¿SESGOS DEL CONTENIDO POR PARES?

David Abián

esLibre 2023

WIKIPEDIA
The Free Encyclopedia

WIKIDATA

stackoverflow

SITIOS DE PRODUCCIÓN POR PARES

participación heterárquica y **distribuida**

anti-credencialismo y equipotencialidad a priori

holoptismo horizontal (todos se ven) y vertical (se ve todo)

reciprocidad **voluntaria** de contribución/explotación

coordinación para superar el valor de las contribuciones individuales
(p. ej. funcionalidades o actividades para prevenir o corregir redundancias)

SITIOS DE PRODUCCIÓN POR PARES

artefactos de contenido **cooperativos**

artefactos de contenido **competitivos**

coordinación para superar el valor de las contribuciones individuales
(p. ej. funcionalidades o actividades para prevenir o corregir redundancias)

¿VACÍOS EN LA PRODUCCIÓN POR PARES?

¿vacíos en contenido?

¿vacíos en participación?

¿vacíos en explotación?

…

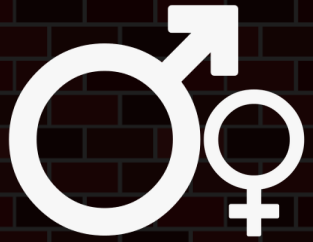
¿VACÍOS EN LA PRODUCCIÓN POR PARES?

→ **¿vacíos en contenido?** ←

¿vacíos en participación?

¿vacíos en explotación?

…



diferencias
por género

18th 19th 20th

diferencias
por época

diferencias
geográficas y
socioeconómicas

...

DATOS

Todas las Wikipedias con al menos 20 000 biografías
tienen más artículos sobre varones que sobre mujeres.

~**77%** de biografías de **Wikipedia en español** son sobre varones.

~**80%** de biografías de **Wikipedia en inglés** son sobre varones.

~**76%** de elementos de **Wikidata** sobre personas son sobre varones.

DATOS

Todas las Wikipedias con al menos 20 000 biografías tienen más artículos sobre varones que sobre mujeres.

~**77%** de biografías de **Wikipedia en español** son sobre varones.

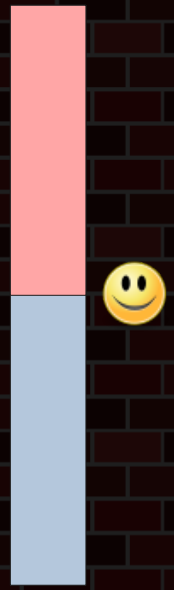
~**80%** de biografías de **Wikipedia en inglés** son sobre varones.

~**76%** de elementos de **Wikidata** sobre personas son sobre varones.

¿Lo deseable?

¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?

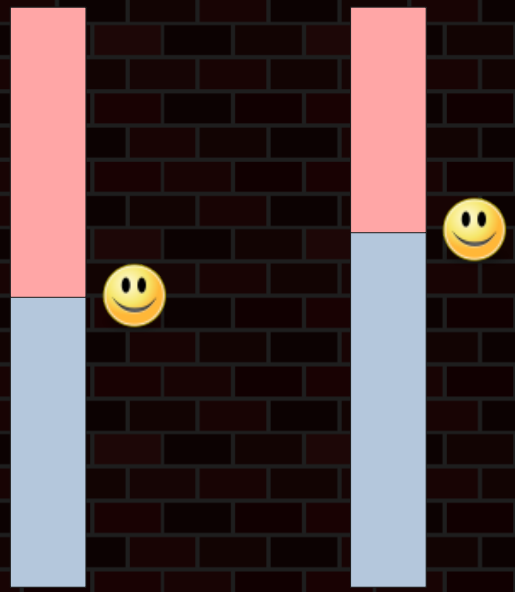
¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?



criterio

50-50

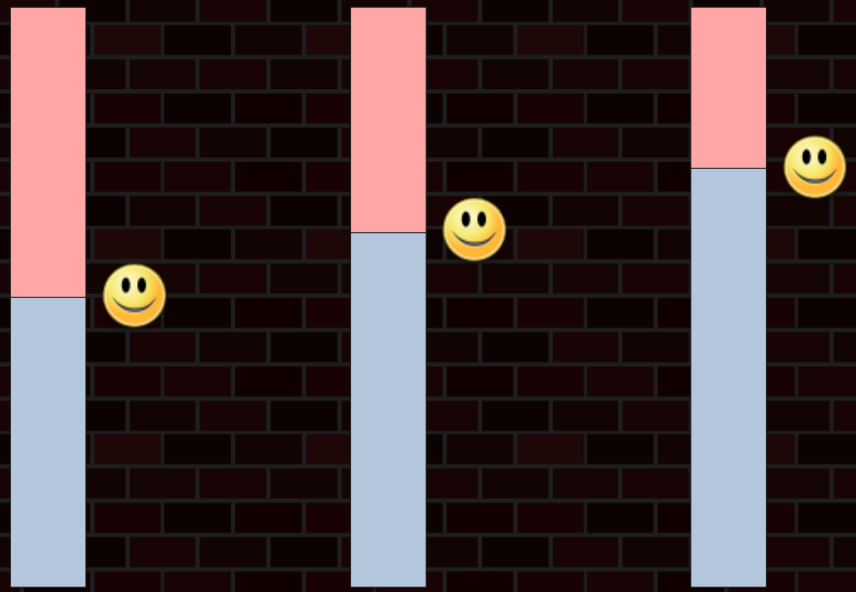
¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?



criterio
50-50

criterio
relevancia

¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?

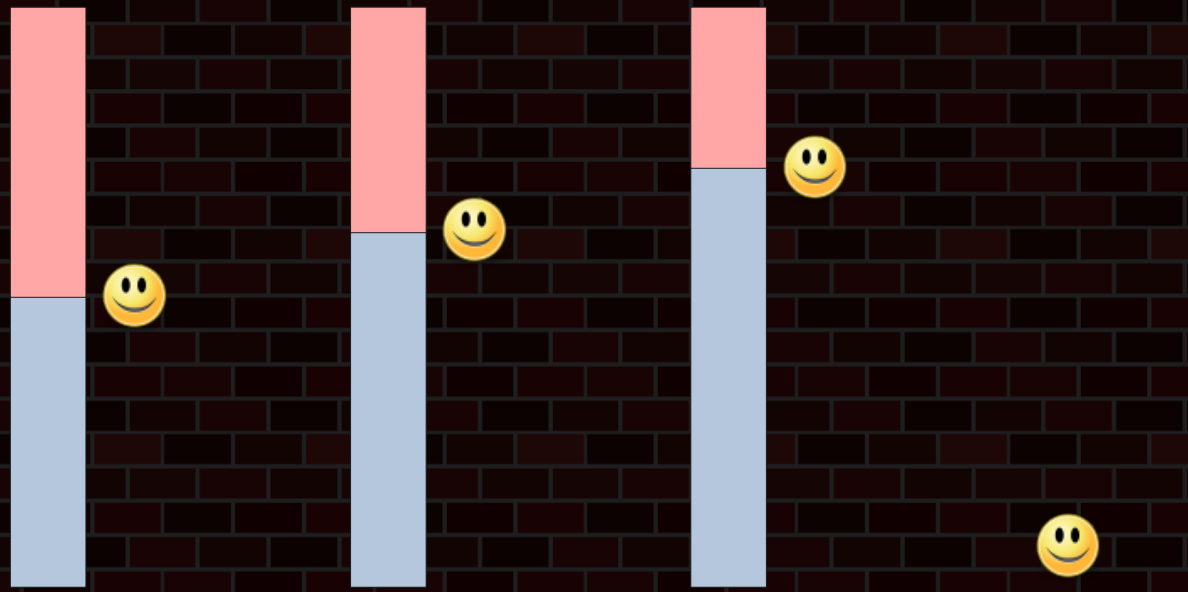


criterio *50-50*

criterio *relevancia*

criterio *relevancia con info. disponible*

¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?



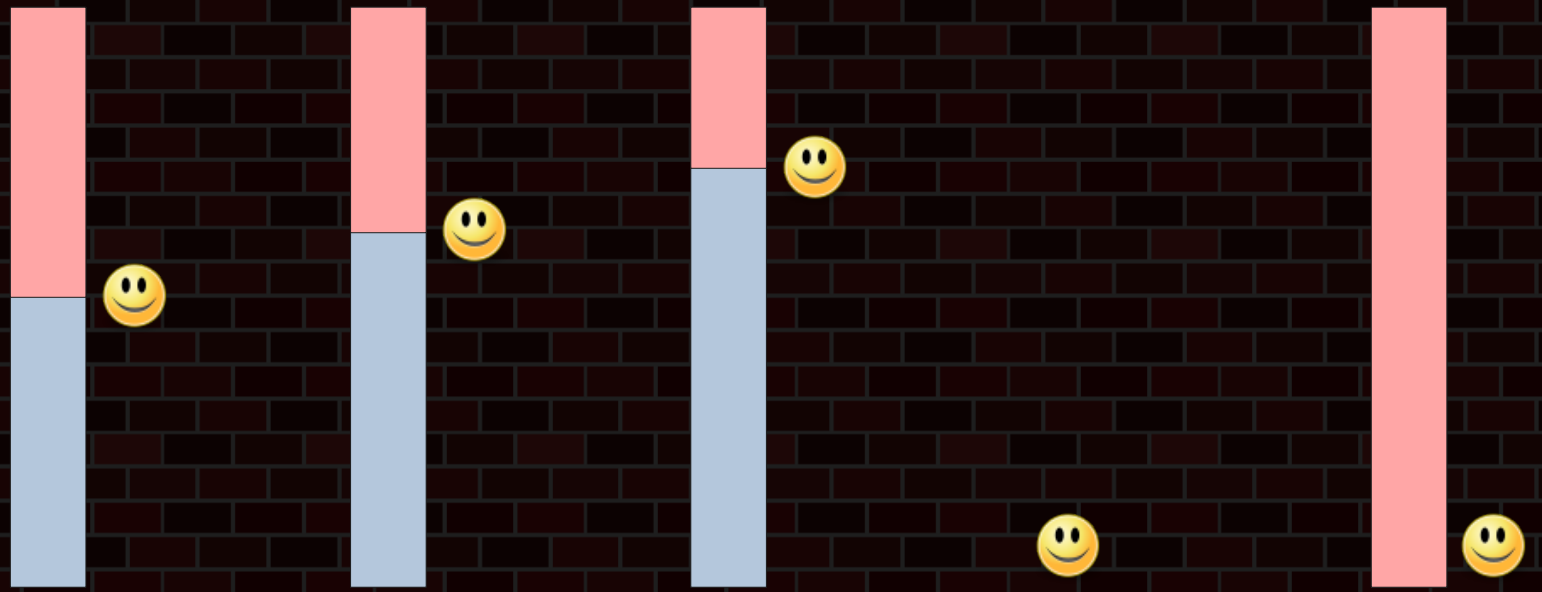
criterio *50-50*

criterio *relevancia*

criterio *relevancia con info. disponible*

criterio *conformismo*

¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?



criterio *50-50*

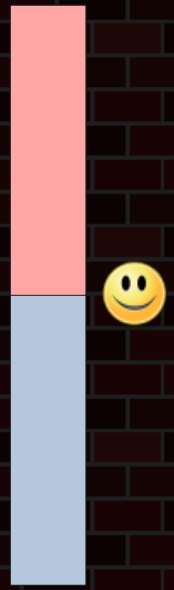
criterio *relevancia*

criterio *relevancia con info. disponible*

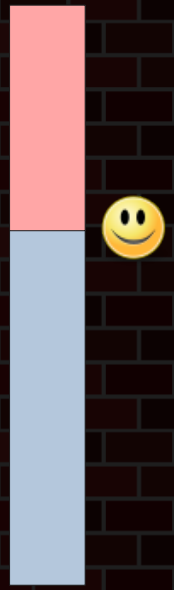
criterio *conformismo*

criterio *inconformismo*

¿EN QUÉ WIKIPEDIA IDEAL NO HABRÍA «BRECHA DE GÉNERO»?



criterio *50-50*



criterio *relevancia*

criterio *relevancia con info. disponible*

criterio *conformismo*

criterio *inconformismo*

… y muchos otros escenarios ideales respetables pero mutuamente incompatibles

CONCLUSIÓN RÁPIDA

En cualquier producto de información podemos afirmar o negar que existe un sesgo o vacío de contenido cualquiera según algún conjunto de criterios y principios respetables que justifiquen el estado ideal de referencia.

DATOS

Todas las Wikipedias con al menos 20 000 biografías
tienen más artículos sobre varones que sobre mujeres.

~**77%** de biografías de **Wikipedia en español** son sobre varones.

~**80%** de biografías de **Wikipedia en inglés** son sobre varones.

~**76%** de elementos de **Wikidata** sobre personas son sobre varones.

DATOS

Todas las Wikipedias con al menos 20 000 biografías tienen más artículos sobre varones que sobre mujeres.

~**77%** de biografías de **Wikipedia en español** son sobre varones.

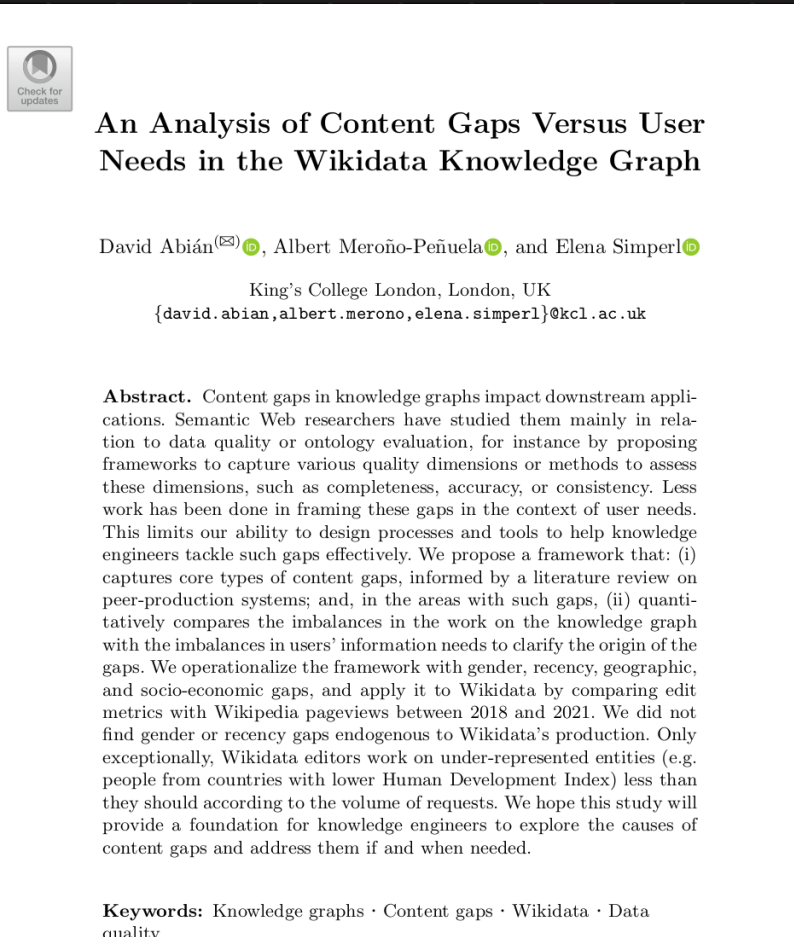
~**80%** de biografías de **Wikipedia en inglés** son sobre varones.

~**76%** de elementos de **Wikidata** sobre personas son sobre varones.

¿Lo deseable?
Tú decides.

¿ENDÓGENAS O EXÓGENAS?

¿ENDÓGENAS O EXÓGENAS?



An Analysis of Content Gaps Versus User Needs in the Wikidata Knowledge Graph

David Abián(✉), Albert Meroño-Peñuela, and Elena Simperl

King's College London, London, UK
{david.abian,albert.merono,elena.simperl}@kcl.ac.uk

Abstract. Content gaps in knowledge graphs impact downstream applications. Semantic Web researchers have studied them mainly in relation to data quality or ontology evaluation, for instance by proposing frameworks to capture various quality dimensions or methods to assess these dimensions, such as completeness, accuracy, or consistency. Less work has been done in framing these gaps in the context of user needs. This limits our ability to design processes and tools to help knowledge engineers tackle such gaps effectively. We propose a framework that: (i) captures core types of content gaps, informed by a literature review on peer-production systems; and, in the areas with such gaps, (ii) quantitatively compares the imbalances in the work on the knowledge graph with the imbalances in users' information needs to clarify the origin of the gaps. We operationalize the framework with gender, recency, geographic, and socio-economic gaps, and apply it to Wikidata by comparing edit metrics with Wikipedia pageviews between 2018 and 2021. We did not find gender or recency gaps endogenous to Wikidata's production. Only exceptionally, Wikidata editors work on under-represented entities (e.g. people from countries with lower Human Development Index) less than they should according to the volume of requests. We hope this study will provide a foundation for knowledge engineers to explore the causes of content gaps and address them if and when needed.

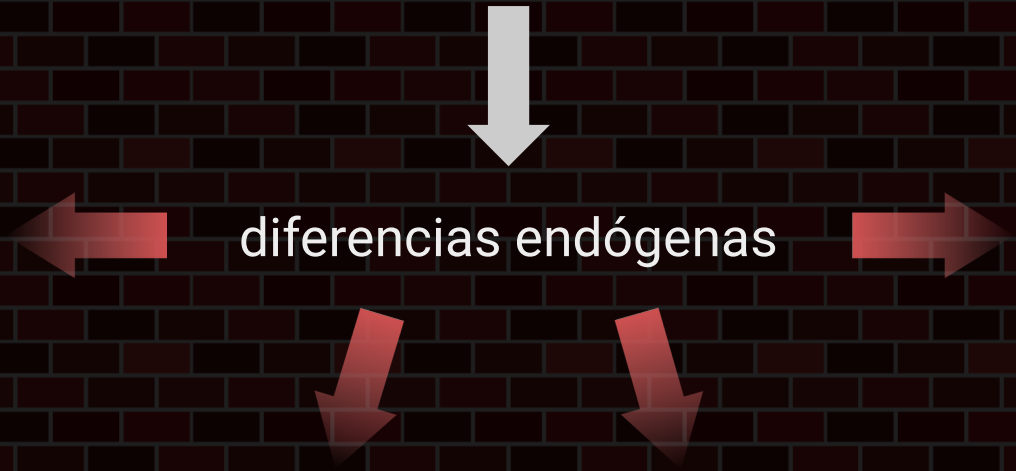
Keywords: Knowledge graphs · Content gaps · Wikidata · Data quality

¿ENDÓGENAS O EXÓGENAS?

$$\text{retorno de la inversión (}\mathbf{ROI}\text{)} = \frac{\text{necesidades de información}}{\text{contribución}}$$

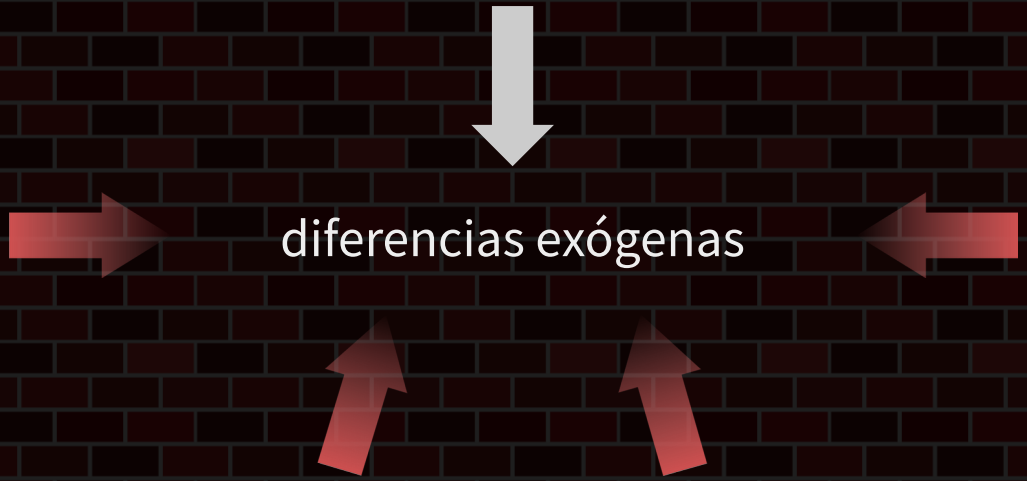
¿ENDÓGENAS O EXÓGENAS?

H: ROI de lo **sobrerrepresentado** < ROI de lo **infrarrepresentado**



¿ENDÓGENAS O EXÓGENAS?

H: ROI de lo **sobrerrepresentado** > ROI de lo **infrarrepresentado**



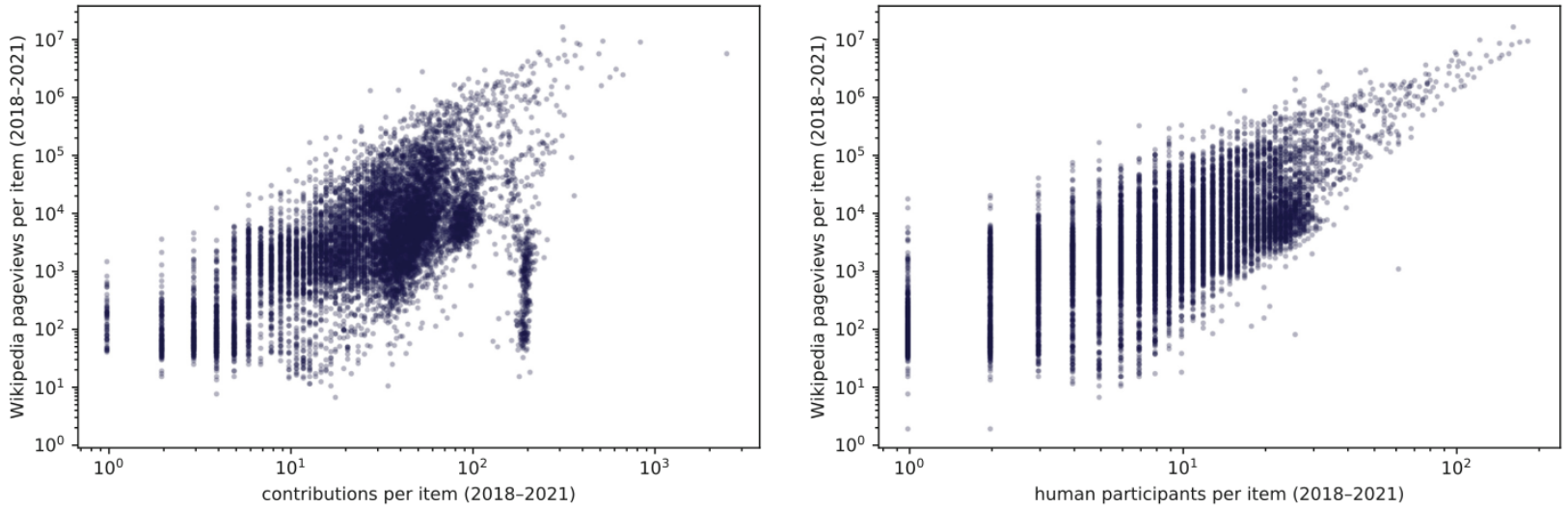
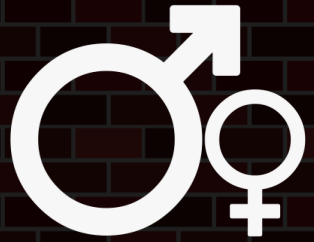


Fig. 1. Random sample of Wikidata items on settlements in two charts as a function of Wikipedia pageviews and two different contribution metrics: on the left, number of (manual and automatic) contributions, revealing clusters of similar entities that received the same automatic treatment (e.g. vertical line around 200 contributions); on the right, number of human editors, which does not provide this insight but better quantifies the actual effort invested and therefore better correlates with pageviews

¿ENDÓGENAS O EXÓGENAS?



diferencias
por género

18th 19th 20th

diferencias
por época

diferencias
geográficas y
socioeconómicas

50,000 artefactos
sobre **personas**

diferencias por género

- género

diferencias por época

- año nacimiento
- año fallecim.

diferencias geográficas y socioeconómicas

- país (IDH…)

50,000 artefactos
sobre **asentamientos**

diferencias geográficas y socioeconómicas

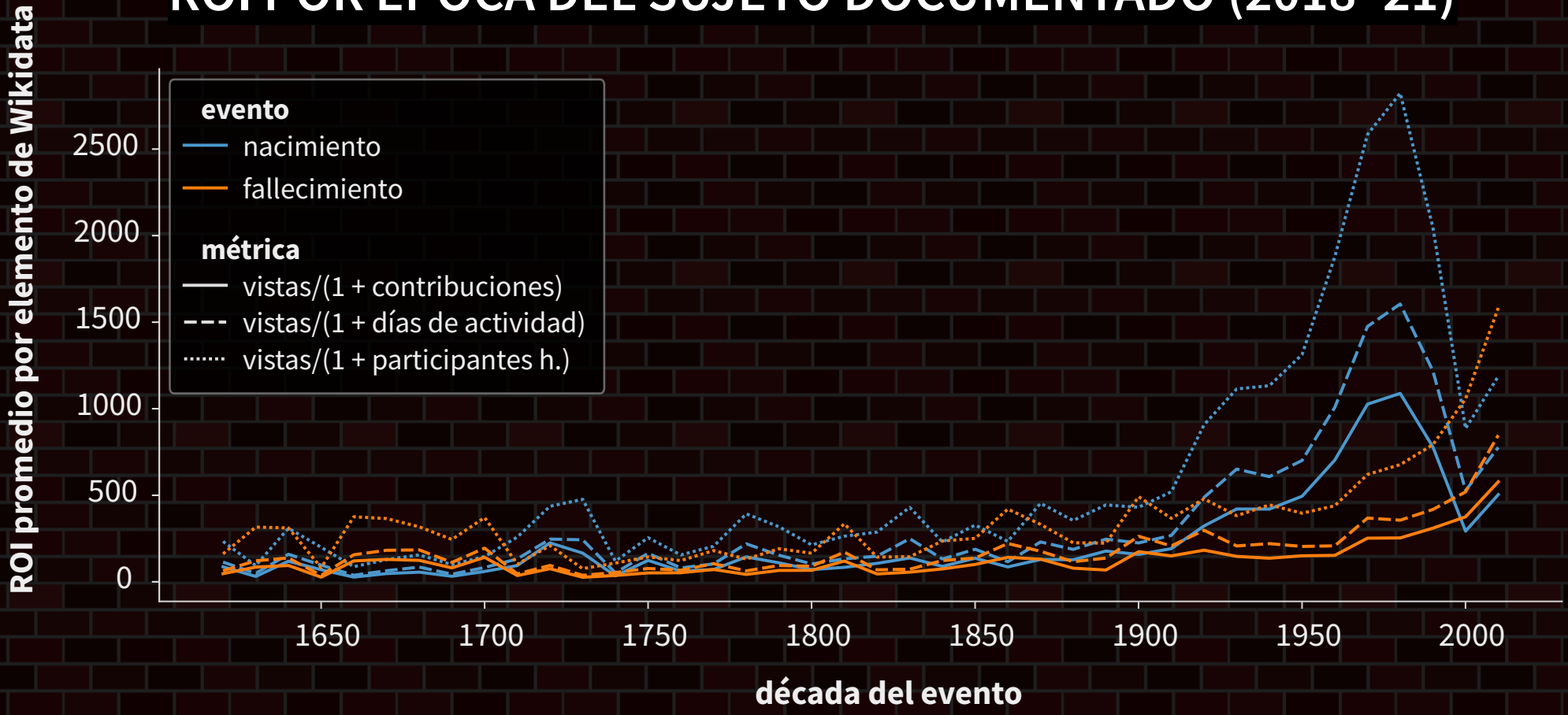
- población
- país (IDH…)

todos los artefactos
sobre **países**

diferencias geográficas y socioeconómicas

- IDH…

ROI POR ÉPOCA DEL SUJETO DOCUMENTADO (2018–21)



POR ÉPOCA DEL SUJETO DOCUMENTADO (2018–21)

corr(año nacimiento ⇔ contribución)	0 0 0
corr(año nacimiento ⇔ demanda)	+
corr(año nacimiento ⇔ ROI)	**+ + +**
corr(año fallecimiento ⇔ contribución)	+ + 0
corr(año fallecimiento ⇔ demanda)	+
corr(año fallecimiento ⇔ ROI)	**+ + +**

significación $p < 0.01$, tamaño de efecto mínimo $r > 0.1$

POR ÉPOCA DEL SUJETO DOCUMENTADO (2018–21)

corr(año nacimiento ⇔ contribución)	+
corr(año nacimiento ⇔ demanda)	+
corr(año nacimiento ⇔ ROI)	+ + +
corr(año fallecimiento ⇔ contribución)	+
corr(año fallecimiento ⇔ demanda)	+
corr(año fallecimiento ⇔ ROI)	+ + +

significación $p < 0.01$, tamaño de efecto mínimo $r > 0.1$

Las diferencias por época son exógenas. Podría rentar invertir aún más en lo actual.

POR POBLACIÓN DEL NÚCLEO DOCUMENTADO (2018–21)

corr(población ⇔ contribución)	+ + +
corr(población ⇔ demanda)	+
corr(población ⇔ ROI)	**+ + +**

significación $p < 0.01$, tamaño de efecto mínimo $r > 0.1$

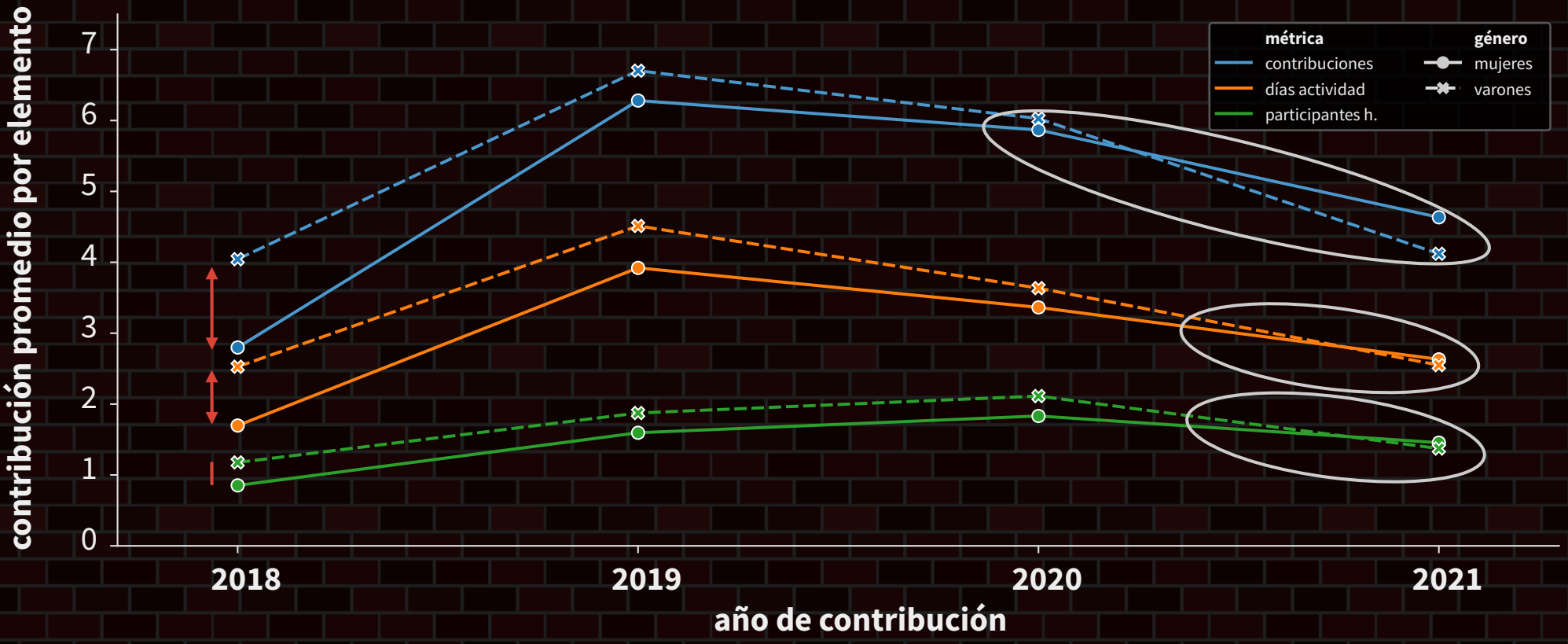
POR POBLACIÓN DEL NÚCLEO DOCUMENTADO (2018–21)

corr(población ⇐ contribución)	+ + +
corr(población ⇦ demanda)	+
corr(población ⇔ ROI)	+ + +

significación p < 0.01, tamaño de efecto mínimo r > 0.1

Las diferencias
entre rural y urbano
son exógenas.
Podría rentar invertir
aún más en lo urbano.

CONTRIBUCIÓN POR GÉNERO DEL SUJETO DOCUMENTADO



POR GÉNERO DEL SUJETO DOCUMENTADO (2018–21)

contrib(varón) > contrib(mujer)	✓ = =
demanda(varón) > demanda(mujer)	✓
ROI(varón) > ROI(mujer)	✓✓✓

significación $p < 0.01$, tamaño de efecto mínimo $r > 0.1$

POR GÉNERO DEL SUJETO DOCUMENTADO (2018–21)

contrib(varón)	> contrib(mujer)
demanda(varón)	> demanda(mujer)
ROI(varón)	> ROI(mujer)

significación p < 0.01, tamaño de efecto mínimo r > 0.1

✓✓✓

Las diferencias por género son exógenas. Podría rentar invertir aún más en contenido sobre varones.

POR IDH DEL PAÍS, NÚCLEO O SUJETO DOCUMENTADO (2018–21)

corr(IDH sujeto ⇔ contribución)	+ + +
corr(IDH sujeto ⇔ demanda)	0
corr(IDH sujeto ⇔ ROI)	**0 0 0**
corr(IDH núcleo ⇔ contribución)	+ + +
corr(IDH núcleo ⇔ demanda)	+
corr(IDH núcleo ⇔ ROI)	**+ + +**
corr(IDH país ⇔ contribución)	+ + +
corr(IDH país ⇔ demanda)	+
corr(IDH país ⇔ ROI)	**+ + +**

significación $p < 0.01$, tamaño de efecto mínimo $r > 0.1$

POR IDH DEL PAÍS, NÚCLEO O SUJETO DOCUMENTADO (2018–21)

corr(IDH sujeto ⇔ contribución)	+ + +
corr(IDH sujeto ⇔ demanda)	0
corr(IDH sujeto ⇔ ROI)	−
corr(IDH núcleo ⇔ contribución)	+ + +
corr(IDH núcleo ⇔ demanda)	+
corr(IDH núcleo ⇔ ROI)	+ + +
corr(IDH país ⇔ contribución)	+ + +
corr(IDH país ⇔ demanda)	+
corr(IDH país ⇔ ROI)	+ + +

Algunas diferencias por esperanza de vida podrían ser endógenas. Podría rentar invertir más en personas de países con menos esperanza de vida.

significación $p < 0.01$, tamaño de efecto mínimo $r > 0.1$

REFLEXIONES FINALES

La producción por pares permite generar rápidamente y mantener productos de información útiles y voluminosos.

Los sitios de producción por pares más populares tienden a alinearse con las necesidades de información.

Las diferencias cuantitativas de contenido más populares existen, pero pueden encontrarse en las necesidades de información y no hay indicios de que se originen o acentúen en la producción por pares.

¿Estas diferencias son deseables? ¿A qué debemos aspirar?
Tú decides.

¿SESGOS DEL CONTENIDO POR PARES?

David Abián

esLibre 2023