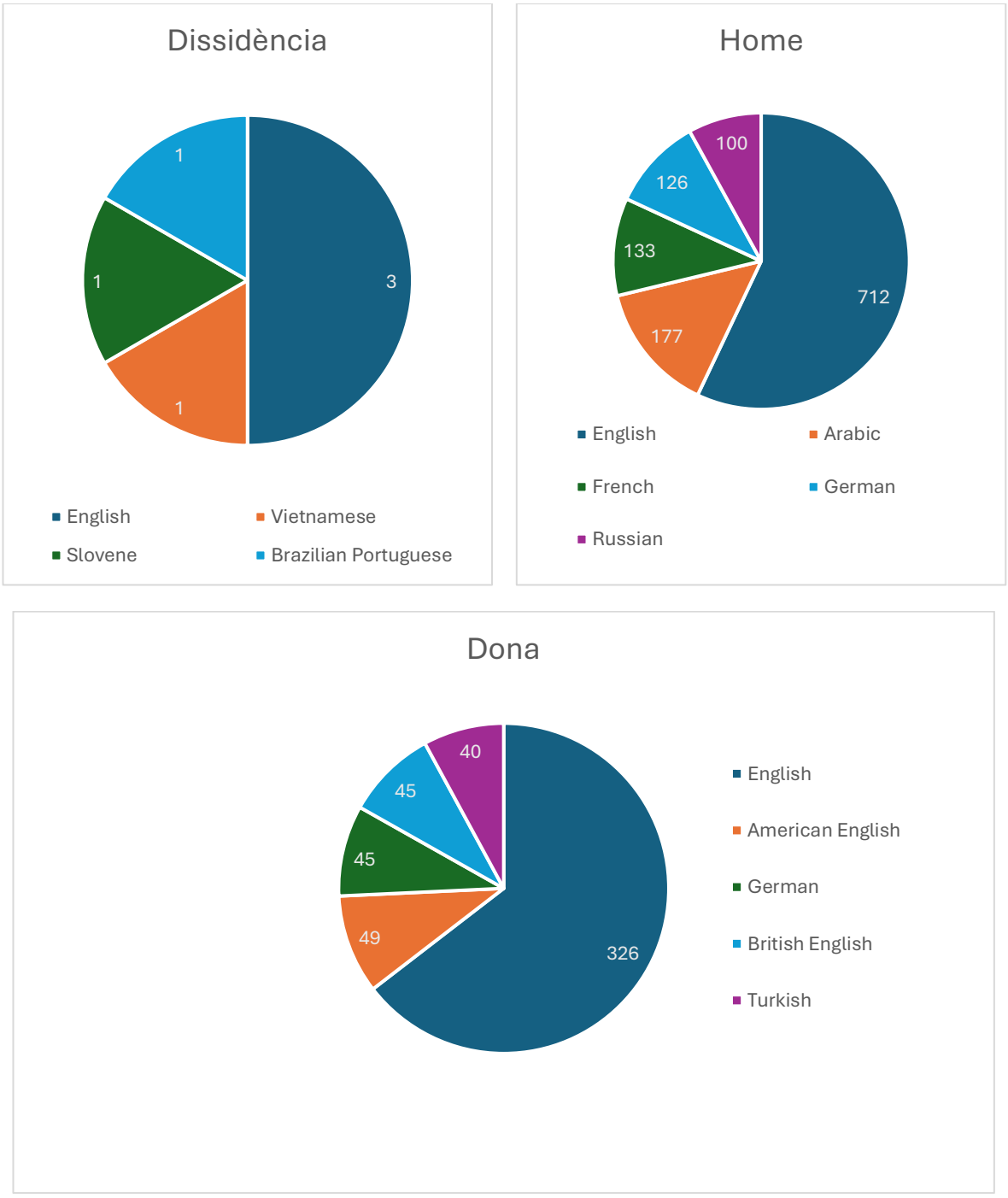

Figura 12. Llengües maternes més freqüents en funció del gènere de les persones que apareixen a les biografies.

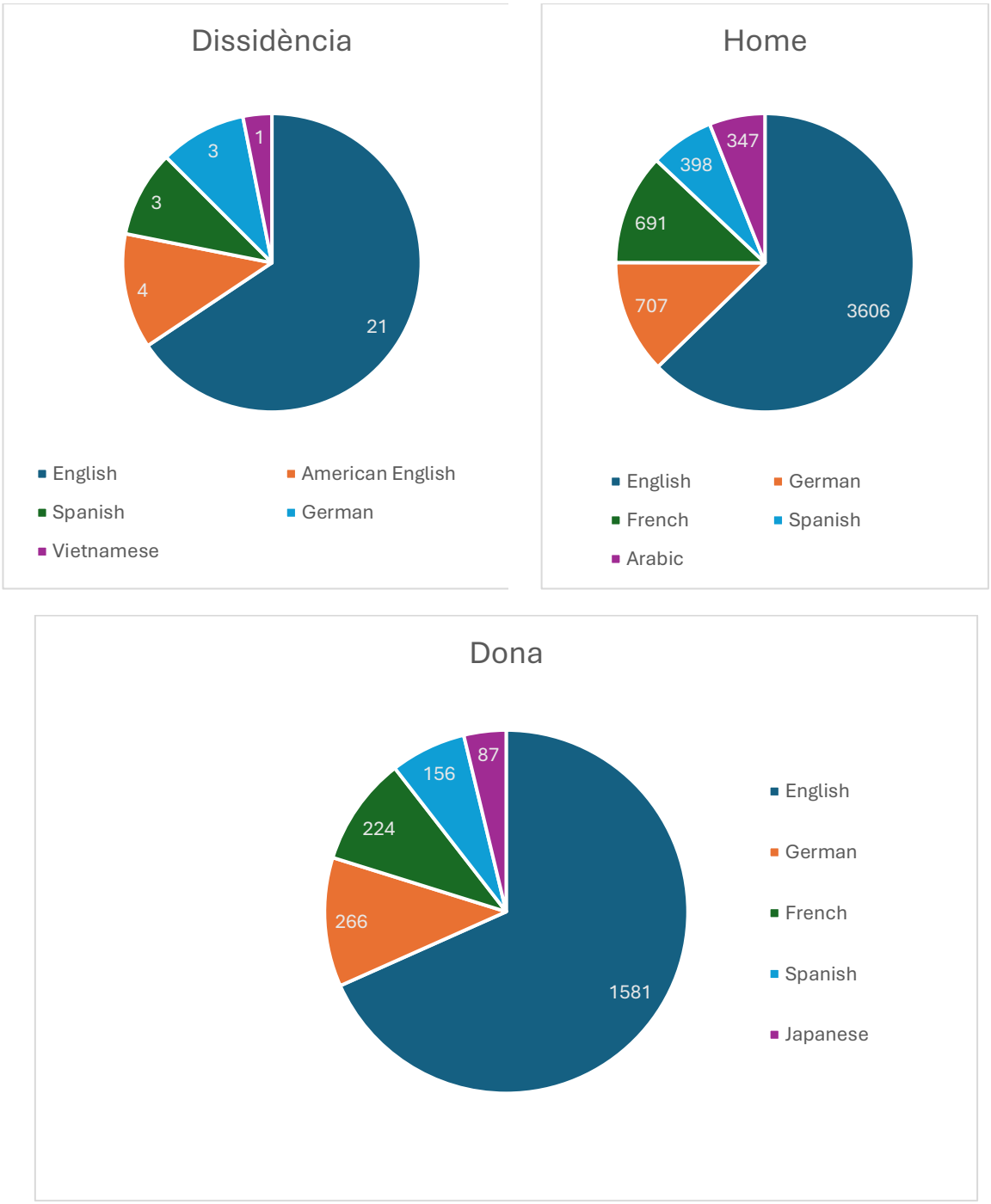

Figura 13. Llengües parlades, escrites o signades més freqüents en funció del gènere de les persones que apareixen a les biografies.

A continuació, també s'ha analitzat la propietat P106 (occupation) per veure quines són les professions més freqüents entre les persones representades a les portades. Segons hem pogut determinar, en els 21007 articles sobre persones que tenen aquesta dada, hi apareixen fins a 1153 professions diferents. Les deu més habituals són: polític (16,9%), escriptor (3,3%), personal militar (2,6%), periodista (1,9%), actor (1,9%), pintor (1,8%), jugador de futbol (1,8%), monarca (1,6%), cantant (1,5%) i sacerdot catòlic (1,4%). Una dada interessant a remarcar és la diferència entre el primer resultat, "polític", amb 3567 resultats, i el segon,

"escriptor", amb 696; és a dir, amb 5,1 vegades menys resultats. I el mateix passa entre el segon resultat i el tercer, "personal militar", amb 541 resultats, que té 155 resultats menys que "escriptor". A la Figura 14 es mostra la distribució quantitativa de les vint professions més freqüents a les portades de l'edició en anglès de Wikipedia.

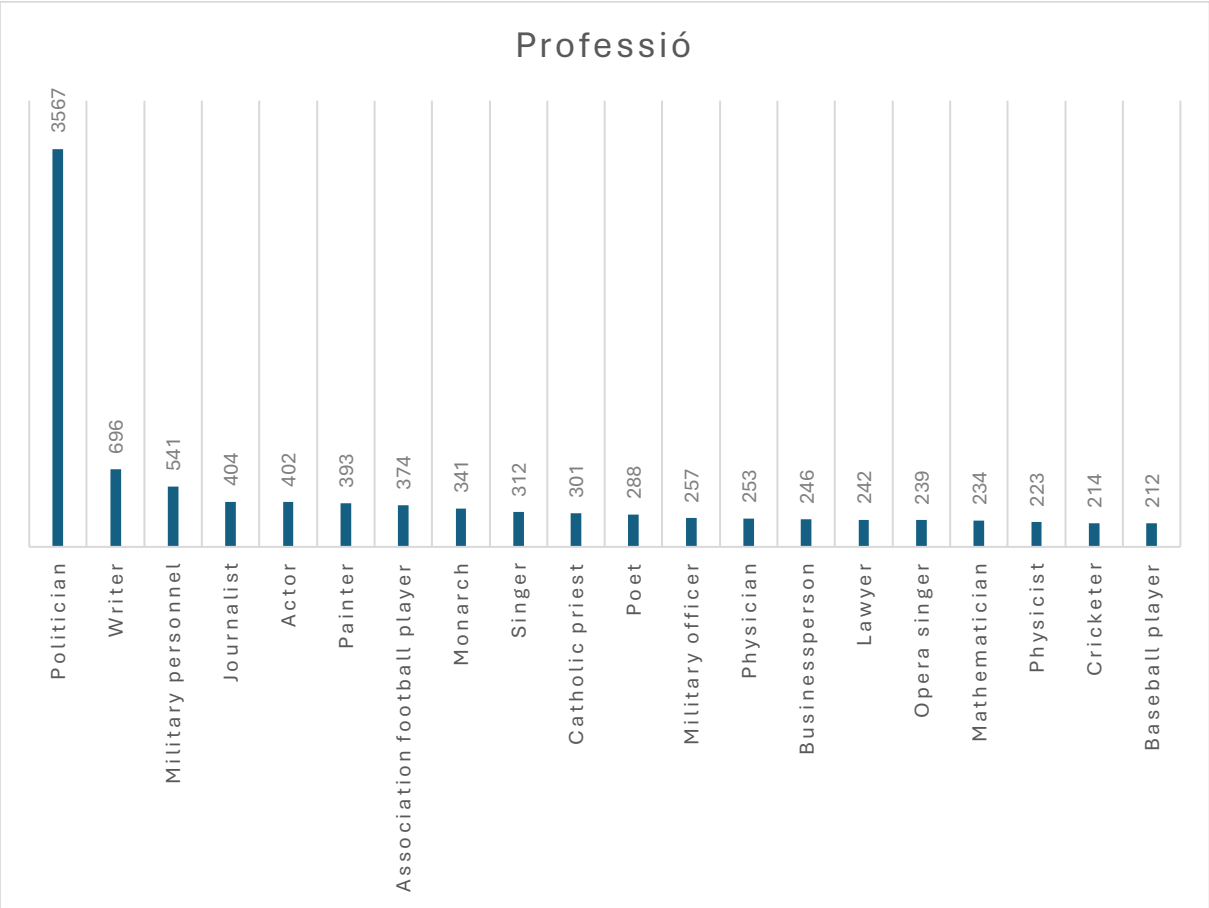

Figura 14. Gràfic quantitatiu de les 20 professions més freqüents entre les persones que apareixen a les portades de l'edició en anglès de Wikipedia.

Si analitzem els resultats d'aquesta propietat en funció del gènere, només podem estudiar 18928 articles, en els quals veiem certes diferències en els resultats d'un grup a l'altre. Una de les professions més habituals entre els homes i les persones de gènere dissident és *personal militar* (amb 2726 resultats en total), mentre que aquesta professió no apareix entre les més freqüents en les dones. Les persones de gènere dissident i les dones comparteixen *cantant* (amb 187 resultats totals) com una de les professions més comunes (entre les dones inclús apareix "cantant d'òpera" com una categoria separada, però també molt comuna, amb 161 resultats). I pel que fa a homes i dones, aquests dos grups tenen en comú la professió d'escriptors (287 homes, 323 dones), mentre que en el grup de dissidència de gènere no apareix aquest resultat com un dels cinc més freqüents. Per altra banda, veiem que els tres grups comparteixen la professió *polític*, amb 2 resultats entre persones de gènere dissident, 2722 entre homes i 482 entre dones. És la professió més freqüent entre homes i dones, i destaca per ser molt més habitual que les que la segueixen (és 5,85 vegades més habitual en el cas dels homes, i 1,5

vegades més habitual en el cas de les dones). En canvi, entre les persones de gènere dissident, la professió més freqüent és *personal militar*, que és una vegada més freqüent que *oficial militar*, en segona posició, i *periodista*, en tercera.

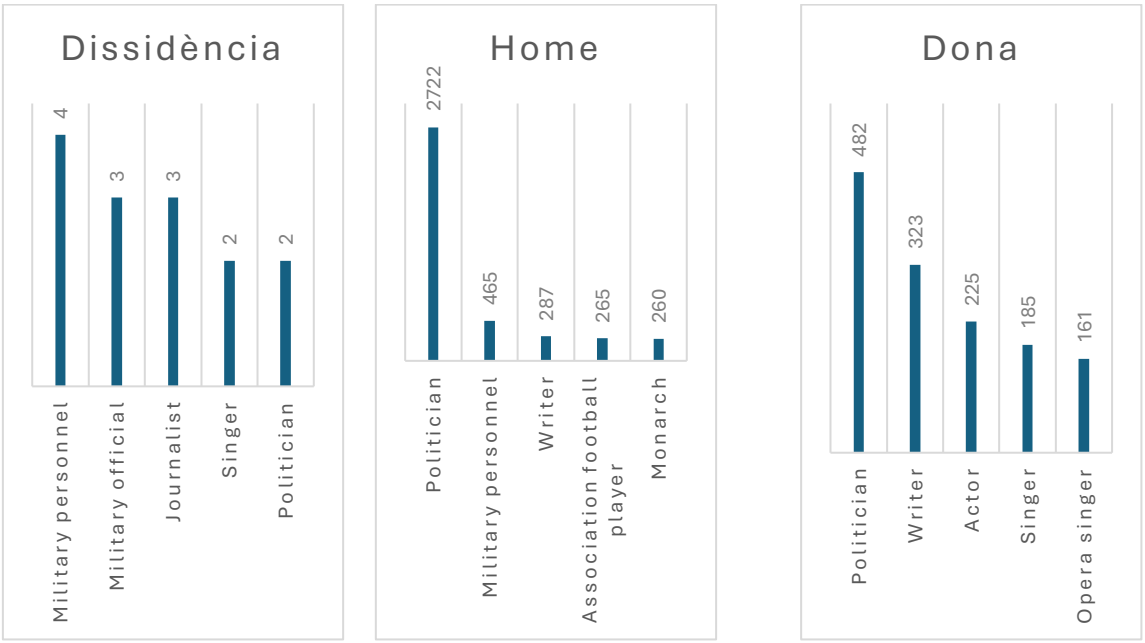

Figura 15. Gràfic quantitatiu de les professions més freqüents entre les persones que apareixen a les portades de l'edició en anglès de Wikipedia, en funció del seu gènere.

Un altre fet que hem analitzat és si s'indica el pare, mare o cònjuge de les persones de les portades per esbrinar si el fet de tenir una persona notable en alguna d'aquestes categories pot ajudar a que aquella persona també sigui categoritzada com a notable i, així, tenir més possibilitats de ser mostrada a la portada. Segons indiquen els resultats de les 16524 biografies analitzades (5715 per la categoria "pare", 4351 per "mare" i 6458 per "cònjuge") i tal com mostra la Figura 16, en totes tres categories és més comú que no consti el pare (16435), mare (17865) o cònjuge (13397), que no pas que consti, de manera que podríem concloure que aquest fet no atorga notabilitat ni afavoreix l'aparició a la portada.

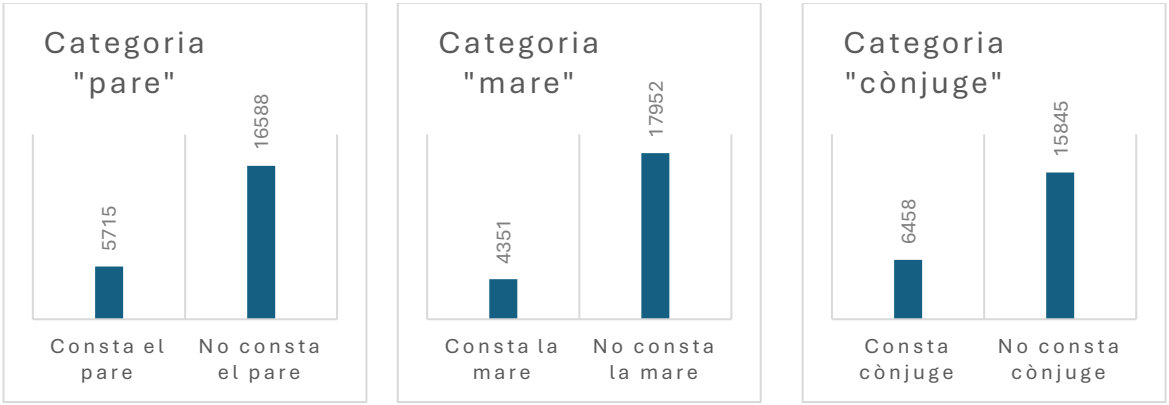

Figura 16. Aparició o no de la categoria "pare" (P22), "mare" (P25) o "cònjuge" (P26) de les persones representades a les portades.

Ens sembla interessant analitzar si aquestes tres propietats consten o no en funció del gènere de les persones. En els 5164 articles en què consta la metadada "pare", els 3939 en què consta la metadada "mare" i els 5834 en què consta "cònjuge", consta més el pare que no pas la mare o cònjuge en les persones de gènere dissident, amb *cònjuge* en segona posició; mentre que entre homes i dones la propietat que més consta d'aquestes tres és *cònjuge* seguida de *pare* i, en darrer lloc, *mare*.

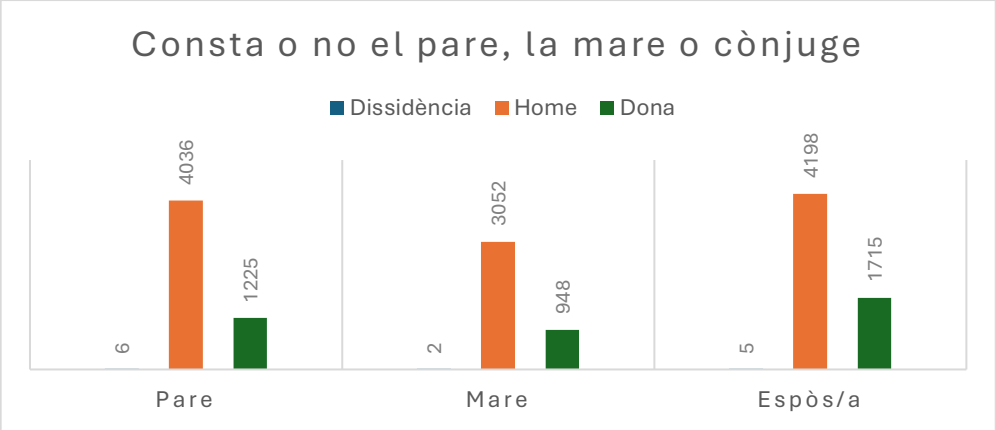

Figura 17. Gràfic sobre si consta o no de la categoria "pare" (P22), "mare" (P25) o "cònjuge" (P26) de les persones representades a les portades en funció del gènere.

Amb relació a la notabilitat que es necessita per tenir un article a Wikipedia, també es pot tenir en consideració la propietat P166 (award received), que indica premis o distincions que s'han donat a aquella persona, la qual cosa podríem considerar un indicador de notabilitat. A l'analitzar aquesta propietat, hem trobat que, entre les biografies de les portades, és més habitual que no hi aparegui aquesta dada, bé sigui perquè no s'ha introduït o perquè aquella persona no ha guanyat cap premi. Segons l'anàlisi realitzada, només en 8050 biografies s'indica alguna distinció o premi que ha rebut aquella persona; en canvi, en les 14253 restants no n'hi ha cap indicació.

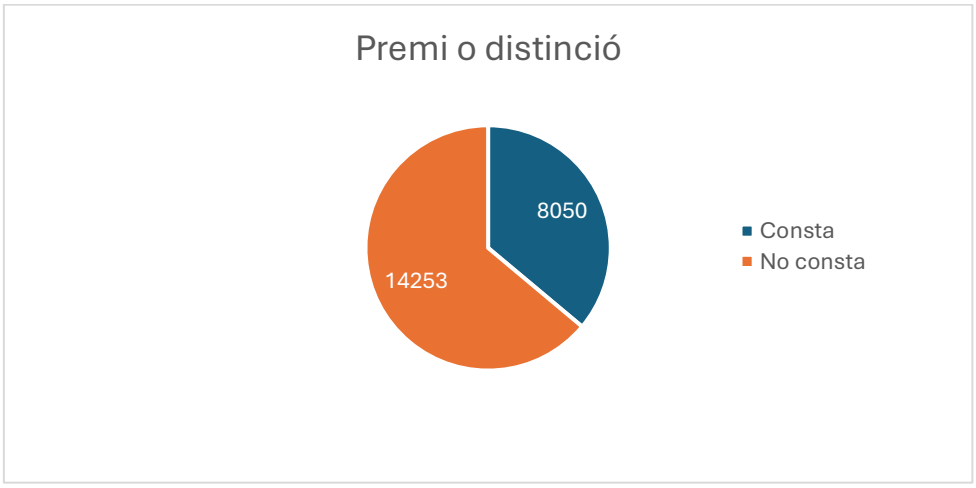

Figura 18. Diferència entre les biografies en les que consta que aquella persona ha rebut un premi o distinció i les que no hi consta.

Si analitzem la constància d'aquesta propietat en funció del gènere, veiem que en els tres grups el més freqüent és que no consti si les persones han rebut un premi o distinció (fet que ja demostrava l'anàlisi general). Ara bé, la diferència quantitativa varia en funció del gènere: hi ha 14 persones de gènere dissident de les quals consta que hagin rebut cap premi o distinció, mentre que n'hi ha 31 de les quals no consta aquesta propietat; en quant als homes, el salt quantitatiu és més gran, atès que hi ha 5437 persones de les quals sabem que han rebut com a mínim un premi, i 8926 de les que no ho sabem (és a dir, 3489 persones de diferència); i pel que fa a les dones, sabem que 1905 dones han obtingut algun premi o distinció, però ho desconeixem de 3679 (1774 dones de diferència).

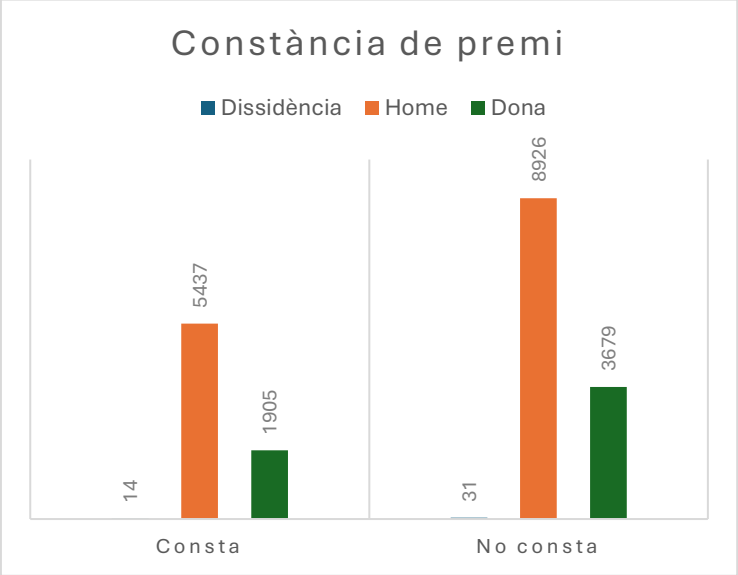

Figura 19. Diferència entre les biografies en les que consta que aquella persona ha rebut un premi o distinció i les que no hi consta, en funció del gènere.

En següent lloc, també hem analitzat les dades geogràfiques de les persones representades. Per a fer-ho, hem utilitzat quatre propietats: P19 (place of birth), P20 (place of death), P27 (country of citizenship) i P30 (continent). Hem analitzat els resultats de 18917 articles per la primera propietat, 12150 per la segona, 18988 biografies per la tercera, i 0 per la quarta propietat, ja que no consta en cap article sobre persona.

Segons hem pogut veure, els llocs de naixement més habituals són grans ciutats dels Estats Units o d'Europa. És a dir, de zones benestants o del que es considera el Nord Global. El mateix passa amb els cinc principals llocs de defunció, que són gairebé tots els mateixos. No obstant, en vuitena posició hi trobem Constantinoble (que ocupa la onzena posició del lloc de naixement), que pot deure's als personatges de l'Antiguitat clàssica; i en novena posició hi trobem Beijing, que forma part d'Àsia, continent que no apareix en les primeres posicions de la gràfica de naixements.

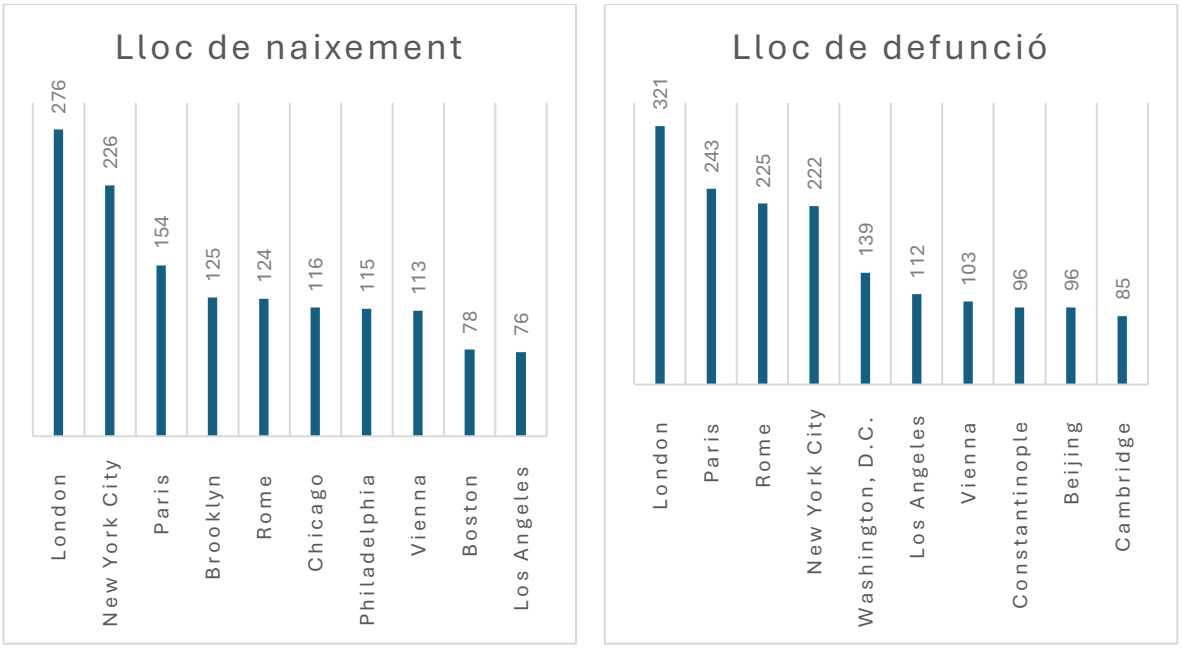

Figura 20. Diferència entre els deu principals llocs de naixement i de defunció de les persones representades a les portades.

Si mirem els principals resultats de la nacionalitat (P27:country of citizenship), veiem que l'hegemonia del Nord Global i dels països més desenvolupats econòmicament també hi és present en les cinc primeres posicions. Tanmateix, a partir de la setena posició, amb Índia (439 persones), se suavitza el predomini de les regions esmentades.

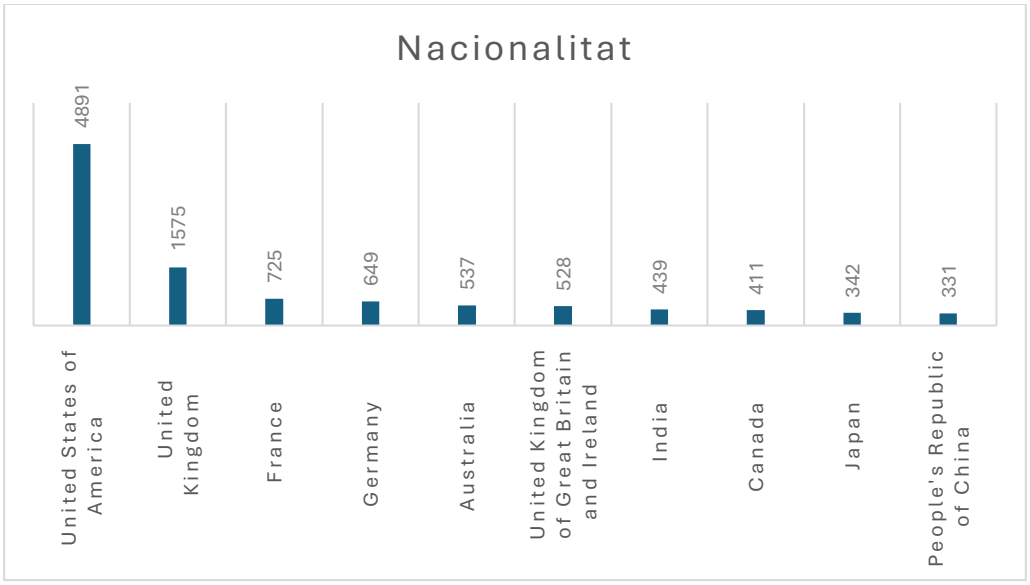

Figura 21. Deu principals nacionalitats de les persones representades a les portades.

L'última categoria que hem volgut analitzar en quant a l'origen geogràfic és, com hem anunciat, la distribució de persones en funció dels continents d'origen. Tanmateix, això no ha estat possible, atès que les dades d'aquesta propietat no consten en cap biografia.

Si analitzem aquestes propietats geogràfiques en funció del gènere, decauen les possibilitats d'estudi, atès que només hi ha 17110 articles en els quals consta el lloc de naixement, 10818 amb el lloc de defunció, i 17129 en els quals s'hi indica la nacionalitat. A continuació, analitzem els resultats.

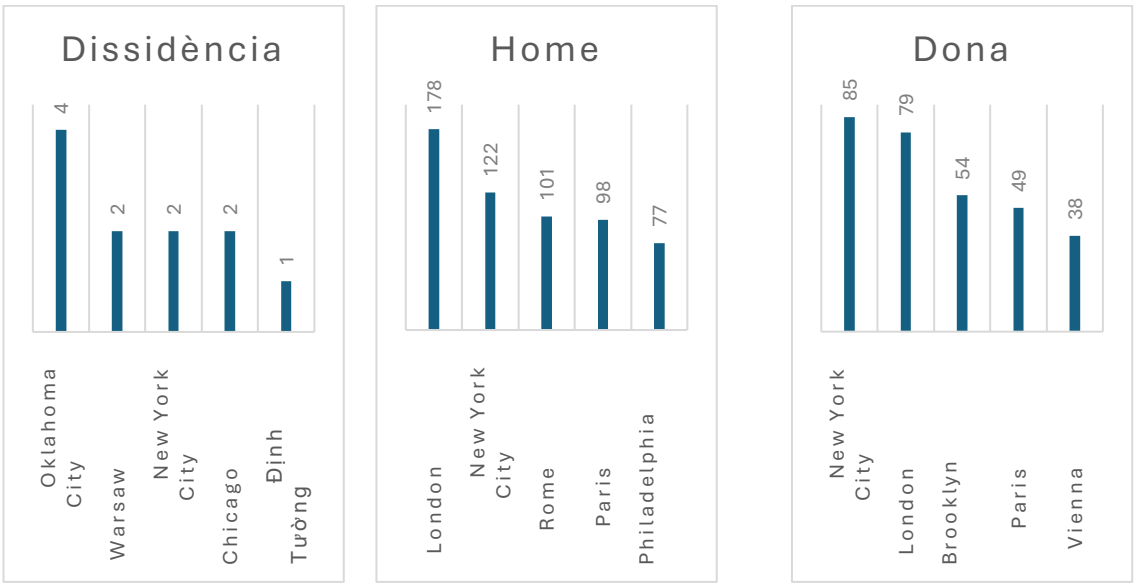

Figura 22. Cinc llocs de naixement més habituals en funció del gènere.

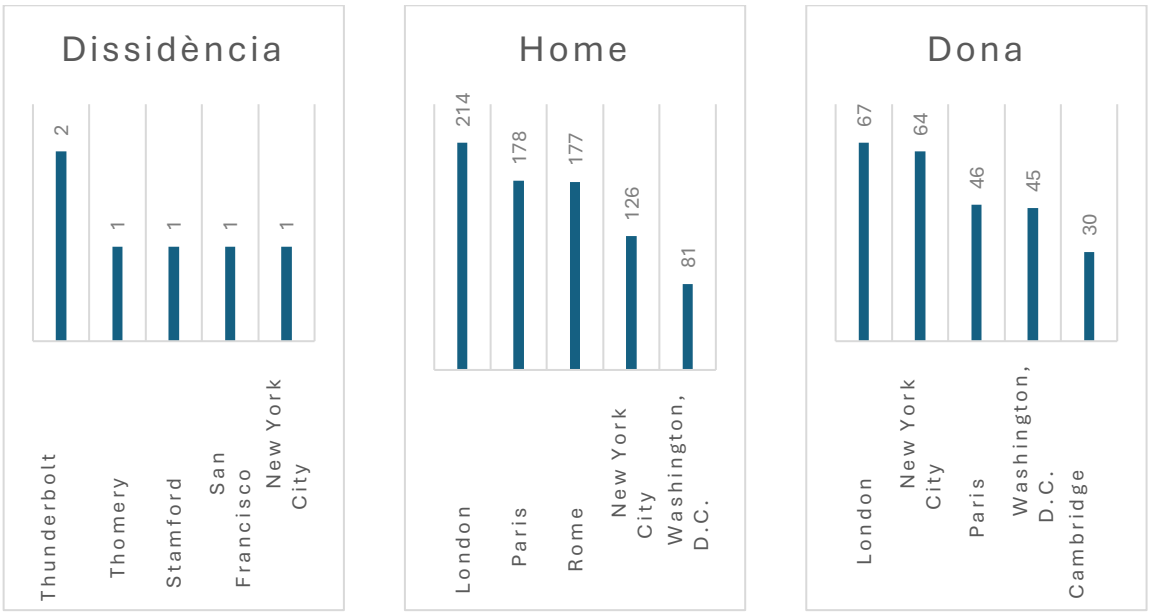

Figura 23. Cinc llocs de defunció més habituals en funció del gènere.

Segons podem veure, en els tres grups el més habitual és néixer i morir a ciutats del Nord Global, d'Europa o d'Amèrica del Nord. Aquest fet es compleix al 100% en el lloc de naixement i de defunció d'homes i de dones, i en el lloc de defunció de persones de gènere dissident. En canvi, pel que fa al lloc de naixement de persones de gènere dissident, veiem que hi ha 8 persones que han nascut a ciutats d'Estats Units d'Amèrica, 2 persones que han nascut

a Polònia i 1 persona que ha nascut a Vietnam. A més, podem veure que en aquesta propietat no hi ha gran salts quantitatius entre el primer resultat i el següent.

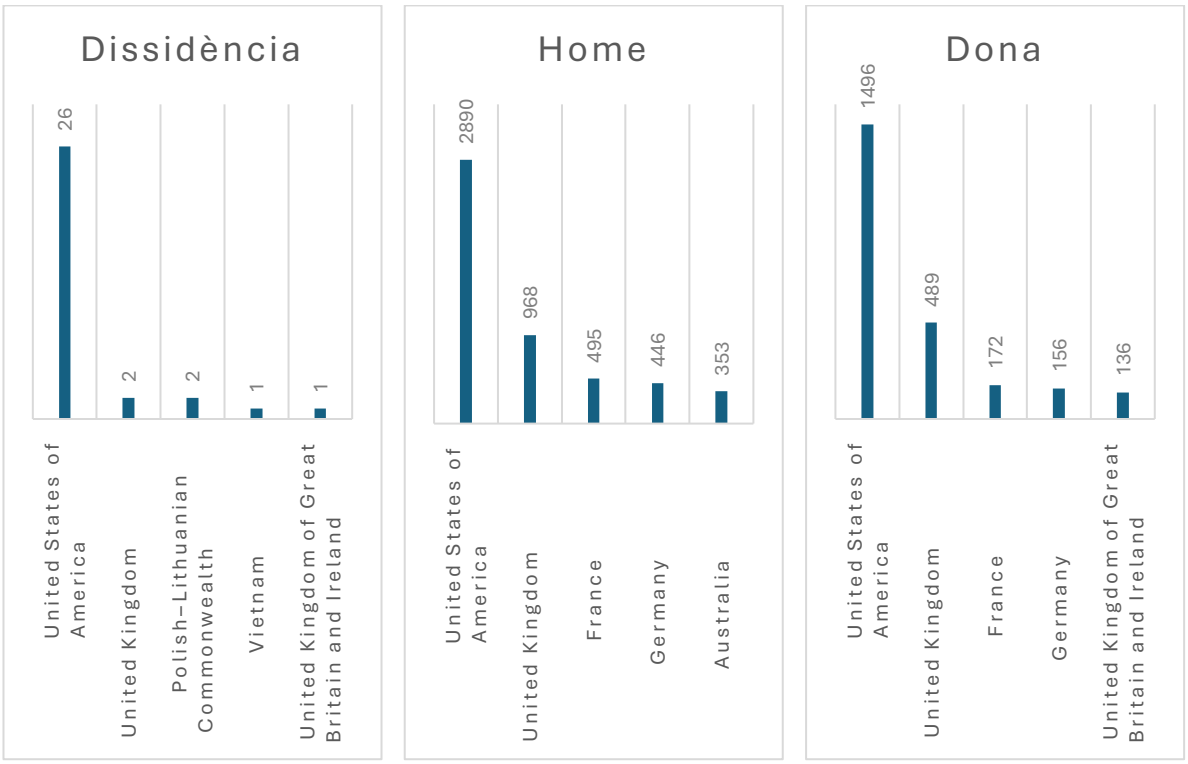

Figura 24. Cinc nacionalitats més comunes en funció del gènere.

Pel que fa a la nacionalitat, aquí trobem altra vegada la mateixa distribució geogràfica que en les propietats de lloc de naixement i de defunció. De nou, els països més habituals són Estats Units d'Amèrica o que es troben a Europa; i, corroborant l'anàlisi anterior, veiem que en el cas de les persones de gènere dissident, n'hi ha 2 que han nascut a la Commonwealth polonesa-lituana (que abans havia considerat com a Polònia, degut a que és el nom actual de l'Estat on es troba) i 1 que ha nascut a Vietnam. Ara bé, en aquesta propietat sí que trobem grans salts entre la primera posició (Estats Units d'Amèrica, en els tres grups) i la segona (Regne Unit, també en els tres grups): 24 persones de diferència en les persones de gènere dissident, 1922 en els homes, i 1007 en el cas de les dones. El país que ocupa la tercera posició més habitual canvia en funció de si les persones són de gènere dissident o si són homes o dones: en el primer cas, és la Commonwealth polonesa-lituana, amb 2 persones, mentre que en el segon i tercer cas hi trobem França, amb 494 homes i 172 dones. Es repeteix la mateixa distribució en la quarta posició: Alemanya és el quart lloc més freqüent entre homes (446) i dones (156), mentre que en el cas de les persones de gènere dissident aquesta posició l'ocupa Vietnam amb 1 persona. Per últim, Regne Unit de Gran Bretanya i Irlanda ocupa la cinquena posició entre persones de gènere dissident (1) i dones (136), mentre que en el cas dels homes l'ocupa Austràlia (353). Respecte aquest resultat, cal remarcar que "Regne Unit de Gran Bretanya i Irlanda" i "Regne Unit" són el mateix país geogràficament, però que la denominació canvia en funció del període

històric —la primera opció correspon a l'estat anglès entre 1801 i 1922, i la segona al període de 1927 fins a l'actualitat.

Per últim, hem volgut analitzar de quin període històric són les persones que apareixen a les portades per a poder detectar si es tendeix a donar més representació a certs períodes o a d'altres. Per a fer-ho, en primer lloc hem analitzat les dates de naixement i de defunció més habituals de les biografies que disposen d'aquesta informació (12972 en el cas de la data de naixement i 10363 de la data de defunció), i hem trobat que les persones més representades a les portades són les que van néixer la segona meitat del segle XX i les que han mort entre finals del segle XX i el segle XXI.

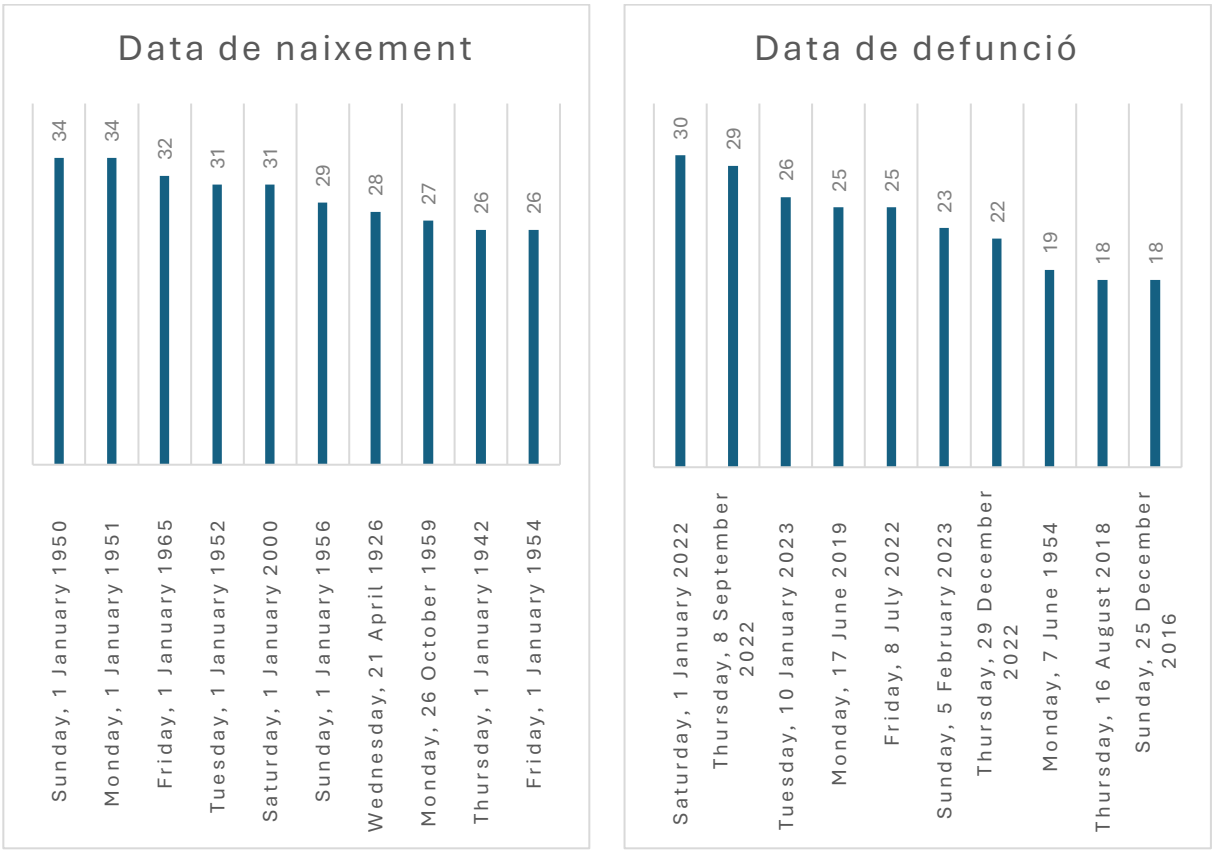

Figura 25. Deu dates de naixement i de defunció més freqüents de les persones que apareixen en portada.

De totes maneres, això canvia si analitzem la propietat del període històric (P2348:time period), que mostra que la majoria de biografies se situen en el període de l'Imperi Romà (188 entre els tres primers resultats), seguides de l'Imperi Bizantí (26), el segle XX amb 23 resultats, el segle XXI amb 22, i l'Antiga Egipte (45 entre els tres resultats que es troben en les deu posicions més comunes).

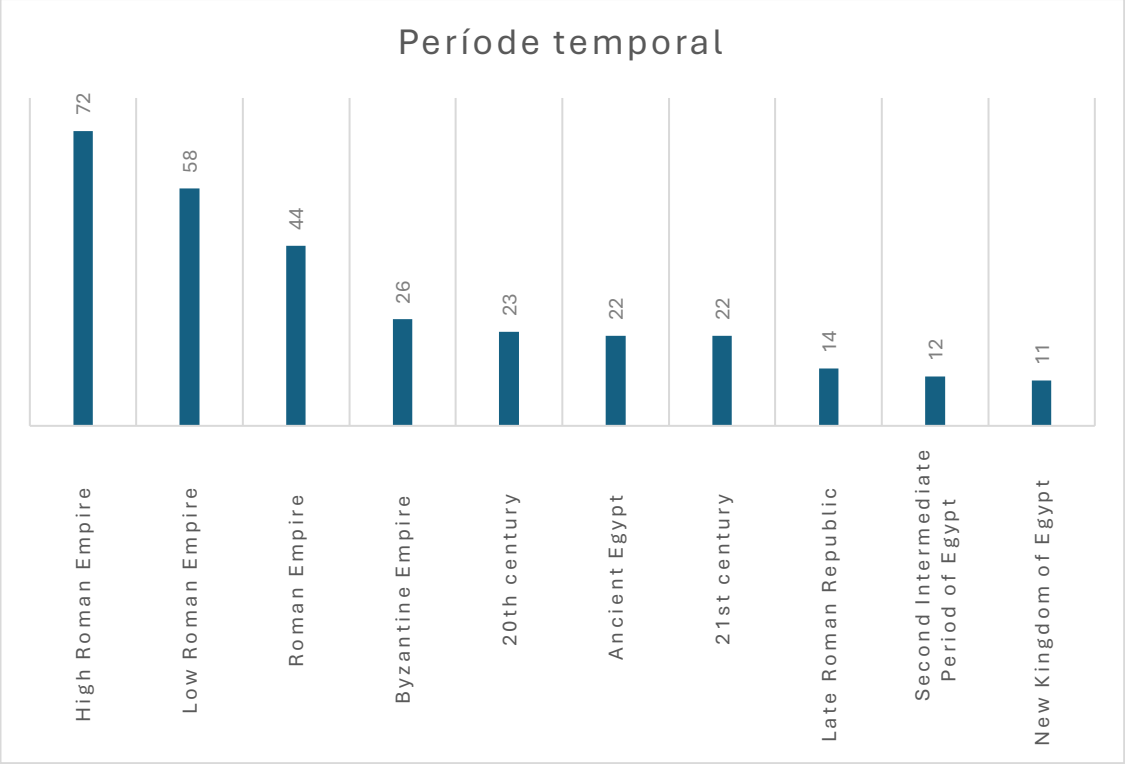

Figura 26. Deu períodes històrics més freqüents que consten en les biografies que apareixen a les portades.

A partir de l'anàlisi que hem fet de les 21577 biografies que tenen tant la metadada de gènere com la de data de naixement (11841), de defunció (9357) o de període històric (379), hem trobat en primera posició 6 persones de gènere dissident que van néixer el 17 de desembre de 1987, 27 homes que van néixer el 26 d'octubre de 1959, i 29 dones que van néixer l'1 de gener de 2000. Pel que fa a les dates de defunció, trobem que cada persona de gènere dissident va morir en una data diferent, de manera que no podem establir una preponderància segons aquest criteri; 25 homes van morir el 8 de juliol de 2022, 25 altres el 17 de juny de 2019, i 25 d'altres l'1 de gener de 2022; i, per últim, veiem que 27 dones van morir el 8 de setembre de 2022.

No hi ha diferències quantitatives entre una posició i la següent, sinó que són totes força seguides i inclús forces d'elles corresponen a la mateixa quantitat de persones que altres dates (hi ha quatre dates de naixement que corresponen a 19 homes cadascuna, 3 que corresponen a 16 dones, tres dates de defunció que corresponen a 25 homes, 3 a 10 dones, etc.). Ara bé, totes aquestes dates són dels segles XX i XXI, la qual cosa fa pensar que es tendeix a donar més visibilitat a figures contemporànies o pròximes a la societat d'avui dia en lloc de donar a conèixer figures de segles anteriors a través de les portades. Per últim, considerem important comentar que la cinquena data de naixement més comuna entre els homes és el 2028, any que encara no ha tingut lloc, de manera que creiem que hi ha un error amb aquesta data.

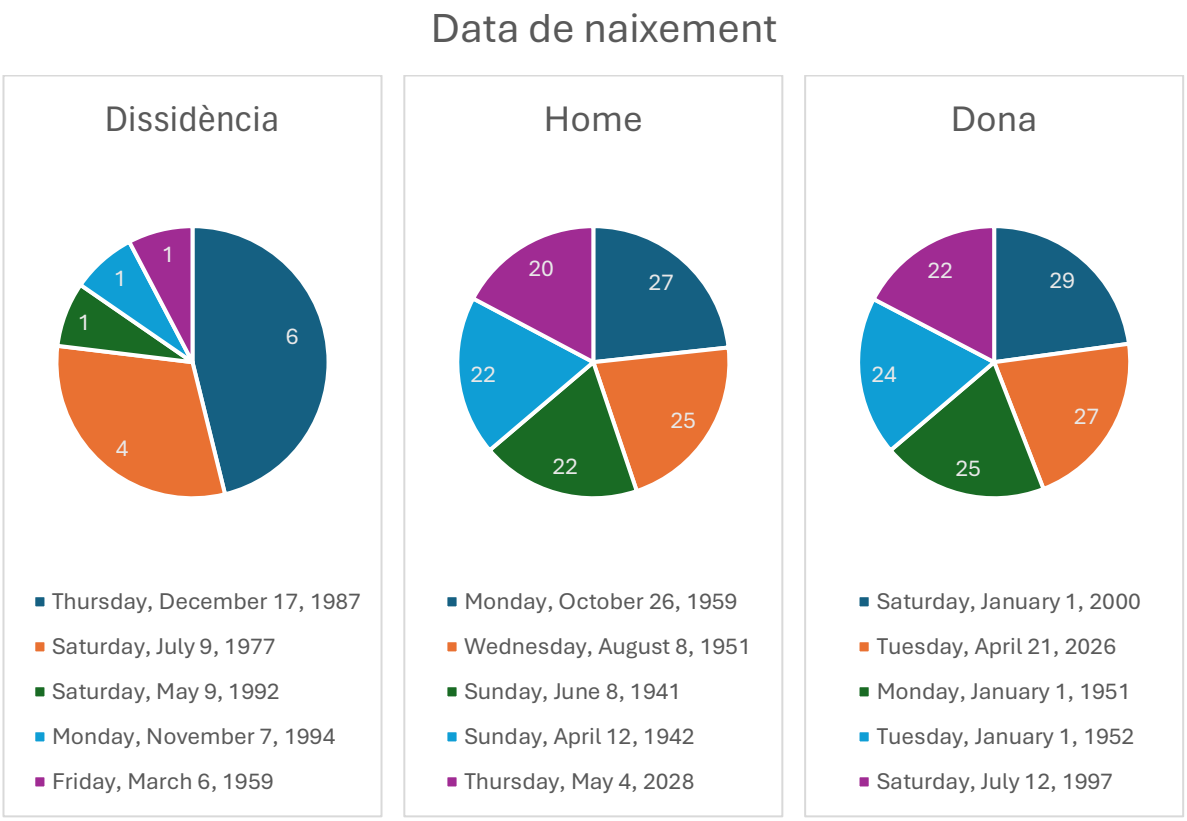

Figura 27. Deu dates de naixement més habituals en funció del gènere.

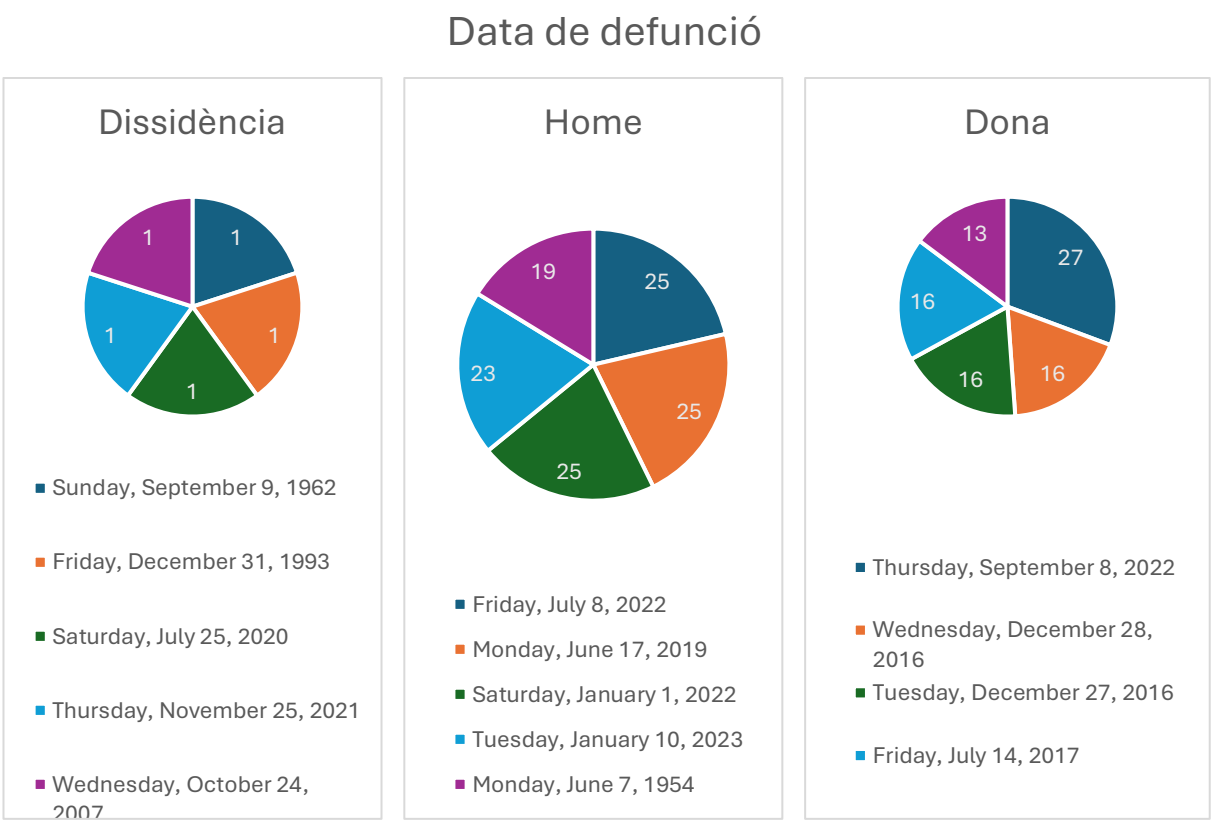

Figura 28. Deu dates de defunció més habituals en funció del gènere.

Pel que fa l'anàlisi de la propietat de període històric segons el gènere, només el podem fer sobre homes i dones, atès que no hi ha dades sobre les persones de gènere dissident. Per a fer-lo, utilitzem les 317 biografies sobre homes i les 62 sobre dones que tenen aquesta informació. Segons es pot veure a la Figura 30, la majoria d'homes que apareixen en portada són de l'Imperi Romà, bé sigui de l'Alt Imperi Romà (59), del Baix Imperi Romà (47) o sense distinció (35), seguits de 21 homes de l'Antic Egipte i 19 de l'Imperi Bizantí. En canvi, la majoria de dones (22) pertanyen al segle XXI, seguides amb un salt a 8 dones del segle XX, 7 de l'Imperi Bizantí, 4 de l'època victoriana i 4 més del Baix Imperi Romà. És a dir, la majoria d'homes que apareixen a les portades de l'edició en anglès de Wikipedia són de l'antiguitat, mentre que les dones tendeixen a ser de períodes històrics més recents o inclús de l'època contemporània, fet que segurament es deu al paper al qual s'han relegat les dones al llarg de la història o a les feines i activitats que se'ls ha permès fer, hipòtesi que pren més força quan veiem el salt entre les dones del segle XXI i les de segles anteriors, a diferència dels resultats sobre els homes.

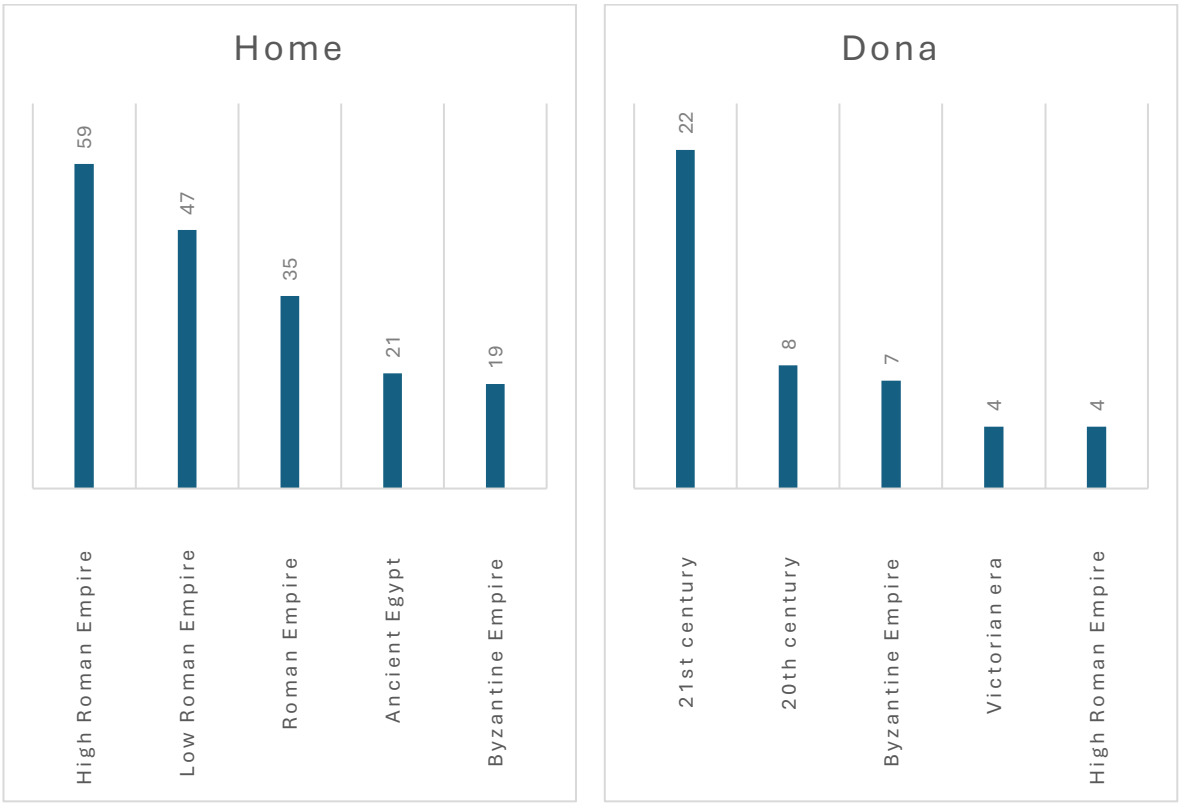

Figura 29. Cinc períodes històrics més habituals dels homes i de les dones que apareixen en portada.

A continuació, proporcionem una taula resum de les biografies analitzades i dels valors recollits de cada propietat.

Propietat	Biografies analitzades	Valors recollits
Instance of	99872	6593
Biografies analitzades	22303	
Sex or gender	19992	9
Personal pronoun	0	0
Sexual orientation	287	9
Ethnic group	1543	193
Religion or worldview	5485	158
Native Language	3436	125
Languages spoken, written or signed	13127	198
Occupation	21007	1153
Father	5715	2883
Mother	4351	2030
Spouse	6458	3170
Award received	8050	1643
Place of birth	18917	6617
Place of death	12150	3705
Country of citizenship	18988	490
Continent	0	0
Date of birth	12972	7143
Date of death	10363	6571
Time period	425	50

Taula 4. Resum de les biografies analitzades i dels valors recollits de cada propietat.

Segons es pot veure a la taula anterior, cap biografia conté les dades de totes les propietats analitzades. Això es reflecteix en el fet que cap de les propietats arriba a les 22303 biografies analitzades. Per tant, cap biografia ens permet conèixer totes les característiques identitàries de la persona segons els eixos interseccionals aplicats. A més, com hem observat en alguns resultats, Wikidata només conté dades dissidents, obviant les majoritàries. Això fa que les biografies incloguin només trets identitaris minoritaris en lloc de totes les dades, independentment de si són identitats privilegiades o oprimides. Per això, considerem necessari adoptar un esquema de metadades mínimes que asseguri que totes les biografies continguin la informació essencial per descriure les característiques identitàries bàsiques de cada persona.

Proposta d'esquema de metadades

En aquesta secció plantegem la nostra proposta d'esquema de metadades necessàries per una biografia en el marc de Wikipedia. En primer lloc, comentarem breument la importància d'adoptar un esquema de metadades; en segon lloc, plantejarem la nostra proposta d'*entity schema* i n'explicarem les propietats de Wikidata que la conformen, i, en darrer terme, il·lustrarem la proposta amb un exemple.

Adoptar un esquema de metadades i propietats mínimes necessàries (o *entity schema*) per biografies assegura la consistència i homogeneïtat de les dades, i millora la precisió de les cerques i facilita el manteniment i actualització del sistema de cares als sistemes de recuperació de la informació. A més, promou la interoperabilitat amb altres sistemes, augmenta l'eficiència en la recuperació d'informació i garanteix que apareguin les dades necessàries, oferint una experiència de cerca eficaç per a les persones usuàries i pels sistemes de recuperació de la informació.

Com diu Leclerc-Olive (2009, p. 12), «no se trata de explorar todo lo que sucedió en el curso de la vida, sino de "comprender" su trayectoria a partir de los acontecimientos más esenciales que la han determinado». Tenint en compte aquesta informació, a continuació presentem la proposta d'esquema de metadades i propietats mínimes necessàries per a la creació de biografies a Wikipedia. Per a fer-ho, ens basem en la proposta d'*entity schema* per a la classe humà (Entity Schema E1074), i a partir d'aquesta base, plantegem les propietats de Wikidata que considerem necessàries que constin en una biografia per assegurar-ne la qualitat.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

start = @<biography>

<biography> EXTRA wdt:P31 {
  wdt:P31 [wd:Q36279];
  wdt:P18 . * ;                    # image (portrait)
  wdt:P21 [wd:Q48270 wd:Q48279 wd:Q179294 wd:Q189125 wd:Q207959 wd:Q301702 wd:Q350374
wd:Q505371 wd:Q660882 wd:Q746411 wd:Q859614 wd:Q1052281 wd:Q1097630 wd:Q1289754
wd:Q1399232 wd:Q2449503 wd:Q3177577 wd:Q3277905 wd:Q6581072 wd:Q6581097 wd:Q7130936
wd:Q12964198 wd:Q15145778 wd:Q15145779 wd:Q18116794 wd:Q27679684 wd:Q27679766
wd:Q52261234 wd:Q93954933 wd:Q93955709 wd:Q96000630 wd:Q25388691 wd:Q56315990]?;   #
gender
  wdt:P6553 . *;                   # personal pronoun
  wdt:P19 . ?;                     # place of birth
  wdt:P20 . ?;                     # place of death
  wdt:P569 . ? ;                   # date of birth
  wdt:P570 . ? ;                   # date of death
  wdt:P2348 . ?;                   # time period
  wdt:P735 . ? ;                   # given name
  wdt:P734 . ? ;                   # family name
  wdt:P1559 . ? ;                  #name in native language
```

```
  wdt:P106 . * ;                  # occupation
  wdt:P27 @<country> *;           # country of citizenship
  wdt:P22 @<human> *;             # father
  wdt:P25 @<human> *;             # mother
  wdt:P3373 @<human> *;           # sibling
  wdt:P26 @<human> *;             # spouse
  wdt:P40 @<human> *;             # child
  wdt:P1038 @<human> *;           # relative
  wdt:P103 @<language> *;         # native language
  wdt:P1412 @<language> *;        # languages spoken, written or signed
  wdt:P172 . ? ;                  # ethnic group
  wdt:P91 . ? ;                   # sexual orientation
  wdt:P140 . * ;                  # religion or worldview
  wdt:P800 . * ;                  # notable work
  wdt:P793 . * ;                  # significant event
  wdt:P166 . * ;                  # award received
  wdt:P8371 . * ;                 # references work, tradition or theory
  wdt:P135 . * ;                  # movement
  rdfs:label rdf:langString+;
}

<country> EXTRA wdt:P31 {
  wdt:P31 [wd:Q6256 wd:Q3024240 wd:Q3624078] +;
}

<language> EXTRA wdt:P31 {
  wdt:P31 [wd:Q34770 wd:Q1288568] +;
}
```

Com es pot veure, la majoria de categories que proposem es corresponen amb les que hem utilitzat per a l'anàlisi de les portades. A més, n'hem afegit d'altres que hem considerat necessàries per representar correctament una persona o per justificar-ne la presència a Wikipedia. Les detallem a continuació:

Nom	Codi	Definició[3]	Justificació
Image	P18	Image of relevant illustration of the subject	La imatge de la persona ajuda a identificar-la fàcilment. A més, pot proporcionar context addicional sobre la seva vida, com l'edat, el gènere, etc., facilitant la comprensió del relat biogràfic a la persona lectora.
Child	P40	Subject has object as child	La història familiar és essencial per comprendre qui som i d'on venim. Els fills, germans i altres familiars

[3] Definicions extretes de la llista de propietats de Wikidata (https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all)

			poden influir en la identitat d'una persona i en els esdeveniments que configuren la seva vida.
Movement	P135	Literary, artistic, scientific or philosophical movement or scene associated with this person or work	La connexió amb un moviment artístic, literari, científic o filosòfic pot revelar les influències que han impactat l'obra i el pensament d'una persona, ja que aquests moviments sovint proporcionen contextos ideològics i estètics que modelen les seves creacions.
Family name	P734	Part of full name of person	El nom i el cognom són elements clau per identificar una persona i distingir-la d'altres.
Given name	P735	First name or another given name of this person	El nom i el cognom són elements clau per identificar una persona i distingir-la d'altres.
Significant event	P793	Significant or notable events associated with the subject	Se puede distinguir entre los "pequeños acontecimientos" –cuya narración no es indispensable para la comprensión de la trayectoria biográfica en su conjunto– y los "acontecimientos significativos", cuya omisión convertiría la historia, en cierta manera, en enigmática. Son los "grandes acontecimientos" los que (...) han constituido momentos de bifurcación o de cambios importantes en la "manera de vivir y de relatar" su vida. (Leclerc-Olive, 2009, pp. 4-5)

Notable work	P800	Notable scientific, artistic or literary work, or other work of significance among subject’s works	Conèixer l'obra notable d'una persona en la seva biografia és essencial per contextualitzar-ne els èxits i entendre'n l'impacte en la societat. Així, aquestes obres revelen les influències i motivacions que han guiat el seu pensament i la seva creativitat, proporcionant una visió més completa de la persona i del seu llegat. A més, ajuden a situar la persona en un context històric i cultural.
Relative	P1038	Family member	La història familiar és essencial per comprendre qui som i d'on venim. Els fills, germans i altres familiars poden influir en la identitat d'una persona i en els esdeveniments que configuren la seva vida.
Name in native language	P1559	Name of a person in their native language	Saber el nom d'una persona en la seva llengua nativa en una biografia és important per reflectir la seva identitat cultural i personal. A més, demostra respecte cap a la persona i la seva cultura, fet que també és important.
Sibling	P3373	The subject and the object have at least one common parent (brother, sister, etc. including half-siblings)	La història familiar és essencial per comprendre qui som i d'on venim. Els fills, germans i altres familiars poden influir en la identitat d'una persona i en els esdeveniments que configuren la seva vida.

References work, tradition or theory	P8371	Creative work, tradition or theory this creative work references by allusion, quote or similar means	És important conèixer les obres, tradicions o teories a què una persona fa referència a la seva biografia per contextualitzar-ne l'obra, comprendre les seves influències i analitzar com adapta aquestes idees al seu propi treball.

Taula 5. Definició de les propietats seleccionades per a l'estudi a partir de Wikidata

Per il·lustrar la proposta, mostrem una aplicació de l'esquema de metadades per a fer una biografia fictícia. Així doncs, si, per exemple, volguéssim escriure la biografia de Maria Soler amb aquesta informació:

P31:biography: Maria Soler
P18:image: [Retrat de Maria Soler]
P21:gender: Femení
P6553:personal pronoun: Ella
P19:place of birth: Barcelona, Catalunya
P20:place of death: París, França
P569:date of birth: 15 de gener de 1925
P570:date of death: 10 d'abril de 2005
P2348:time period: Segle XX
P735:given name: Maria
P734:family name: Soler
P1559:name in native language: Maria Soler
P106:occupation: Novel·lista, Dramaturga
P27:country of citizenship: Espanya
P22:father: Joan Soler
P25:mother: Maria Borrell
P3373:sibling: Arnau Soler
P26:spouse: Valentí Anglada
P40:child: Emili Soler Anglada
P1038:relative: Pere Soler (oncle)
P103:native language: Català

P1412:languages spoken, written or signed: Català, Castellà, Francès

P172:ethnic group: Caucàsic

P91:sexual orientation: Heterosexual

P140:religion or worldview: Agnòstica

P800:notable work: *Deu narracions*

P793:significant event: Premi Sant Jordi de Novel·la

P166:award recived: Premi Sant Jordi de Novel·la

P8371:references work, tradition or theory: Literatura Postmoderna

P135:movement: Postmodernisme

L'esquema de metadades proposat quedaría així:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

start = @<biography>

<biography> EXTRA wdt:P31 {
  wdt:P31 [wd:Q36279 wd:Q5];
  wdt:P18 . * ;                                  # image (portrait)
  wdt:P21 [wd:Q6581072]?;                        # gender
  wdt:P6553 [wd:Q1270787] *;                     # personal pronoun
  wdt:P19 [wd:Q1492]  ?;                         # place of birth
  wdt:P20 [wd:Q90]  ?;                           # place of death
  wdt:P569 [wd:Q69265756]   ? ;                  # date of birth
  wdt:P570 [wd:Q22663058]  ? ;                   # date of death
  wdt:P2348 [wd:Q6927] ?;                        # time period
  wdt:P735 [wd:Q325872]  ? ;                     # given name
  wdt:P734 [wd:Q30330126]  ? ;                   # family name
  wdt:P1559 [wd:Q116029301]  ? ;                 #name in native language
  wdt:P106 [wd:Q6625963 wd:Q487596]  * ;         # occupation
  wdt:P27 @<country> [wd:Q29] *;                 # country of citizenship
  wdt:P22 @<human> [wd:Q116029302] *;            # father
  wdt:P25 @<human> [wd:Q116029302] *;            # mother
  wdt:P3373 @<human> *;                          # sibling
  wdt:P26 @<human> [wd:Q115942461] *;            # spouse
  wdt:P40 @<human> [wd:Q110520174]  *;           # child
  wdt:P1038 @<human> [wd:Q112546180] *;          # relative
  wdt:P103 @<language> [wd:Q7026] *;             # native language
  wdt:P1412 @<language> [wd:Q7026 wd:Q1321 wd:Q150] *;
                                                 # languages spoken, written or signed
  wdt:P172 [wd:Q7129609]  ? ;                    # ethnic group
  wdt:P91 [wd:Q1035954]  ? ;                     # sexual orientation
  wdt:P140 [wd:Q288928]  * ;                     # religion or worldview
  wdt:P800 . * ;                                 # notable work
  wdt:P793 [wd:Q2233927]  * ;                    # significant event
  wdt:P166 [wd:Q2233927]  * ;                    # award received
  wdt:P8371 [wd:Q113013]  * ;                    # references work, tradition or theory
  wdt:P135 [wd:Q47783]  * ;                      # movement
```

```
  rdfs:label rdf:langString+;
}

<country> EXTRA wdt:P31 {
  wdt:P31 [wd:Q6256 wd:Q3024240 wd:Q3624078] +;
}

<language> EXTRA wdt:P31 {
  wdt:P31 [wd:Q34770 wd:Q1288568] +;
}
```

Conclusions

Aquest estudi ha proporcionat una anàlisi amb perspectiva de gènere i interseccional de les bretxes de contingut i de la diversitat de representació a les portades de l'edició en anglès de Wikipedia, amb focus especial en les biografies. Els resultats han demostrat que, tot i la missió d'oferir un accés lliure i equitatiu al coneixement, existeixen desigualtats notables en la representació de gèneres, ètnies, orientacions sexuals, orígens geogràfics i altres aspectes identitaris a la plataforma.

Segons hem pogut veure, els trets dins dels eixos interseccionals en situació de privilegi (Rodó-Zárate, 2021) són els que no tenen la metadada recollida. És a dir, Wikidata inclou en les interseccionalitats les dades dissidents, però obvia les majoritàries, donant a entendre que aquestes últimes són la norma i que, si no s'indica el contrari, les persones responen a aquests trets identitaris. Per tant, si es vol que Wikidata, i conseqüentment Wikipedia, siguin igualitàries, diverses i inclusives, és essencial recollir les dades referents a qualsevol orientació sexual, ètnia, i altres eixos interseccionals. És per això que proporcionem un esquema de metadades i propietats mínimes necessàries amb la voluntat de recollir les metadades bàsiques per descriure un ésser humà.

Les biografies ocupen aproximadament una quarta part dels articles que apareixen publicats diàriament a la portada de la Wikipedia en anglès a les seccions "From today's featured article", "Did you know..." i "On this day"; la resta d'articles fan referència a d'altres accepcions no relacionades amb entrades de persones. Dins d'aquestes biografies, hi ha una clara predominança d'homes i una representació desigual de dones i persones no-binàries. L'anàlisi de l'orientació sexual ha revelat una presència destacada de persones no-heterosexuals, seguides de persones heterosexuals, persones que no han etiquetat la seva orientació i una persona asexual. Pel que fa a l'ètnia i la religió, s'ha trobat una gran diversitat de resultats, però també una concentració en certs grups i religions, amb una notable subrepresentació d'altres.

La distribució geogràfica de les persones representades mostra una clara inclinació cap a les grans ciutats dels Estats Units i Europa, destacant una preponderància del Nord Global i una infrarepresentació d'Àfrica, Amèrica del Sud i Amèrica Central. Això es reflecteix també en les nacionalitats i els continents d'origen de les persones destacades. Aquesta inclinació pot afectar la percepció pública i la visibilitat de persones de regions menys representades i, per tant, cal plantejar mesures per revertir el biaix geogràfic.

L'anàlisi de les professions i dels períodes històrics més representats evidencia una tendència a destacar figures polítiques, escriptors i professionals del sector artístic, especialment del segle XIX i XX. Això podria limitar la diversitat de narratives i perspectives presentades a la plataforma, motiu pel qual convé analitzar aquesta bretxa amb més profunditat i plantejar-ne solucions.

Aquest projecte ha posat de manifest la necessitat d'una major inclusió i equitat en la representació de les persones a Wikipedia. Per fer front al biaix de contingut que s'ha detectat, recomanem l'adopció de l'esquema estandarditzat de metadades proposat, amb la finalitat d'assegurar una descripció completa i precisa de les persones. Al mateix temps, subscrivim la recomanació d'alguns dels estudis analitzats de participar en projectes i fer esforços conscients per destacar una major diversitat de persones i perspectives a les portades de Wikipedia. Això contribuirà a crear una plataforma més justa, representativa i inclusiva, reflectint millor la diversitat humana i promovent una cultura del coneixement més equitativa i representativa per a tothom.

Bibliografia

Aliaga, Juan Vicente; Mayayo, Patricia (ed.) (2013). *Genealogías feministas en el arte español: 1960-2010*. This Side Up.

Bejarano, Andrés (2023). *El sesgo cultural en las portadas de la Wikipedia en español e inglés: un estudio sobre colonialismo digital* (Treball Final de Màster). Universitat de Barcelona, Catalunya.

Beytía, Pablo; Wagner, Claudia (2022). Visibility layers: a framework for systematising the gender gap in Wikipedia content. *Internet Policy Review*, *11* (1). https://doi.org/10.14763/2022.1.1621

Centelles, Miquel; Ferran-Ferrer, Núria (2024). Assessing knowledge organization systems from a gender perspective: Wikipedia taxonomy and Wikidata ontologies. *Journal of Documentation, 80* (7), pp. 124-147. https://doi.org/http://10.1108/JD-11-2023-0230

Expósito, Carmen (2012). ¿Qué es eso de la interseccionalidad? Aproximación al tratamiento de la diversidad desde la perspectiva de género en España. *Investigaciones Feministas, 3*, pp. 203-222.

Fahimnia, Fatemeh; Damerchiloo, Mansoureh; Khandan, Mohammad; Eltemasi, Mahshid (2022). A Framework for Assessing the Quality of Wikipedia Articles: A Meta-Synthesis of the Literature. *International Journal of Information Science and Management, 20* (1), pp. 91-118.

Fan, Angela; Gardent, Claire (2022). Generating Full Length Wikipedia Biographies. The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies. A *60th Annual Meeting of the Association for Computational Linguistics, 2*, pp. 8561-8576. Dublin, Ireland. 10.18653/v1/2022.acl-long.586

Ferran-Ferrer, Núria; Boté-Vericad, Juan-José; Minguillón, Julià (2023). Wikipedia gender gap: a scoping review. *Profesional de la información, 32* (6). https://doi.org/10.3145/epi.2023.nov.17

Ferran-Ferrer, Núria; Miquel-Ribé, Marc; Meneses, Julio; Minguillón, Julià (2022). The Gender Perspective in Wikipedia: A Content and Participation Challenge. A *Companion Proceedings of the Web Conference 2022 (WWW '22). Association for Computing Machinery*, pp. 1319-1323. New York, USA. https://doi.org/10.1145/3487553.3524937

Grant, Maria J.; Booth, Andrew (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26* (2), pp. 89-168.

Konieczny, Piotr; Klein, Maximilian (2018). Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator. *New Media & Society, 20* (12), pp. 4608-4633. https://doi.org/10.1177/1461444818779080

Leclerc-Olive, Michèle (2009). Temporalidades de la experiencia: las biografías y sus acontecimientos. Iberófo*rum. Revista de Ciencias Sociales de la Universidad Iberoamericana, 4* (8), pp. 1-39. https://www.redalyc.org/articulo.oa?id=211014822001

Lewoniewski, Włodzimierz; Węcel, Krzysztof; Abramowicz, Witold (2019). Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. *Computers, 8* (3). https://doi.org/10.3390/computers8030060

Martini, Franziska (2023). Notable enough? The questioning of women's biographies on Wikipedia. *Feminist Media Studies*, pp. 1-17. https://doi.org/10.1080/14680777.2023.2266585

Meyer, Christine (2022). *"If You Want to Change the World, Edit Wikipedia": Mitigating the Gender Gap and Systemic Bias on Wikipedia* (Tesi Doctoral). University of Idaho, USA.

Miquel-Ribé, Marc; Laniado, David (2020). The Wikipedia Diversity Observatory. A project to identify and bridge content gaps in Wikipedia. A *Proceedings of the International Symposium on Open Collaboration (OpenSym 2020)*. ACM, New York, NY, USA.

Miquel-Ribé, Marc; Laniado, David (2021). The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces. *Journal of Internet Services and Applications, 12* (10).

Moyano, Jun (Dir.). (2023). *Guia gramatical de llenguatge no-binari* (2a ed.). Raig Verd Editorial.

Perez-Vaisvidovsky, Nadav (2019). Enemies, allies or citizens? The subject positions of men in the making of birth leave for fathers in Israel. *Families, Relationships and Societies, 20* (10), pp. 1-19.

Rodó-Zárate, Maria (2021). *Interseccionalitat. Desigualtats, llocs i emocions*. Tigre de Paper.

Roy, Dwaipayan; Bhatia, Sumit; Jain, Prateek (2022). Information asymmetry in Wikipedia across different languages: A statistical analysis. *Association for Information Science and Technology, 73* (3), pp. 347-361. https://doi.org/10.1002/asi.24553

Sefidari, Maria (2022). Equidad de conocimiento y sesgos: un análisis cuantitativo del contenido destacado en la Portada de Wikipedia. *IC Revista Científica de Información y Comunicación, 19*, pp. 141-163. https://dx.doi.org/10.12795/IC.2022.I19.07

Stranisci, Marco Antonio; Damiano, Rossana; Mensa, Enrico; Patti, Viviana; Radicioni, Daniele; Caselli, Tommaso (2023). WibiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. A *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, *1*, pp. 12370-1284. Toronto, Canada. 10.18653/v1/2023.acl-long.691

Tripodi, Francesca (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *Sage, 25* (7), pp. 1687-1707. https://doi.org/10.1177/14614448211023772

Wikipedia:Prime objective. Recuperat 6 abril 2024, de https://en.wikipedia.org/wiki/Wikipedia:Prime_objective

Annex 1. Índex de figures i de taules

Índex de figures

Índex de taules

Annex 2. *Scoping review*

Enllaç al document Annex 2: Annex 2. Scoping Review.xlsx

Annex 3. Extracció de contingut de les portades

Enllaç al document Annex 3: Annex 3. Extraccion_contenido_portadas_wikipedia.pdf

Annex 4. Scraping

Enllaç al document Annex 4: Annex 4. Scraping.xlsx