Aliança per la presència digital del català

Informe sobre la pèrdua de visibilitat dels continguts en llengua catalana als resultats de cerca web

6 de juny 2023

Acció Cultural del País Valencià · Amical Wikimedia · fundació .cat · Institut d'Estudis Catalans · lllll institut ramon llull · Universitat de València · Òmnium Llengua Cultura País · Plataforma per la Llengua · SoftCatalà · WACCAC

Taula de continguts

Introducció

Des de mitjans del 2022, els internautes han observat que els resultats orgànics de les seves cerques web prioritzen els continguts en espanyol, fins i tot quan el mateix contingut o similar està disponible en català i han configurat el seu dispositiu, sistema operatiu, navegador o compte perquè prioritzi els resultats en català. Abans, aquestes preferències lingüístiques es respectaven, però ara s'ignoren.

Aquest problema s'ha plantejat als principals proveïdors de cerca web, però s'han negat a donar cap explicació ni solució eventual, i han demanat que el problema es documenti més enllà de la percepció subjectiva dels usuaris.

Així doncs, a petició del Govern de la Generalitat de Catalunya, l'Aliança per la presència digital del català s'ha encarregat d'aportar la documentació sobre aquesta qüestió i, liderada per Fundació .cat., membre de l'Aliança, ha analitzat les dades de trànsit web de més de 600 webs multilingües per tal de fer un seguiment de l'evolució del trànsit de les seves versions en català al llarg del temps en comparació amb altres idiomes.

Les nostres principals conclusions són que el 66,5 % (més de dos terços) dels llocs web s'han vist (i encara es veuen) afectats pel problema i han perdut trànsit a les seves versions en català. A més, hi ha una forta correlació (una mitjana del 80 %) entre el trànsit català i l'espanyol, la qual cosa significa que la versió en espanyol guanya gairebé una pàgina vista per cadascuna que el català perd. Aquesta correlació és molt més feble (0,25) entre el català i l'anglès.

El problema no afecta per igual tots els llocs web, de manera que aquest informe examina diversos perfils comuns de llocs web, inclosos alguns que no s'han vist afectats de cap manera.

En qualsevol cas, molts dels llocs web afectats es troben entre les organitzacions en català més visitades i rellevants, incloses les administracions, l'àmbit acadèmic, els mitjans de comunicació i els sectors empresarials del domini català que publiquen els seus continguts web en català.

Aquest informe es posarà a disposició dels principals proveïdors de cerca perquè l'utilitzin en els seus esforços per restaurar la visibilitat que han perdut els continguts en català. També estarà disponible per al públic en general i els mitjans de comunicació, així com per a certs diputats del Parlament Europeu que treballen en qüestions relacionades amb les llengües minoritàries de la UE, perquè el puguin fer servir a les seves iniciatives legislatives.

Aquest informe arriba en un moment crític, quan l'aparició dels *chatbots* d'IA està canviant la forma en què els usuaris cerquen i interactuen amb el contingut

digital, i aquests chatbots semblen aprofitar principalment el contingut en els idiomes majoritaris. Per tant, és imprescindible restablir la presència adequada del contingut original en aquests idiomes no tan majoritaris abans que els *chatbots* prenguin el relleu de la cerca web habitual.

Resum Executiu

Impacte en el posicionament del contingut en català als cercadors

Des de mitjans de l'any 2022 s'identifica un fenomen nou que té a veure amb el contingut en català i els cercadors d'Internet. En el cas de **pàgines web multilingües, el contingut en català desapareix de les primeres posicions dels cercadors malgrat la cerca s'hagi realitzat en català i fins i tot si es té configurat l'entorn de navegació per donar preferència al català.**

Abans que això passés, el contingut en català de llocs web multilingües sí que es podia trobar a les primeres posicions de cerca quan el mateix cercador interpretava (ja fos per les configuracions i preferències de l'usuari o per la llengua usada en la pròpia cerca) que aquesta era la llengua en què l'usuari feia la consulta.

No s'ha determinat la data exacta d'inici d'aquest comportament ni tampoc se'n coneix el motiu, tot i les consultes realitzades a les empreses de cerca.

Realitzem un estudi amb dades

Des de l'Aliança per la presència digital del català, i amb l'encàrrec que ens fa la Generalitat de Catalunya, volem poder conèixer millor el fenomen, poder-lo quantificar i observar, saber com impacta en el trànsit dels llocs web, i així poder tenir més i millors arguments per reclamar als actors implicats (les empreses de cerca d'internet, principalment) **que es pugui revertir la situació i es tornin a respectar les preferències dels usuaris a l'hora de decidir en quin idioma volen rebre els continguts de les cerques.** Addicionalment, potser podrem rebatre o confirmar les hipòtesis que tenim sobre què està passant, i per què.

Ens plantegem fer l'estudi mitjançant l'obtenció de dades, i per això **sol·licitem col·laboració a diversos organismes i entitats** que gestionen llocs web multi-idioma (que inclou el català) per demanar-los que ens **facilitin informació sobre el seu trànsit web** que hagin registrat durant els

En quines dades ens hem basat

Hem analitzat el **trànsit web orgànic procedent de cercadors de 639 llocs web** multilingües que inclouen el català i una -o més- llengües.

Tots aquests webs procedeixen d'entitats i organismes dels sectors públic, acadèmic, mediàtic i empresarial catalans que han accedit a col·laborar amb l'estudi de l'Aliança per la presència digital del català, a petició -en nom seu- de la Fundació .cat, que ha realitzat la recol·lecció de dades i s'ha encarregat de l'elaboració de l'informe, amb l'ajut de tots els membres de l'Aliança.

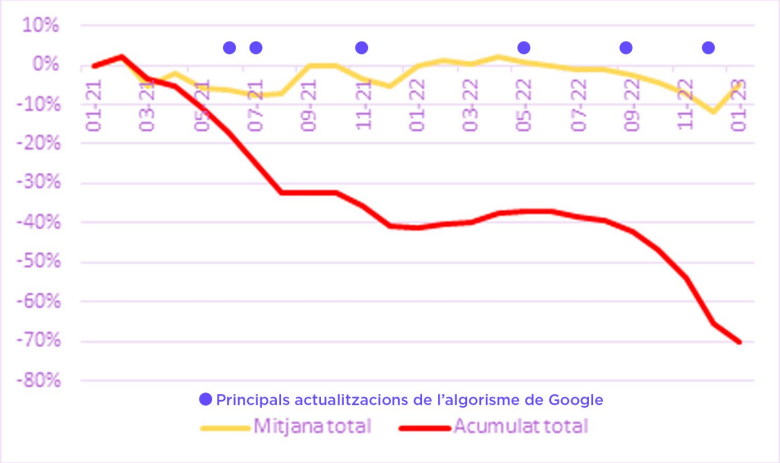
Algunes d'aquestes entitats no fan pública la seva col·laboració i d'altres sí. Agraïm la col·laboració de totes elles.

Principals resultats de l'estudi

Hi està havent pèrdua de visibilitat del contingut en llengua catalana?

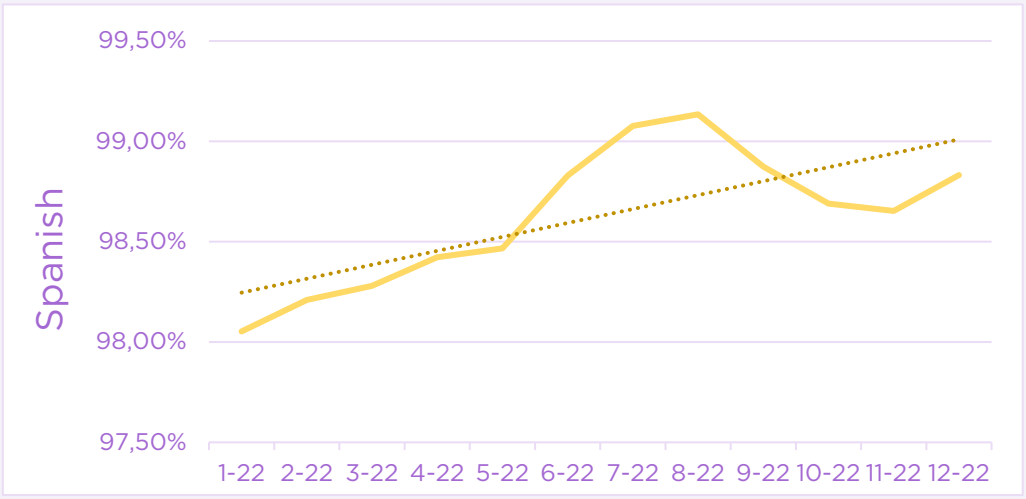
Es constata efectivament que el canvi de tendència del trànsit aportat des del cercador de Google (en el qual es centra l'estudi) es detecta de forma generalitzada durant la primavera del 2022 i persisteix fins a l'actualitat.

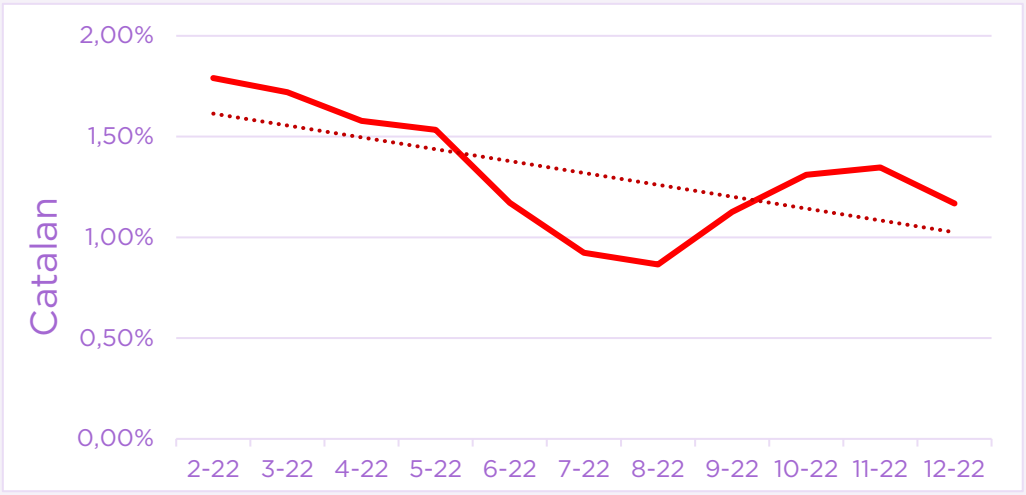
Això s'evidencia en el gràfic següent, que mostra com ha evolucionat la proporció de visites a continguts en català i en castellà per al conjunt de llocs web analitzats durant el període de dos anys considerat en l'estudi.



També ens hem centrat en cadascun dels llocs web afectats, observant l'evolució del trànsit per cada idioma.

Exemple contribuïdor de dades núm. 13

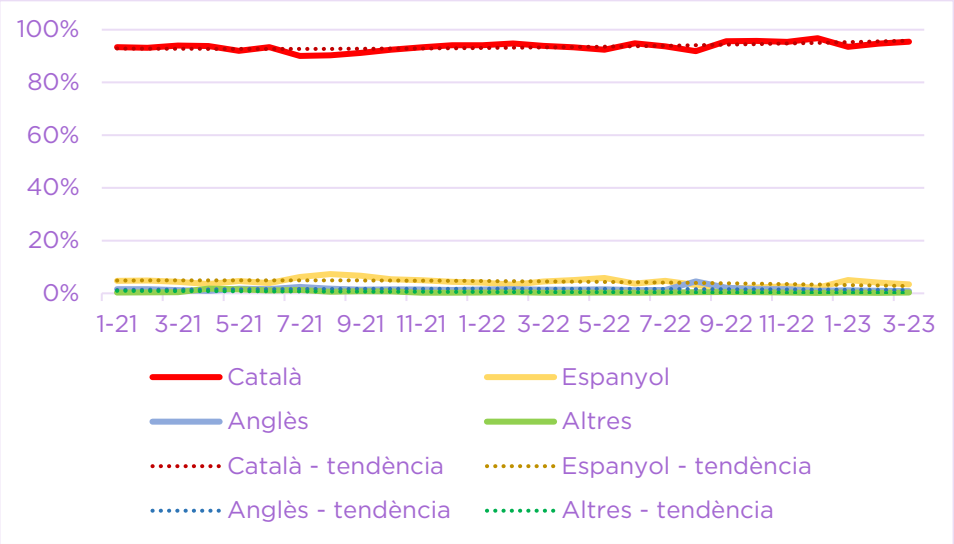




L' afectació és a tots els webs? A quin percentatge de llocs web afecta?

No afecta a tots els llocs web. Segons l'estudi el 66,5% dels llocs web analitzats s'han vist afectats pel problema, havent així perdut tràfic a les seves versions en català.

Exemple contribuïdor de dades núm. 8 (sense afectació)

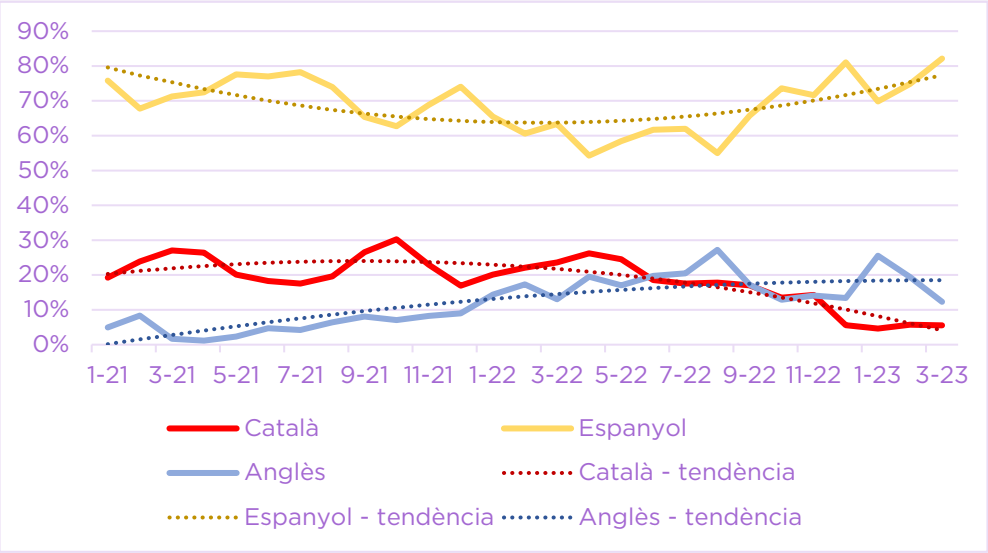


Existeix alguna correlació entre la pèrdua de visibilitat del contingut en català i l'augment de visites a pàgines en altres idiomes, principalment el castellà?

Hi ha una forta correlació d'un 80% entre el tràfic en català i el tràfic en castellà: això vol dir que la versió en castellà guanya gairebé una pàgina vista per cada una que perd la versió en català.

No és que baixin les visites en català únicament, ni les totals, sinó que s'observa com moltes de les visites que abans eren en català ara són visites al contingut en castellà.

Exemple contribuïdor de dades núm. 3



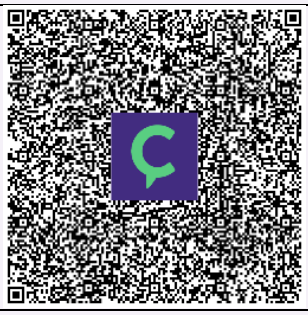
Podem certificar que no es respecten les preferències de l'usuari pel que fa a l'idioma preferent?

Els cercadors han deixat de respectar les preferències lingüístiques explícites dels usuaris. El contingut en català ha perdut visibilitat als resultats de les cerques amb independència de la configuració d'idioma del dispositiu, el navegador i el perfil d'usuari. Per tal de comprovar-ho, hem configurat el nostre propi entorn de proves controlat:

Equip informàtic utilitzat

SO / versió	Windows 10 Pro 2H22 19045.2846		
Navegador / versió	Chrome 113.0.5672.126		
Context de navegació			
IDIOMA ACCEPTAT	ca-ES,ca;q=0.9		
USER-AGENT	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/113.0.0.0 Safari/537.36		
Context de galetes	Sense galetes prèviament acceptades ni generals ni de Google		
Geolocalització			
IP pública	141.166.99.42	Localització de Google	43470 La Selva del Camp

Cerca realitzada: barcelona

Cerca literal	barcelona		
URL de cerca	Enllaç Degut a la llargada de la URL facilitem un enllaç directe.	QR de la URL	

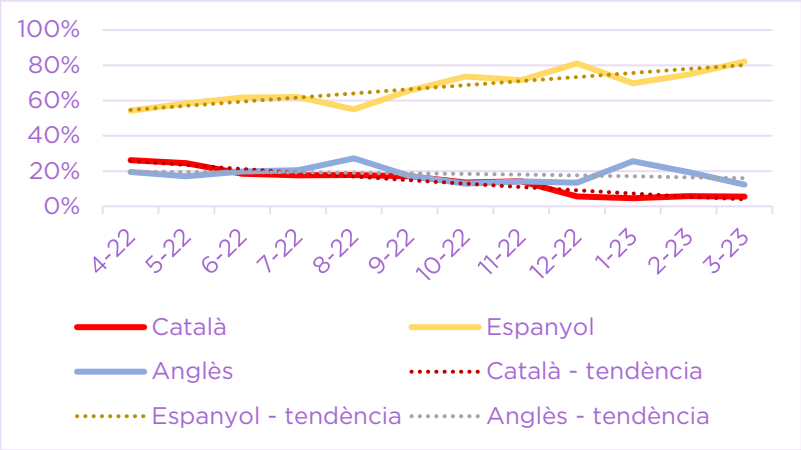
Resultats de la cerca

Captura de pantalla dels primers resultats (no inclou contingut publicitari)			
# primer contingut en català	No n'apareix cap a la primera pàgina de resultats de la cerca	# primer resultat en Espanyol / altra llengua	1

Quan va iniciar-se el problema?

A l'estudi observem que el canvi de tendència del trànsit aportat des del cercador de Google es pot identificar de forma generalitzada **durant la primavera del 2022 i persisteix** fins a l'actualitat.

Exemple contribuïdor de dades núm. 3



- Es constata que l'afectació es manté tant a webs amb .com, .org, .cat o .es, de manera que **l'autoritat del domini no és un punt rellevant en aquest aspecte.**
- **La disponibilitat de contingut no és un problema.** No hi ha hagut canvis en la quantitat ni en la qualitat dels continguts en català.

Podria ésser degut a un etiquetatge incorrecte dels llocs web?

El paràmetre *hreflang* s'ha assenyalat com a una possible causa del problema. Però després de comprovar la utilització de l'*hreflang* que fan tots els llocs web analitzats, hem comprovat que alguns dels que es veuen afectats tenen configurat aquest paràmetre correctament, mentre que alguns dels que no estan afectats no tenen la configuració correcta.

Context

El català és una llengua romànica amb deu milions de parlants en un domini que abasta quatre estats europeus (Andorra, Espanya, França i Itàlia). És la novena llengua més parlada a la Unió Europea, juntament amb el grec, el txec i el portuguès i per davant del suec i el danès. Segons el Baròmetre de les llengües al món[1] del Ministeri de Cultura francès, el català és la 12a llengua més influent a nivell mundial. A més, és la segona llengua a Espanya per nombre de parlants, molt per davant de l'anglès.

El català és molt present a Internet: tot i que és la 75a llengua del món per nombre de parlants, se situa constantment entre la 10a i la 20a posició pel que fa a la presència a la xarxa. La satisfacció produïda per la presència de la llengua catalana a Internet (almenys a la xarxa) ha donat pas, en els darrers mesos, a una frustració generalitzada entre els internautes catalanoparlants per la marginació dels continguts digitals en català als resultats de la cerca web: quan una pàgina està disponible en català i en espanyol, la versió en català apareix per sota de la versió en espanyol als resultats de la cerca web, és a dir, subordinada, i **això passa independentment de les preferències de l'usuari**. El fenomen doncs afavoreix així els clics a les versions dels llocs web en espanyol i afecta també els altres motors de cerca. És més visible a Google perquè és el cercador més utilitzat, però també passa a tota la resta. Tot indica que l'origen del problema, encara no diagnosticat, podria estar en un altre lloc.

La visibilitat digital de la llengua catalana a la xarxa es veu amenaçada per aquest fenomen. Continuen publicant-se continguts en català i s'hi pot accedir directament com sempre, però perden visibilitat gradualment perquè bona part dels internautes hi arribaven a través de cerques web, i per tant, una incidència tècnica està provocant que tant Google com altres cercadors (Bing, DuckDuckGo, Qwant...) prioritzin als seus resultats la versió en espanyol de les webs que també en tenen una en català.

Per exemple, quan se cerca «Merce Rodoreda» sense l'accent, el primer resultat és un article sobre l'escriptora a la Viquipèdia, però tant Google com Bing porten a l'edició en espanyol i mostren l'article de la Viquipèdia en català en segon lloc. Altres cerques, incloses les de llocs web corporatius i organismes oficials, es comporten de la mateixa manera. Això també passa encara que l'usuari hagi indicat explícitament que prefereix veure primer els resultats en català, a través de la configuració del seu dispositiu, navegador i compte personal (de Google i de Microsoft, respectivament).

[1] https://www.culture.gouv.fr/en/Thematic/French-and-French-languages/Acting-for-languages/Innovation-in-language-and-digital/Supporting-and-encouraging-linguistic-diversity-in-the-digital-domain/2022-World-Language-Barometer

No obstant això, la situació actual és encara més sorprenent pel que fa a la marginació del català, i això fa encara més incomprensible aquest problema. Tant Google com Bing mostren l'anomenat *snippet*, el quadre de resum de dades que apareix destacat a la dreta (als ordinadors) o a la part superior (als dispositius mòbils) quan es cerquen persones, empreses i topònims, entre d'altres, en el nostre idioma si hem configurat així el navegador. En canvi, si la cerca es realitza amb l'ortografia correcta («Mercè» en lloc de «Merce»), el primer resultat orgànic de la llista és l'article de la Viquipèdia en català. Tanmateix, hi ha cerques de termes en català (xucrut) que retornen pàgines en espanyol com a primer resultat que no contenen el terme real que hem introduït, sinó la traducció. Pel que fa a l'alimentació, algunes recerques de «pollastre» proposen la definició de la paraula... al Diccionari de la Real Academia Española!

Així doncs, el problema sembla complex i depèn d'una gran varietat de combinacions, però la situació general és que, avui dia, les pàgines en català tenen menys visibilitat que abans i d'alguna manera, es prioritzen els equivalents en espanyol per sobre del català, la qual cosa implica una menor visibilitat i menys clics. Això suposa un **empitjorament respecte** del passat recent, quan els catalanoparlants tècnicament declarats rebien, en primer lloc, els resultats d'enllaços en català sempre que n'hi haguessin de disponibles. Però, sobretot, es tracta d'un problema de futur, ja que, a mida que cliquem els resultats de la cerca, entrenem l'algoritme del cercador indicant quina pàgina de les que proposa ens ha interessat més. Si no és la de català, aquesta versió del contingut cada cop quedarà més enterrat.

La disponibilitat de contingut no és el problema

Cal destacar que no es tracta d'un problema d'abastiment. No hi ha hagut canvis en la quantitat ni la qualitat dels continguts en català. La satisfacció esmentada abans sobre la presència del català a la xarxa està justificada. Des de fa dues dècades, els voluntaris de l'associació WICCAC (Webmàsters Independents en Català, de Cultura i d'Àmbits Cívics) elaboren un baròmetre mensual[2] que detalla el percentatge d'ús del català als llocs web de centenars d'empreses, entitats i institucions amb seu o activitat al nostre àmbit lingüístic. La xifra, que ha crescut gradualment del 41 % l'agost del 2002 al 66 % el desembre del 2022, varia en funció del sector d'activitat i té un valor relatiu perquè no es pondera a partir del trànsit de cada web: la pàgina d'una agència immobiliària local compta igual que la d'un diari digital generalista. No obstant això, el baròmetre ofereix una imatge molt completa de la situació i fa que sigui més fàcil identificar on cal concentrar els esforços de millora.

De la mateixa manera, un estudi recent[3] de Softcatalà (una altra associació de voluntaris) demostra que entre el mig milió de llocs webs més populars

2 http://wiccac.cat/webscat.html
3 https://github.com/jordimas/crux-top-lists-catalan

d'Internet, n'hi ha prop de 470 que tenen versió en català. Sis d'ells es troben entre els 1000 primers, set es troben entre els 5000 primers i dotze entre els 10.000 primers. Alguns són previsibles, com Booking, Google, Facebook, Outlook i Twitter, però també hi ha webs menys coneguts però amb moltes visites que tenen versió en català, com la pàgina d'escacs Chess.com i la bíblica Bible.com. Cal tenir en compte que aquesta anàlisi es va basar en una font de dades inesperada: la llista d'adreces que els usuaris de Google Chrome visiten i guarden a la memòria cau del navegador, que Google acumula i agrega mensualment per a ús públic.

Aquest estudi de Softcatalà conclou que el català té una presència digital especialment forta al sector públic i l'acadèmic. També demostra que l'existència del domini de primer nivell «.cat» és un signe d'identitat: en quinze anys de treball, la Fundació .cat l'ha convertit en la segona extensió més utilitzada per a continguts en català (141 dels 470 llocs web) només per darrere del «.com» (178 webs) globalment i superant-lo en nombre de webs dissenyades a Catalunya; en tot cas, molt per sobre del «.es» (40 webs).

Un problema greu que requereix l'atenció adient

Cal tenir en compte que el contingut en català ha anat perdent preeminència en els resultats de la cerca web respecte del contingut en espanyol, que molts webs multilingües han detectat un canvi sobtat en les visites a les seves versions en català, que aquest canvi s'ha produït sense cap modificació significativa a l'arquitectura, el contingut o la qualitat dels llocs, i que molts llocs web han observat que el trànsit que abans es dirigia al català ara passa a l'espanyol. Ateses aquestes circumstàncies, creiem que la llengua catalana s'enfronta a un greu problema de visibilitat a la xarxa i que cal que els proveïdors de cerca web prenguin mesures ràpides per restaurar la situació anterior.

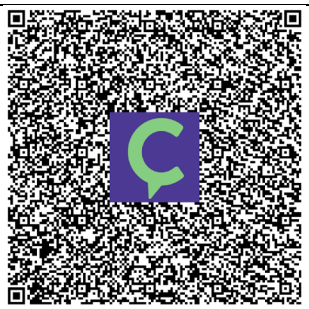
També cal tornar a subratllar que els proveïdors de motors de cerca han començat a ignorar les preferències d'idioma específiques que cada usuari estableix al seu dispositiu, sistema operatiu, navegador o compte d'usuari. Independentment dels nombrosos elements que influeixen als rànquings dels motors de cerca, exigim que els motors de cerca tornin a respectar les preferències dels usuaris com ho feien abans del maig del 2022.

Exemples

Més enllà de la situació que ja han fet palesa diferents usuaris a les xarxes socials, cal demostrar de manera neutral el biaix dels resultats de cerca. Per això, establim un entorn de proves basat en el sistema operatiu Windows on escollim el català com a principal preferència per a les interfícies i la navegació. Les especificacions d'aquest entorn de prova es defineixen a la taula següent.

SO/versió	Windows 10 Pro 2H22 19045.2846		
Navegador/versió	Chrome 113.0.5672.126		
Context del navegador			
ACCEPT-LANGUAGE	ca-ES,ca;q=0.9		
USER-AGENT	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, com Gecko) Chrome/113.0.0.0 Safari/537.36		
Context de les galetes	No hi ha galetes generals ni de Google establertes prèviament		
Context de geolocalització			
IP pública	141.166.99.42	Ubicació de Google	43470 La Selva del Camp

Un cop llest, aquest entorn es va utilitzar per executar diversos conjunts de proves i per avaluar-ne els resultats. A continuació, es descriuen tres de les situacions més rellevants:

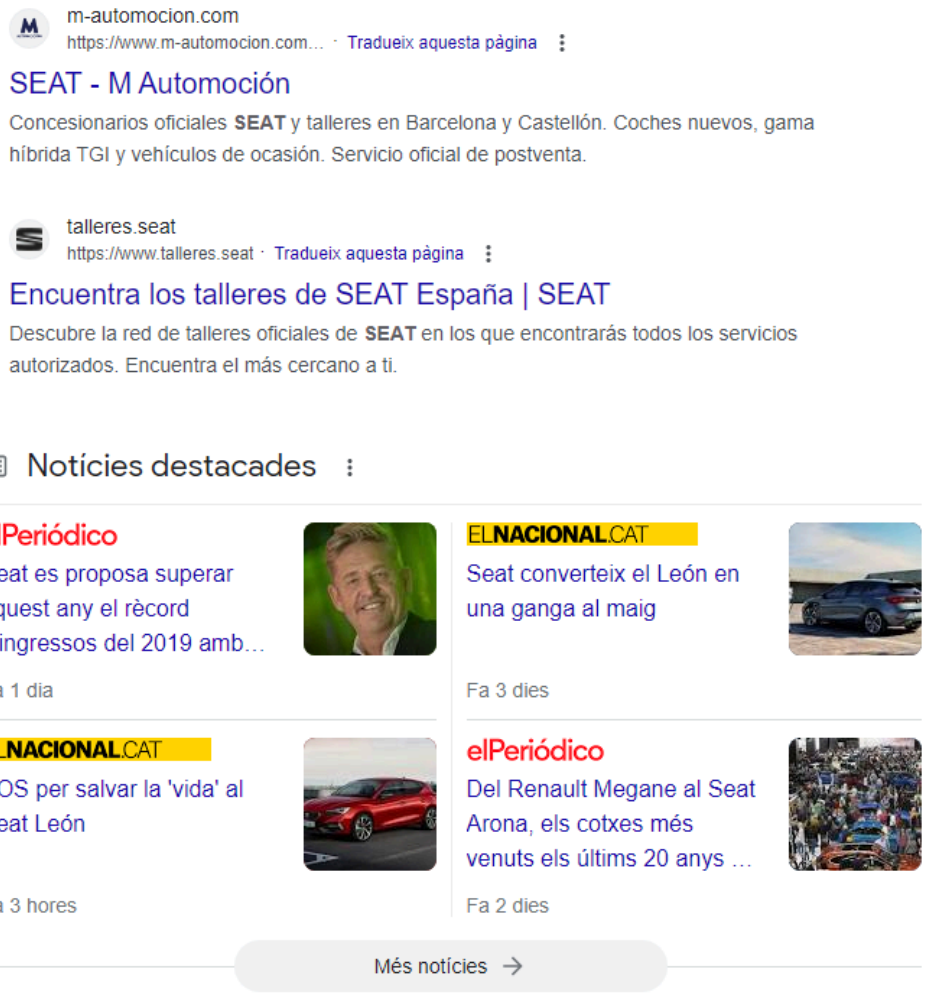
Consulta literal	barcelona		
URL de consulta	Enllaç A causa de la longitud de l'URL, és més fàcil i pràctic utilitzar un enllaç directe.	QR de l'URL	

Captura de pantalla dels resultats (a dalt, contingut no promocionat)			
Núm. de primers resultats en català	Cap a la primera pàgina de resultats.	Núm. de primers resultats en espanyol/un altre idioma	1

Consulta literal	sagrada familia		
URL de consulta	Enllaç A causa de la longitud de l'URL, és més fàcil i pràctic utilitzar un enllaç directe.	QR de l'URL	
Captura de pantalla dels resultats (a dalt, contingut no promocionat)			
Núm. de primers resultats en català	2 (subnivell)	Núm. de primers resultats en espanyol/un altre idioma	1

L'última situació és una mica diferent a les anteriors: en aquest cas, la consulta força els resultats en català. Aquest paràmetre es va definir mitjançant l'opció

que hi ha disponible a la interfície de cerca a Google que permet mostrar resultats amb un idioma específic.

<table>
<tr><td>Consulta literal</td><td colspan="3">seat</td></tr>
<tr><td>URL de consulta</td><td>Enllaç
A causa de la longitud de l'URL, és més fàcil i pràctic utilitzar un enllaç directe.</td><td>QR de l'URL</td><td></td></tr>
<tr><td>Captura de pantalla dels resultats (a dalt, contingut no promocionat + notícies destacades)</td><td colspan="3"> 
</td></tr>
<tr><td>Núm. de primers resultats en català</td><td>10 (resultats)/1 (notícies)</td><td>Núm. de primers resultats en espanyol/un altre idioma</td><td>1 (resultats)/no trobat (notícies)</td></tr>
</table>

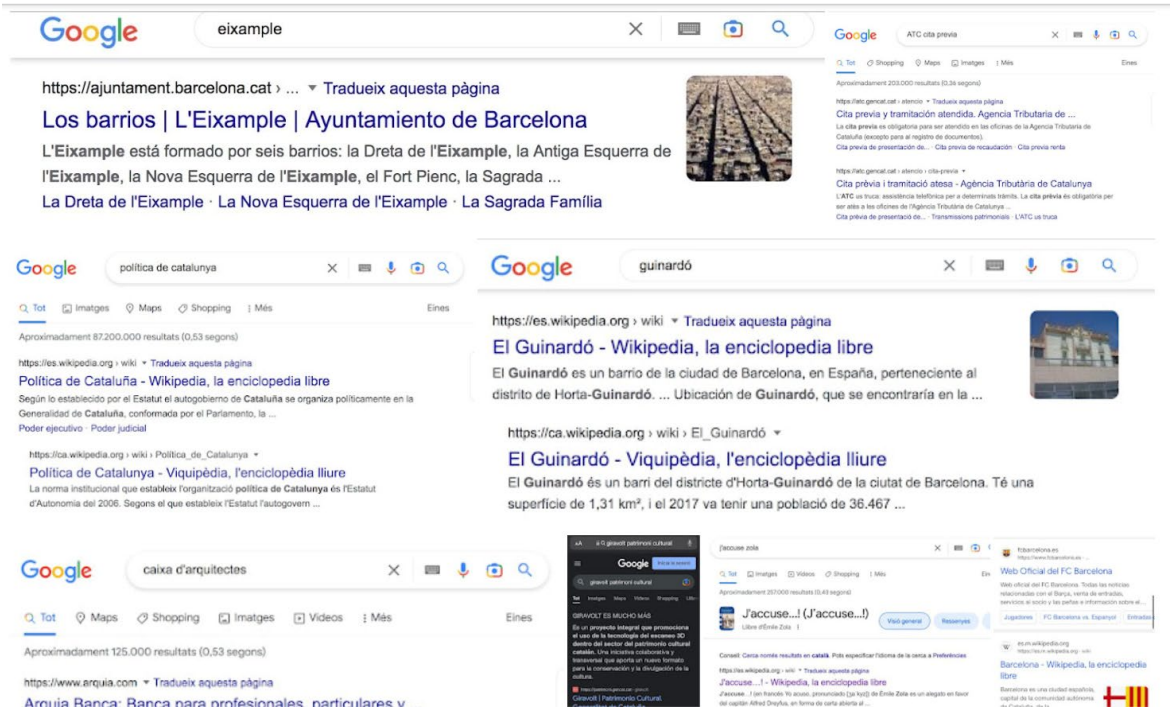
Aquesta primera etapa de l'estudi va concloure tal com mostren aquests exemples: Google ignora la preferència lingüística de l'usuari, almenys pel que fa a la llengua catalana. Els resultats de les cerques convencionals prioritzen els continguts en llengua castellana, encara que s'especifiqui el català com a idioma preferit per als resultats.

Les proves també mostren que Google coneix bé la diferència entre el català i l'espanyol, ja que les notícies destacades es mostraven exclusivament en l'idioma preferit quan es va definir específicament.

Exemples aportats pels usuaris

Tan bon punt hi va haver una percepció pública del problema de la visibilitat dels continguts en català, nombrosos usuaris van penjar captures de pantalla dels seus resultats de cerca a les xarxes socials. N'hem recollit alguns, sense cap ordre particular, en aquest àlbum de Google Fotos:

https://photos.app.goo.gl/mSqZtXiBFd6fCCR27



Objectius d'aquest informe

Les principals organitzacions cíviques de promoció i defensa de la llengua catalana han constituït[4] l'Aliança per la presència digital del català (APDC). Acció Cultural del País Valencià, Amical Wikimedia, la Fundació .cat, l'Institut d'Estudis Catalans, l'Institut Ramon Llull, Obra Cultural Balear, Òmnium Cultural, la Plataforma per la Llengua, Softcatalà i WICCAC han posat en comú els seus recursos tècnics, coneixements i capacitat de mobilització per abordar el problema dels resultats de la cerca web.

Mitjançant contactes informals amb les principals empreses proveïdores de cerca web, les mateixes empreses han reconegut aquest problema i han sol·licitat dades sobre el seu abast per tal d'intentar resoldre'l. Membres de l'APDC, liderats per la Fundació .cat, han elaborat aquest informe que rastreja i detalla la pèrdua de visibilitat dels continguts en llengua catalana als resultats de la cerca web al llarg del temps, de manera que les empreses implicades puguin diagnosticar millor el problema a partir de fets en comptes de percepcions.

Hem reunit desenes d'organitzacions del sector públic, acadèmic, mediàtic i empresarial del domini de la llengua catalana que publiquen el contingut web en català perquè ens subministrin dades històriques de trànsit de més de 600 llocs web (dominis principals i subdominis) que provenen de resultats de cerca orgànics (exclosos, per tant, els resultats patrocinats). Després, hem recopilat i analitzat aquestes dades per poder quantificar la creixent degradació del contingut en català als resultats de trànsit web, així com poder identificar i revertir qualsevol element relacionat amb els canvis tècnics i d'infraestructura.

Aquest informe es posarà a disposició de la ciutadania i dels mitjans de comunicació a través del web de l'APDC (https://aliançadigital.cat). També es facilitarà als principals proveïdors de cerca, ja sigui directament o a través de la Generalitat de Catalunya, que ha manifestat interès per ajudar a resoldre el problema i, fins i tot, s'ha posat en contacte amb una de les empreses. Aquest informe es lliurarà a certs diputats del Parlament Europeu que treballen en qüestions relacionades amb les llengües minoritàries de la UE, perquè el puguin fer servir a les seves iniciatives legislatives.

Creiem que aquest informe arriba en un moment crític, ja que l'aparició de *chatbots* d'IA (com ara ChatGPT, Bing Chat i Google Bard) està a punt de canviar la manera en què els usuaris cerquen i interactuen amb el contingut digital; aquests *chatbots* s'aprofitarien principalment del contingut en els idiomes majoritaris (tal como ha reconegut OpenAI), encara que siguin perfectament capaços de dialogar amb l'usuari en molts idiomes, inclòs el català. Per tant, és imprescindible restablir la presència adequada del contingut original

[4] https://aliançadigital.cat/wp-content/uploads/2023/03/NdP_Alianca-per-la-presencia-digital-del-catala_.pdf

en aquests idiomes no tan majoritaris abans que els *chatbots* prenguin el relleu de la cerca web habitual.

Metodologia

Per tal de diagnosticar el problema dels resultats en català a la cerca web i ajudar a resoldre'l, hem aplicat un enfocament de dos nivells. D'una banda, hem intentat avaluar el grau de descens del posicionament del català respecte de la situació anterior. D'altra banda, hem demanat a diversos experts la seva opinió sobre les possibles causes d'aquest descens, de manera que els socis implicats no hagin d'explorar pel seu compte aquestes possibilitats partint de zero.

L'annex 1 conté una còpia de les especificacions tècniques que vam facilitar a tots els col·laboradors a qui vam contactar i que podien estar disposats a subministrar les seves dades de trànsit web als efectes d'aquest informe.

Caracterització general dels col·laboradors

Aquest informe exposa l'anàlisi de les dades de trànsit web proporcionades per tretze col·laboradors, que cobreixen un conjunt agregat de 639 dominis i subdominis, sobre els quals hem rebut dades individuals o agregades.

Aquests col·laboradors s'han seleccionat no només en funció dels criteris tècnics requerits (multilingüisme del lloc, comparació, etc.), sinó també de la rellevància. Es tracta de llocs web importants pel que fa al nombre de visites i alguns estan gestionats per institucions públiques de primer nivell.

Aquests col·laboradors s'han classificat en tres grups:

- **Grup 1.** Llocs web afectats pel canvi de comportament de l'algoritme de classificació de Google. Aquest grup inclou un conjunt de vuit col·laboradors, que abasten 425 dominis/subdominis.
- **Grup 2.** Llocs web no afectats pel canvi anterior. Els webs d'aquest grup no s'han vist afectats pel canvi en l'algoritme de classificació, de manera que mantenen el seu comportament i les tendències anteriors sense cap canvi notable. Aquest grup inclou tres col·laboradors, que abasten 212 dominis/subdominis.
- **Grup 3.** Aquest grup conté llocs web les dades dels quals no es poden analitzar, a causa del seu comportament o la seva estructura tècnica. Està format per dos col·laboradors, que representen dos dominis/subdominis.

Totes les dades proporcionades pels col·laboradors s'han normalitzat a un format de pes relatiu per idioma i període de temps (mensual), perquè aquest és el format que prioritzen els col·laboradors que trien l'estructura més restrictiva.

A més, els llocs web dels col·laboradors utilitzen diferents dominis de primer nivell, de manera que els llocs web analitzats poden acabar en «.com», «.cat», «.org» i «.es», entre d'altres.

Pel que fa als cercadors web

Les dades següents fan referència a les visites al web proporcionades específicament pels resultats orgànics (no de pagament) del motor de cerca web de Google als llocs web dels col·laboradors. Hem centrat la nostra recerca en aquest cas perquè és el que ha provocat la problemàtica actual per un canvi en el seu comportament habitual. Així mateix, Google ha estat (i encara és) el cercador web dominant al mercat espanyol i mai no ha tingut una quota de mercat inferior al 95 % durant el període considerat[5].

[5]Estadístiques globals d'StatCounter: https://gs.statcounter.com/search-engine-market-share/all/spain

També hem confirmat amb cada col·laborador que no s'han fet canvis importants en l'arquitectura i/o el contingut del lloc web que hagin pogut afectar directament la indexació i el posicionament del lloc als motors de cerca durant el període considerat. La majoria d'ells han estat realitzant un treball actiu en aquests àmbits, però sense grans canvis d'estratègia ni de recursos.

Podríem repetir aquest mateix procés amb altres cercadors en els propers mesos.

Resultats acumulats

Com veurem més endavant, el trànsit orgànic procedent de les cerques a Google de cada lloc web que ha participat en aquest estudi s'ha analitzat separadament.

Tot i això, per tenir una visió global de l'evolució de les visites que arriben al contingut web en català o castellà procedent de cerques de Google, també hem calculat la proporció entre les que arriben a cada idioma, mes per mes i per a cada lloc web participant. Després hem normalitzat aquesta ràtio basant-nos en el valor inicial de la sèrie temporal, que vam establir al gener de 2021. Finalment, hem fet la mitjana de les ràtios dels diferents llocs web (línia groga) i hem acumulat aquests canvis al llarg del temps (línia vermella).

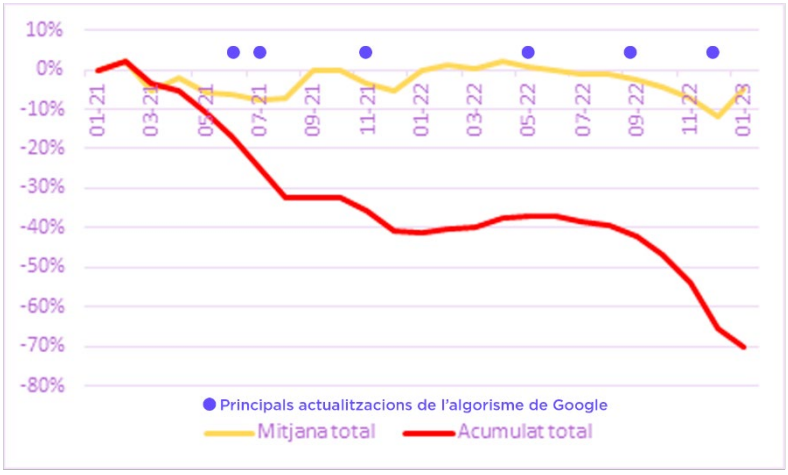
Ho hem fet únicament amb els llocs web afectats per la pèrdua de trànsit a les seves versions en llengua catalana, que és la situació que ens interessa analitzar en el context d'aquest estudi.

Atès que no tots els llocs web participants han proporcionat el seu trànsit absolut de dades, no hem pogut ponderar els resultats en funció del seu volum. Degut a això, hem diferenciat dues situacions. La primera considera que cada lloc web participant té el mateix pes. La segona considera només els llocs web que hem comprovat que tenen més de 500.000 sessions al llarg de la sèrie temporal sencera. En ambdós gràfics hem indicat els moments en què, segons les informacions que ja ha revelat[6] Google, s'han produït canvis importants en l'algorisme del seu motor de cerca.

Pèrdua de trànsit del contingut en català en favor del contingut en espanyol pel conjunt de llocs web analitzats

El gràfic següent mostra com la ràtio de visites al contingut en català -i en castellà- ha evolucionat amb el temps pel conjunt de llocs web d'aquest estudi.

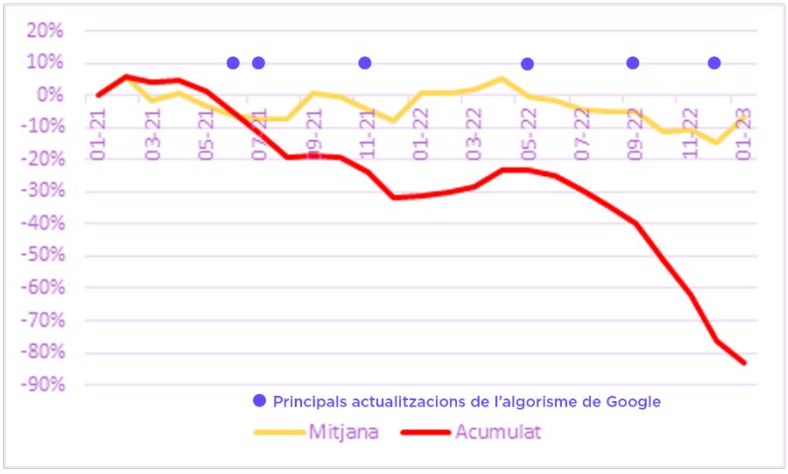
[6] https://status.search.google.com/products/rGHU1u87FJnkP6W2GwMi/history



Després d'un breu període inicial, que aparentment mostra un creixement del trànsit cap a la versió catalana dels llocs web afectats, la situació es converteix de sobte en una tendència a la baixa, que durant 12 mesos es manté majoritàriament estable al voltant del 40%. A partir d'aquí, entre finals de l'estiu del 2022 i principis de la tardor del 2022, torna a caure amb força fins a arribar a una pèrdua del 70% a principis del 2023.

Pèrdua de trànsit del contingut en català en favor del contingut en espanyol en els llocs web més visitats

D'altra banda, quan del nostre conjunt de dades descartem els llocs web menys visitats, i considerem només els que tenen més trànsit (més de 40 milions de sessions en un dels casos), la ràtio entre el contingut en català i en espanyol evoluciona de la manera que mostra el gràfic següent:



La pèrdua de visites procedents de cerques de Google pel contingut en català en el cas dels llocs web més visitats és encara més notable que a la vista anterior (pel conjunt sencer de llocs web analitzats): el creixement inicial dels continguts en llengua catalana roman uns mesos més, però a partir del maig del 2021 hi ha una caiguda del voltant del 20% de pèrdua de visites a continguts en català.

Aquesta situació es manté estable, amb petites variacions, fins a abril – maig de 2022, quan comença una caiguda sostinguda fins i tot superant el 80% de les visites perdudes de manera acumulada.

Fins i tot deixant de banda la primera part de la sèrie temporal i centrant-nos en l'última meitat l'any 2022, les pèrdues mes a mes es mantenen en ambdues vistes, la qual cosa provoca una pèrdua crítica de trànsit de continguts en llengua catalana.

Resultats específics de llocs web

Comportament lingüístic dels llocs web analitzats

Tots i cadascun dels llocs web coberts per aquesta anàlisi són multilingües, és a dir, **tots ofereixen contingut en més d'un idioma** als seus visitants. A la pràctica, això vol dir que la interfície d'usuari d'aquests llocs i, almenys, una part del seu contingut, s'ofereix en català i en espanyol; alguns poden incloure també altres llengües com l'occità, l'anglès o el francès, entre d'altres.

Molts llocs web multilingües no són simètrics, és a dir, no ofereixen el mateix contingut exacte en tots els idiomes que tenen inclosos a la seva interfície; per això, en cas de dubte, hem escollit aquells que mostren una correspondència més ajustada del contingut.

Ens hem centrat en el comportament relatiu dels continguts en català i en espanyol, i hem observat que **el conjunt de casos mostren una correlació negativa**, més o menys forta, entre aquestes dues llengües:

	Correlació CA – ES	Correlació CA - EN *
Col·laborador 1	-0,98	N. a.
Col·laborador 2	No valorable	No valorable
Col·laborador 3	-0,52	-0,41
Col·laborador 4	-0,41	-0,02
Col·laborador 5	-0,99	-0,67
Col·laborador 6	-0,87	-0,17
Col·laborador 7	No valorable	No valorable
Col·laborador 8	-0,85	-0,42
Col·laborador 9	-0,96	-0,23
Col·laborador 10	-0,71	-0,46
Col·laborador 11	-0,97	-0,24
Col·laborador 12	-0,55	0,29
Col·laborador 13	-1,00	N. a.
Mitjana	**-0,80**	**-0,25**
Màx.	-1,00	-0,67
Mín.	-0,41	0,29

* No tots els llocs web dels col·laboradors estan disponibles en anglès.

Aquesta forta correlació negativa fa que un augment de les visites al lloc web en espanyol impliqui una pèrdua de visites al lloc web en català. Això valida la percepció que ja han expressat diversos administradors web. Per tant, no només estan disminuint les visites al contingut en català, sinó que moltes de les visites que abans anaven al català passen ara al contingut en espanyol.

A més, aquesta ràtio tendeix a -0,80, la qual cosa és una correlació molt forta: per cada visita a la versió en espanyol, la versió catalana perd gairebé una visita.

D'altra banda, si es compara amb la correlació entre el català i l'anglès, la ràtio mitjana és només de -0,25 i fins i tot esdevé neutra o positiva en alguns casos (col·laboradors 4 i 12, respectivament). Aquesta ràtio mitjana fa que només es perdi una visita a la versió catalana del lloc per cada quatre visites a la versió en anglès.

Significat en el context d'aquest informe

A nivell pràctic, aquests resultats mostren que el trànsit de cerca procedent de Google funciona amb una relació gairebé exclusiva entre el català i l'espanyol: cada visita a la secció en espanyol del lloc web equival a una visita perduda a la secció en català; per tant, una pèrdua de trànsit al contingut en català suposa un creixement del trànsit al contingut en espanyol.

Per tant, el fet que Google prioritzi els resultats de cerca dels continguts en espanyol sobre el català, sovint fins i tot quan les paraules clau de cerca són termes comuns catalans, ha provocat que el lloc web darrere dels enllaços hagi experimentat un canvi en el perfil de les seves visites, fet que, en la majoria dels casos, també ha suposat una variació de tendència d'ús de les llengües que ofereixen.

Així, hem confirmat que els canvis que ha realitzat Google als resultats de la cerca han tingut un impacte important en molts dels llocs web analitzats.

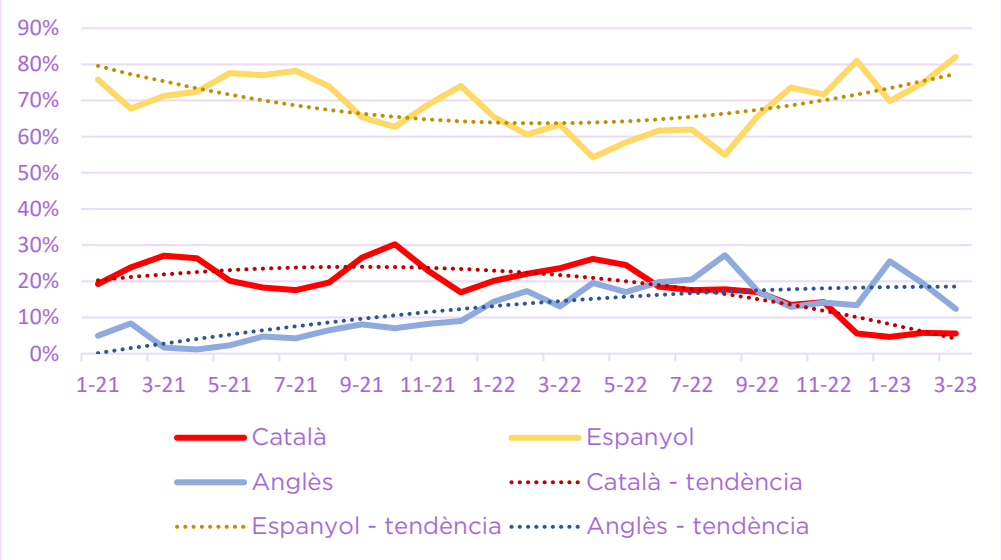
Cal destacar que **el 66,5 % dels llocs web analitzats s'han vist afectats**, per la qual cosa l'efecte és molt rellevant atès el volum de webs afectats. Tanmateix, ens sorprèn que aquest problema no impacti en tots els llocs web.

Finalment, els gràfics següents mostren una clara coincidència temporal dels incidents, en forma d'interrupció/ruptura de tendències. Això podria indicar que el problema es pot explicar per algun canvi sobtat en la indexació de Google. No totes les tendències es trencarien alhora si fossin causades per canvis interns als llocs web.

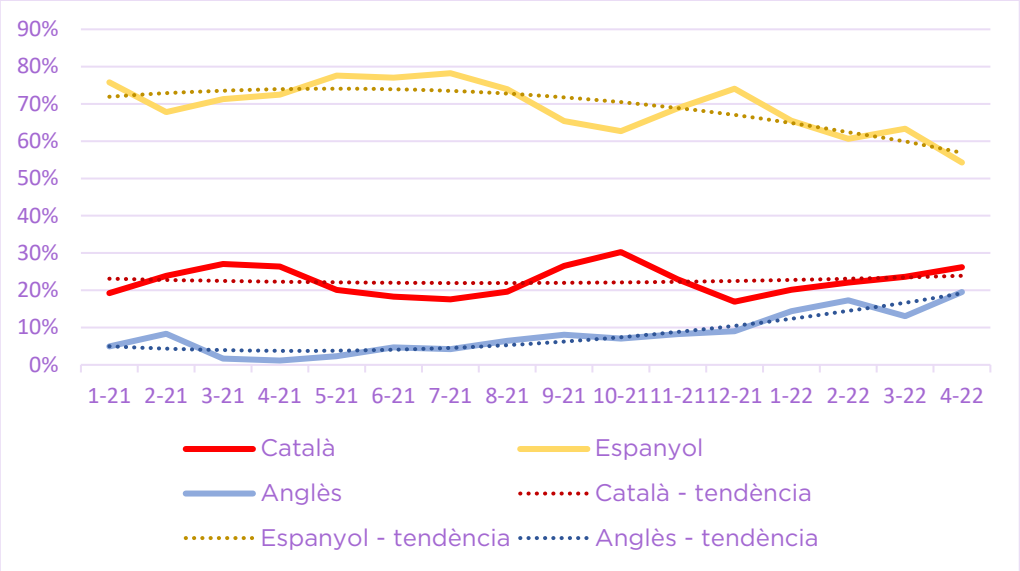
Grup 1. Llocs web on s'ha detectat un impacte

Per tal d'il·lustrar aquesta qüestió, a continuació repassem individualment quatre de les situacions analitzades.

Col·laborador 3

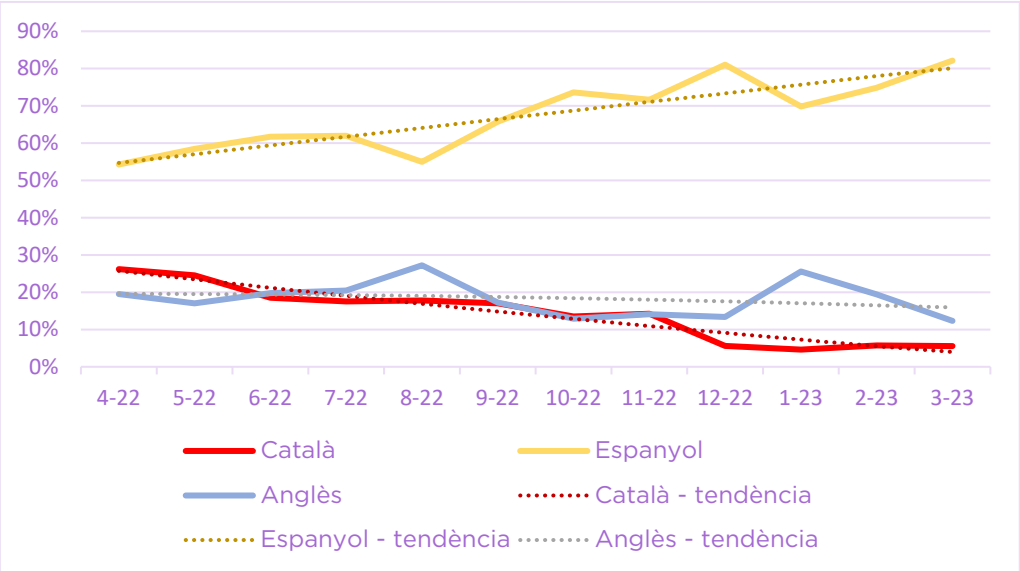
<table>
<tr><td colspan="4">Targeta de col·laborador</td></tr>
<tr><td>Descripció</td><td colspan="3">Aquest primer cas és una organització amb un doble enfocament, local i internacional, que es comunica activament en diversos idiomes com a element indispensable de la seva activitat.</td></tr>
<tr><td>Correlació CA – ES</td><td>-0,52</td><td>Correlació CA - EN</td><td>-0,41</td></tr>
<tr><td>Té «hreflang»?</td><td colspan="3">Sí, sense errors. Possible millora: afegir x-default.</td></tr>
<tr><td>Sèrie completa</td><td colspan="3"> 
</td></tr>
</table>

El primer gràfic mostra la situació *abans* que es detectés l'anomalia del trànsit procedent de Google:



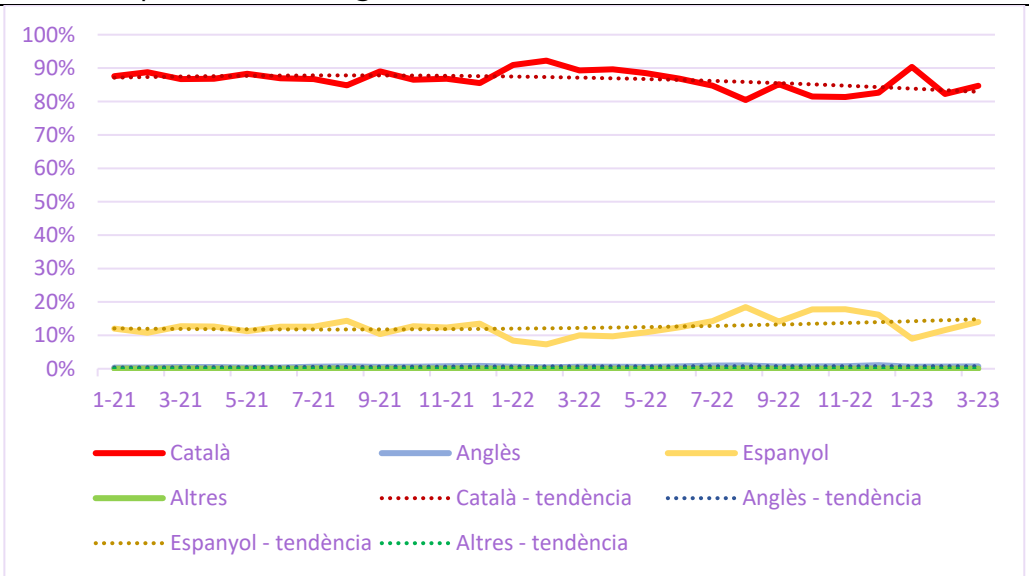
Es pot observar una clara tendència de setze mesos en què l'anglès creix clarament mentre que el català es manté estable, amb breus variacions al voltant del 10 %. Mentrestant, l'espanyol mostrava una tendència evident a la baixa, que començava amb el 75 % de les visites i arribava al 55 % al final del període.

Durant aquest període, també s'observa el punt màxim de visites a la versió en català l'octubre del 2021, així com la diferència mínima entre la versió en català i en espanyol: 28 % l'abril del 2022.

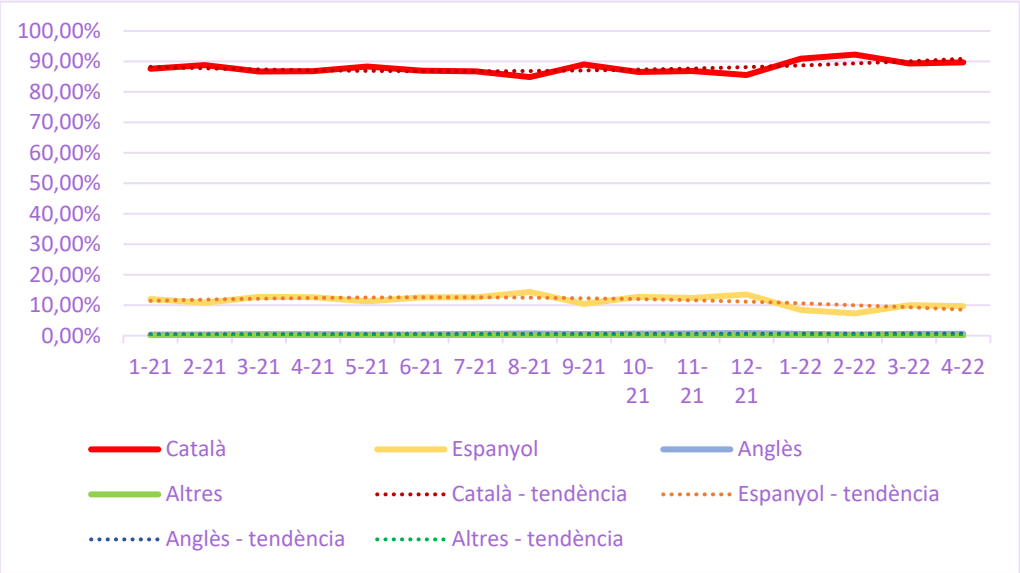


A partir d'aleshores, la sèrie continua des d'abril del 2022 i podem veure un marcat canvi de tendència: l'espanyol comença a créixer ràpidament fins arribar al cim, per sobre del 80 % al final de la sèrie. Durant el mateix període, el català deixa de ser estable i arriba al seu mínim, lleugerament per sota del 5 % del trànsit, el gener del 2023. Aquest és també el moment de la diferència més gran amb l'espanyol: una distància de més del 75 % el desembre del 2022. Finalment, l'anglès també experimenta un canvi de tendència, passant d'un creixement marcat a una disminució lenta.

Col·laborador 5

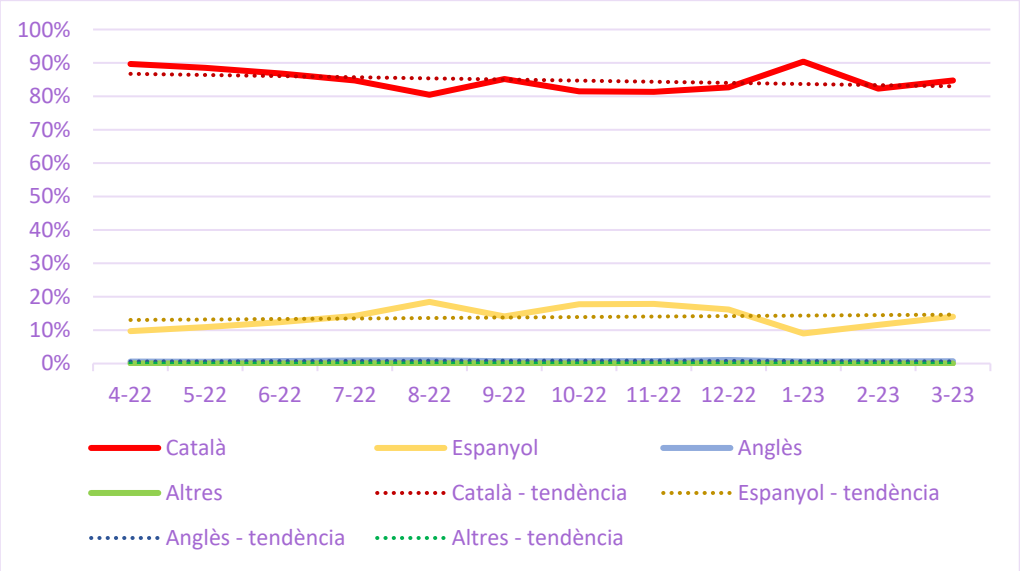
Targeta de col·laborador			
Descripció	El segon cas que hem analitzat és un lloc web específic d'una administració pública centrat principalment en el públic de parla catalana. No obstant això, també ofereix continguts en espanyol i en anglès, així com un cas regional concret.		
Correlació CA – ES	-0,99	Correlació CA - EN	-0,67
Té «hreflang»?	Sí, però la referència entre llengües no està ben configurada. Millores possibles: afegir la referència i x-default.		
Sèrie completa			

El gràfic següent mostra el 2021 complet i la primera part del 2022, fins al moment en què es detecta un canvi de tendència:



Com era d'esperar d'aquest lloc web en concret, el català té un volum de visites molt superior al de la resta d'idiomes, de manera que destaca com la llengua principal per una gran diferència. Durant aquest període, el català mostra un creixement lent però sostingut i arriba al que és, probablement, el seu llindar màxim, mentre que la resta de llengües es mantenen estables o disminueixen lentament, en el cas de l'espanyol.

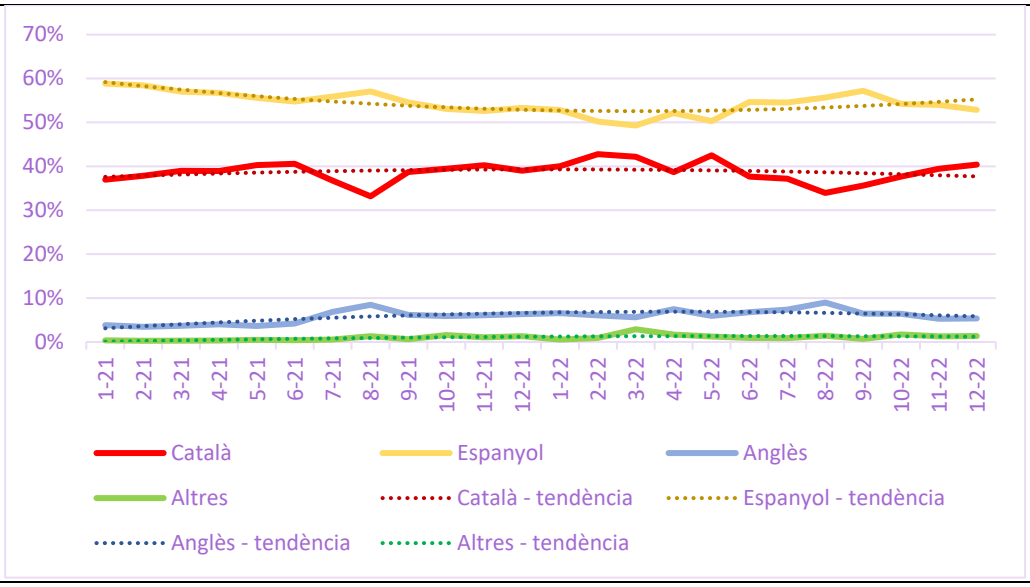
El màxim de visites al contingut en català procedents dels cercadors es produeix el febrer del 2022, fins superar el 92 % del trànsit total i registrar la diferència màxima entre el català i l'espanyol: gairebé un 85 %. L'espanyol assoleix el seu nivell màxim durant aquest primer període l'agost del 2021, amb una mica més del 14 % dels visitants.



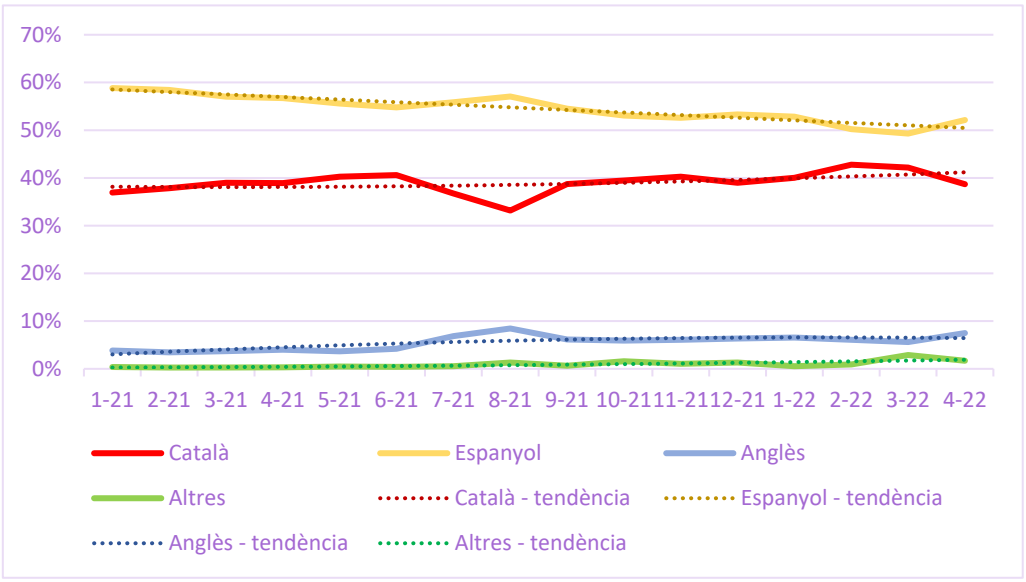
A partir d'abril del 2022 la tendència canvia, especialment en la ràtio català/espanyol. Les visites a continguts en espanyol augmenten fins superar el 18 % l'agost del 2022, el seu nivell màxim dins la sèrie; alhora, es registra la distància mínima amb el català, amb només un 62 %. No es tracta d'un augment puntual, ja que l'espanyol es manté estable en aquest nivell superior i, amb un únic descens el gener del 2023, la tendència de creixement es manté el 2023.

Col·laborador 10

Targeta de col·laborador			
Descripció	Es tracta d'una institució pública amb un gran nombre de llocs web que s'han recollit com a conjunt agregat. L'aspecte multilingüe és necessari tant per la diversitat del públic objectiu com, en menor mesura, pel turisme.		
Correlació CA – ES	-0,71	Correlació CA - EN	-0,46
Té «hreflang»?	El 25 % dels llocs el tenen i estan ben formats. La resta varia però estem considerant que no en tenen.		

Sèrie completa	

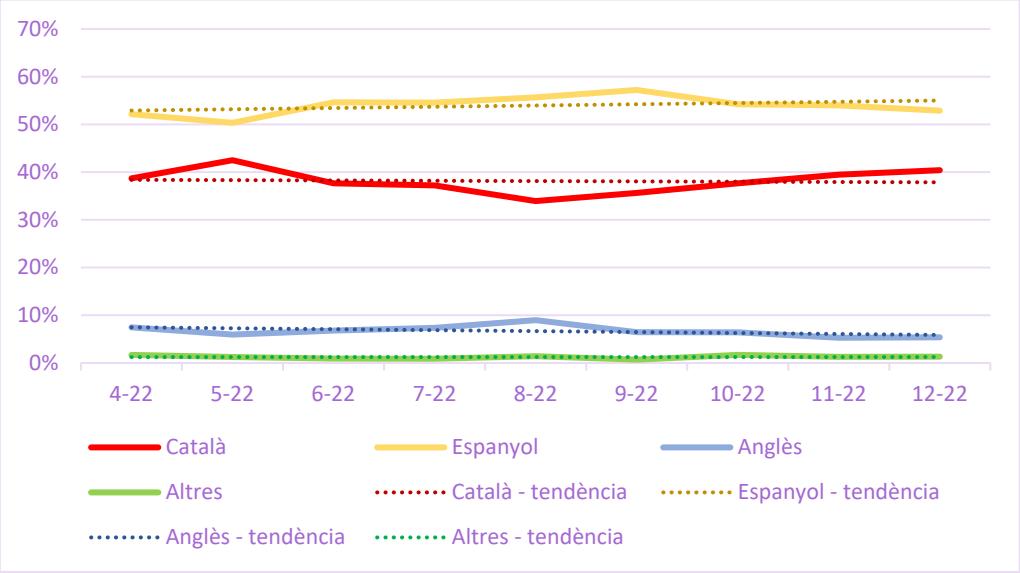
Seguint el mateix model que el cas anterior, el nostre primer gràfic mostra com ha evolucionat el trànsit originat per Google al conjunt de llocs web que hem analitzat:



En aquest cas, els continguts en espanyol també mostren una tendència decreixent mentre que el català mostra el creixement més marcat. Tanmateix, l'anglès i, en menor mesura, la resta d'idiomes disponibles també mostren una tendència de creixement més lenta. Cal destacar que tant l'espanyol com l'anglès mostren un creixement sobtat els mesos de juliol i agost, mentre que el català disminueix.

El nivell màxim de trànsit cap a pàgines en català es produeix en aquest període: febrer de 2022, fins gairebé el 43 % de les visites. La diferència mínima entre les

visites al català i l'espanyol es produeix també en aquest període: al voltant del 7 % el març del 2022.



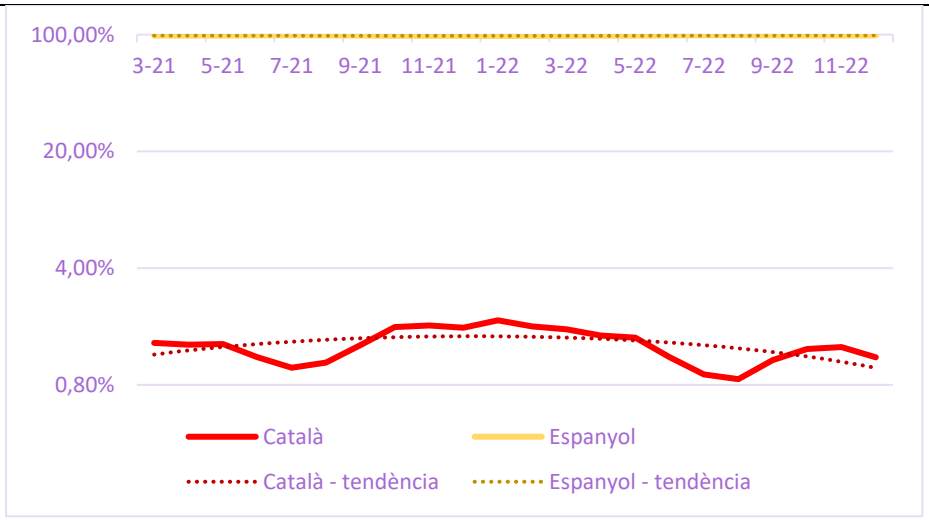
Ara passem al següent període. Una vegada més, les tendències canvien: la disminució de l'espanyol s'alenteix mentre que la resta de llengües deixen de créixer i comencen a disminuir. Pel que fa al català contra l'espanyol, el mateix canvi sobtat que es va produir l'estiu anterior es converteix ara en un sotrac del qual el català no es pot recuperar ni tan sols a finals d'any.

En xifres, el català parteix d'un nivell màxim de gairebé el 43 % del trànsit el febrer i després baixa a menys del 34 % l'agost. D'altra banda, l'espanyol creix fins a més del 57 % el setembre del 2022, apropant-se al seu propi nivell màxim en la sèrie completa, del 59 %.

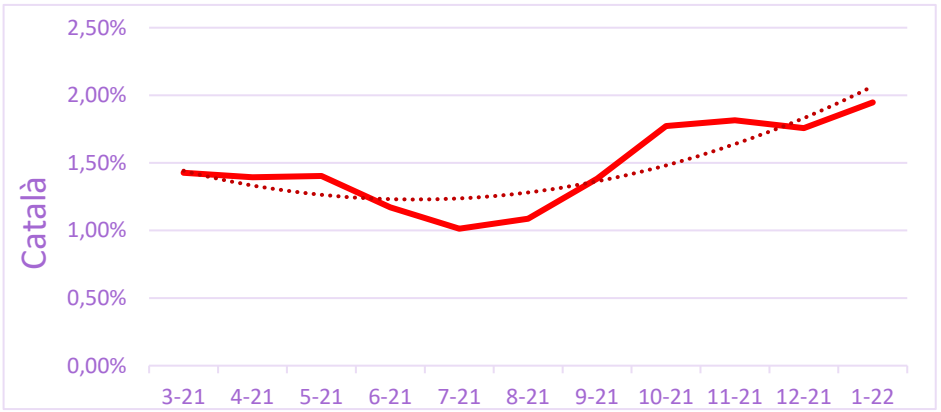
Aquest col·laborador ens ha fet saber que la tendència durant el 2023 ha estat inestable fins ara i esperem que ens faci arribar dades actuals per confirmar-ho.

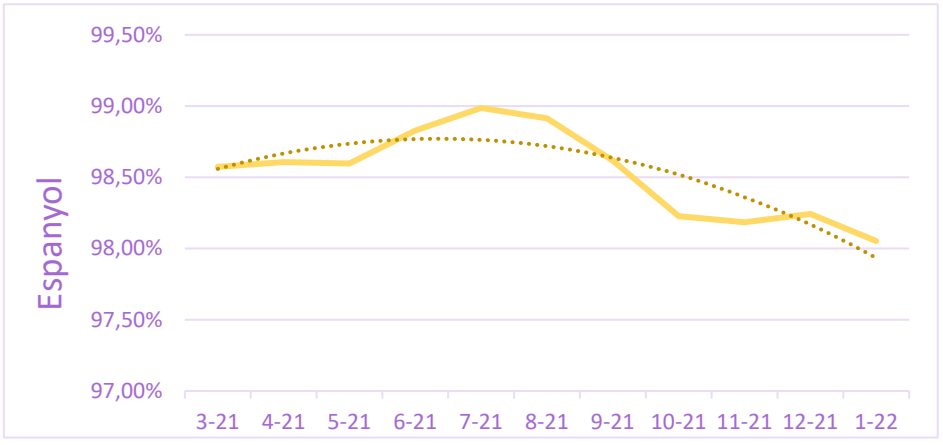
Col·laborador 13

Targeta de col·laborador			
Descripció	Organització mundial que facilita les seves dades de trànsit en català i en espanyol. Opera un dels llocs més visitats d'Internet i dona servei a tot tipus de perfils d'usuari. Diversos motors de cerca n'estan citant directament el contingut.		
Correlació CA – ES	-1,00	Correlació CA - EN	N. a.
Té «hreflang»?	No, però té un tractament específic per indexar-se correctament.		

Sèrie completa	 Nota: aquest gràfic utilitza una escala Y logarítmica a causa de la diferència entre els nivells de trànsit de cada idioma.

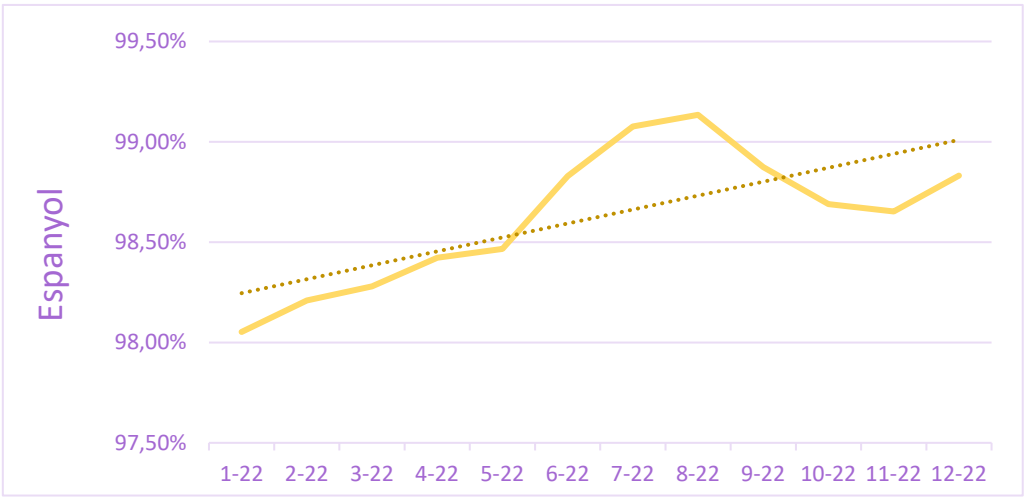
Aquest cas descriu una de les afectacions més marcades entre tots els casos que hem analitzat. Com que es tracta d'un web mundial, hi ha una gran diferència entre el nombre d'usuaris en català i en espanyol, motiu pel qual oferim gràfics separats per a cada idioma que cobreixen la primera part de la sèrie temporal.

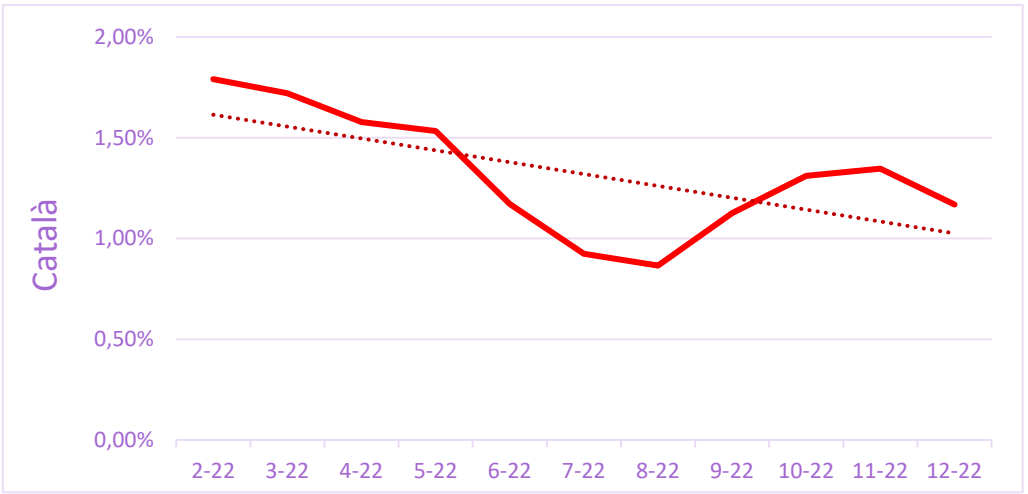




Comencem subratllant que aquest cas sembla tenir un tractament especial per part de Google, probablement per la seva singularitat. Així, les variacions del trànsit originat per la cerca comencen una mica abans, entre gener i febrer del 2022, en lloc de març i abril de 2022 com a la resta de casos.

Durant el primer període (gràfics anteriors), el català arriba a un màxim proper al 2 % el gener del 2022. Alhora, s'observa la diferència més petita entre les dues llengües, lleugerament per sobre del 96 %. Hi ha una clara tendència de creixement del català i una disminució de l'espanyol, ambdós lents. Cal tenir en compte que aquests gràfics mostren una ampliació significativa de la sèrie.



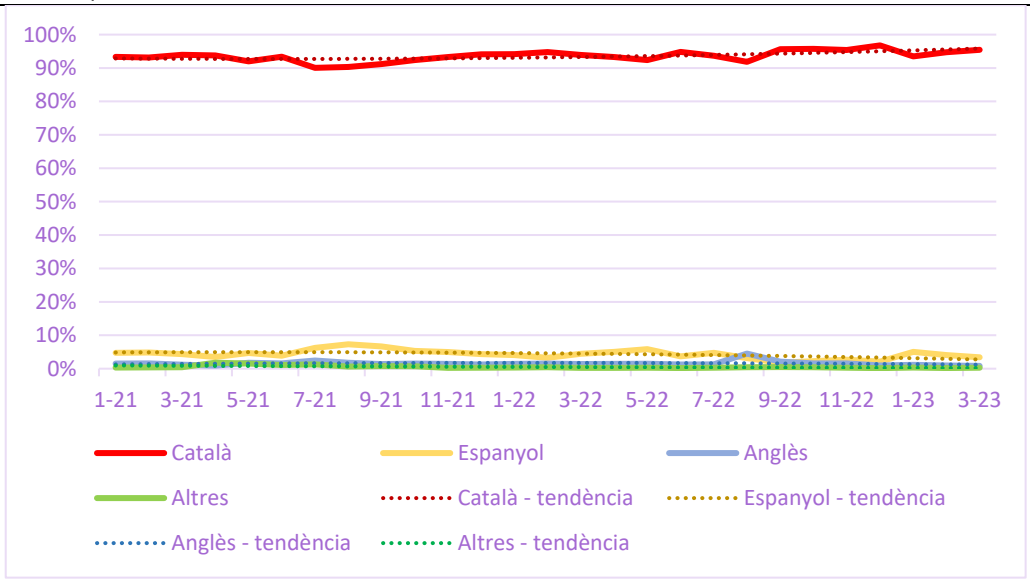


D'altra banda, durant la segona part de la sèrie, el contingut català aconsegueix el seu trànsit mínim, per sota del 0,9 %, mentre que el contingut espanyol assoleix el seu nivell màxim, per sobre del 99 %. Descartant el cicle d'estacionalitat aparent, el mínim de català a l'agost del 2022, que normalment es troba en un punt baix, encara es troba un 0,2 % per sota del 2021. A més, el mateix mes del 2022, observem la distància màxima entre l'espanyol i el català.

Pel que fa a les tendències, fins i tot amb oscil·lacions, els continguts en català prenen un camí a la baixa amb més força que l'anterior augment, mentre que l'espanyol puja i arriba al seu nivell màxim.

Grup 2. Llocs on no s'ha detectat cap impacte

Col·laborador 8

Targeta de col·laborador			
Descripció	Organització centrada en la promoció i la defensa de la cultura i la llengua catalanes. El seu web té una doble finalitat: promocionar les activitats pròpies de l'organització en català i difondre la realitat del dia a dia a l'exterior, oferint també continguts en altres idiomes.		
Correlació CA – ES	-0,85	Correlació CA - EN	-0,42
Té «hreflang»?	Parcial. Els idiomes i les variables estan definits, però falten els enllaços de retorn.		
Sèrie completa			

Com es mostra al gràfic de la sèrie completa anterior, el nivell màxim de visites a continguts en català procedents dels motors de cerca es produeix just quan fa mesos que Google aplica un biaix lingüístic. El desembre del 2022, gairebé el 97 % de les visites procedents d'aquest cercador són en català. Mentrestant, l'espanyol arriba al seu punt àlgid l'agost del 2021 i l'anglès ho fa l'agost del 2022. Per tant, creiem que aquest lloc web no s'ha vist afectat per les variacions en les visites procedents dels resultats de la cerca de Google.

Els valors atípics: llocs web afectats que han aplicat contramesures

El problema ja ha durat més d'un any, fet que ha provocat que els llocs afectats hagin notat canvis sostinguts en el perfil de les seves visites procedents de Google.

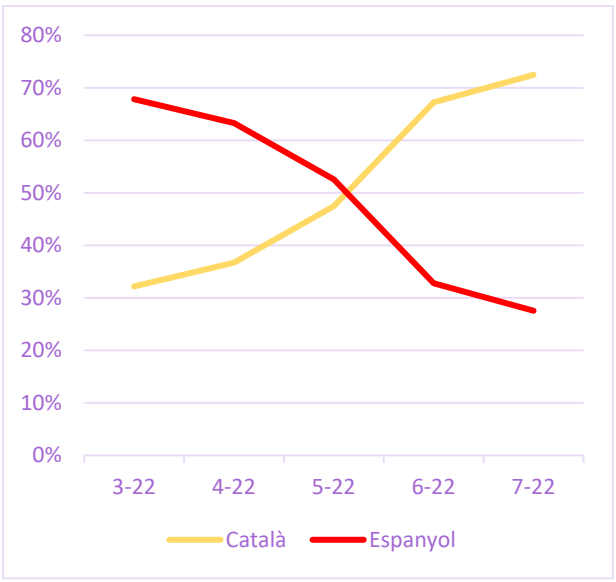
Més enllà de les crítiques a alguns llocs web informatius per part d'usuaris que es troben ara amb continguts en espanyol on abans hi havia continguts en català, en alguns casos aquest problema podria perjudicar el model de negoci o els aspectes principals d'una marca.

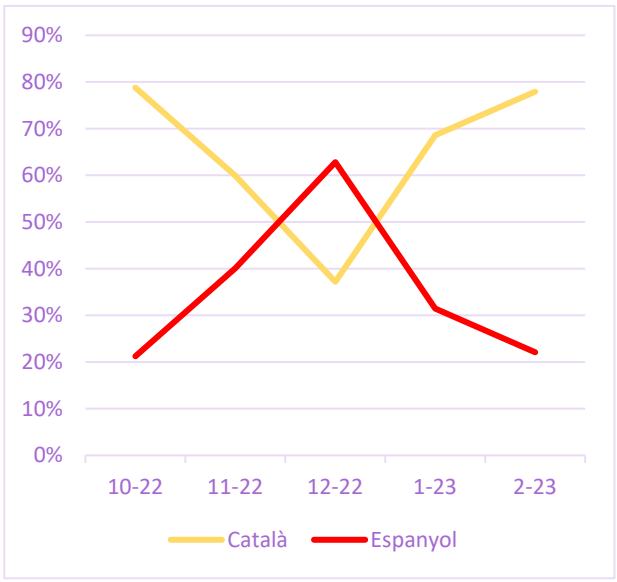
A causa d'això, hem trobat situacions en què els propietaris del lloc web han pres contramesures forçades per evitar que aquesta situació ocasioni encara més problemes. A continuació, descrivim un d'aquests casos.

Col·laborador 1

Targeta de col·laborador			
Descripció	Empresa catalana amb cinquanta anys de trajectòria que gestiona una plataforma de comerç electrònic. Només tenen llocs en català i espanyol perquè no venen a l'estranger. Alhora, això els fa més sensibles a qualsevol canvi de comportament del seu públic objectiu.		
Correlació CA – ES	-0,98	Correlació CA - EN	N. a.
Té «hreflang»?	Sí, utilitzant el format i els enllaços adequats.		
Sèrie completa	Cap valor específic més enllà de l'efecte de les contramesures.		

En aquest cas, hi ha una tendència històrica d'indexació de continguts que dona prioritat a l'espanyol, encara que l'empresa estigui prioritzant activament els continguts en català al lloc web. En base a això, la situació es va tornar insostenible el maig del 2022, motiu pel qual van decidir desindexar tot el contingut en espanyol del lloc web, per aconseguir un canvi de comportament immediat.





Després d'una primera intervenció de desindexació, el trànsit català es va estabilitzar durant uns mesos, fins que, de sobte, va començar a baixar el novembre del 2022. Això va obligar a repetir l'acció anterior i tornar a desindexar de manera forçada els continguts en espanyol el desembre del 2022.

Conclusions preliminars

L'impacte no és general
La nostra anàlisi dels llocs web mostra que no tots els llocs es veuen afectats de la mateixa manera, independentment de l'idioma principal que utilitzen per al seu públic objectiu, tal com es mostra als exemples anteriors. Tanmateix, aquest impacte està present en la gran majoria dels casos, al voltant de dos terços dels llocs que hem pogut examinar.

La força de l'impacte varia
De la mateixa manera que l'impacte no és general, hi ha casos en què la variació del volum de trànsit procedent dels cercadors és menor que altres casos en què aquesta variació acaba modificant el perfil dels usuaris que passen per aquest procés.

No hi ha relació amb el domini (TLD)
TLD (per les seves sigles en anglès) fa referència al domini de primer nivell que el lloc web utilitza al seu nom principal. L'impacte es produeix als llocs web .com, .org, .cat i .es; per això, creiem que l'autoritat del domini no és un factor rellevant en aquest problema.

Hi ha una relació inversa entre el català i l'espanyol
Tal i com es sospitava quan l'Aliança per la presència digital del català es va fer pública, l'augment del trànsit de continguts espanyols procedents de Google implica una disminució del trànsit en català. Aquesta tendència s'ha confirmat en tots els casos analitzats.

Un dels col·laboradors ajuda a confirmar l'anterior: tan bon punt desindexa per força els seus continguts en castellà (una contramesura extrema que apliquen per motius comercials), el trànsit dels seus continguts en català recupera tota la seva visibilitat a les cerques web. Quan la versió castellana es torna a indexar, la catalana torna a baixar.

Per què passa això? Algunes hipòtesis

Complementant l'anàlisi de les dades de trànsit web originades per la cerca de col·laboradors, hem demanat a diversos experts que facin una pluja d'idees sobre els motius que podrien estar afectant la visibilitat del contingut en català als resultats de la cerca. Aquest exercici pretén ajudar els proveïdors de cerca web a descartar algunes vies possibles a l'hora d'investigar el problema. La majoria es refereixen a Google perquè és el motor de cerca dominant (95,9 % de la quota de mercat a Espanya l'abril del 2023 segons StatCounter GlobalStats).

Motius polítics
Quan es va fer evident la pèrdua de visibilitat de la llengua catalana a les cerques web, a les xarxes socials, molts usuaris ho van atribuir a una voluntat de Google de penalitzar qualsevol contingut en català. Aquesta teoria s'ha vist potenciada per la manca de reconeixement o suport públic o institucional per part de qualsevol representant de Google Espanya.

Tanmateix, descartem qualsevol animositat de Google contra el català perquè molts dels seus productes i serveis de consum ja estan disponibles en català.

Impacte dels clics
Generalment, els internautes de parla catalana fan clic als resultats en català i en espanyol indistintament. Per tant, el contingut web en espanyol acostuma a obtenir més clics que el contingut web en català, motiu pel qual Google decideix que el contingut en espanyol és més rellevant.

El problema és que Google no compleix les preferències dels usuaris que s'estimen més veure pàgines en català, de manera que, quan un contingut està disponible en diversos idiomes, com passa als webs multilingües, Google descarta la preferència d'idioma al navegador o al perfil d'usuari i dona més rellevància al lloc amb més visites, que acaba rebent encara més visites.

Codificació de l'idioma
S'ha suggerit que Internet i les potències de la WWW (ICANN, IETF, W3C) han aplicat algun canvi a la codificació de l'idioma del contingut que ha agafat per sorpresa els motors de cerca.

No obstant, sembla ser que aquestes potències no estan aplicant aquest tipus de canvis. Van fer recomanacions que els motors de cerca i els navegadors web poden seguir, però no detallen en quin ordre un motor de cerca ha de mostrar els seus resultats. Per exemple, el W3C defineix etiquetes amb els atributs «hreflang» i «lang» que s'utilitzen per enllaçar pàgines en diversos idiomes al mateix lloc web. Google té en compte aquestes etiquetes, però prefereix

detectar l'idioma per si mateix, sense tenir en compte el codi ISO que s'especifica a les etiquetes:

https://developers.google.com/search/docs/specialty/international/localized-versions

A més d'això, no podem afirmar que el problema també afecti altres idiomes, tot i que han aparegut a Twitter algunes queixes sobre l'ucraïnès.

Google confon el català amb l'espanyol
S'ha dit que Google podria identificar erròniament el contingut en català com a espanyol, possiblement per algun canvi en el motor d'IA que detecta idiomes.

Ho descartem perquè Google no confon el català i l'espanyol i, de fet, pot proporcionar resultats en català si això es força. A més, Google ofereix l'opció (que requereix un clic addicional) de cercar específicament contingut en català a la barra d'eines superior.

Discrepància entre els resultats orgànics i les fitxes informatives
La mateixa cerca produeix resultats diferents als resultats orgànics i les fitxes informatives a la banda dreta (la part superior als dispositius mòbils): els resultats orgànics penalitzen el català mentre que les fitxes informatives respecten la preferència de l'usuari. De vegades, les fitxes informatives fins i tot semblen obsoletes: quan es va realitzar la prova, la cerca de «GSMA» encara mostrava Stéphane Richard com a president de l'associació, mentre que l'article enllaçat de la Viquipèdia ja esmentava José María Álvarez Pallete, l'actual president.

Sembla ser un carreró sense sortida, perquè Wikidata no és l'única font de les fitxes informatives, sinó que també es poden extreure de llocs web oficials i altres repositoris, com IMDb o FilmAffinity, entre d'altres.

El més important és que moltes fitxes informatives mostren text en català, però no perquè extreguin la informació d'una font catalana: en realitat, l'estan agafant d'un lloc en anglès i la tradueixen al català, fet que dona una falsa impressió de normalitat. Aquesta impressió és encara més evident als dispositius mòbils, perquè el fragment/la bústia d'informació es mostra a la part superior dels resultats de la cerca.

El problema d'AdWords
Un especialista en màrqueting SEO/SEM destaca el cas estrany i encara inexplicable (https://twitter.com/EvaOlivaresb/status/1618646946713055232) d'un anunciant que compra paraules clau en català però els usuaris de la cerca veuen els resultats de la cerca en espanyol.

En relació amb l'anterior, s'ha plantejat la hipòtesi que Google se centra en els idiomes majoritaris per impulsar el seu negoci publicitari, aplicant alguna combinació de posicionament i escala que prioritzi l'espanyol en un territori que suposa un mercat rellevant per a l'empresa.

Actualització principal del motor de cerca

Google va realitzar una actualització principal del seu sistema de cerca el maig del 2022, just quan va començar el problema. Tal com afirma l'empresa, aquest tipus d'actualitzacions canvien substancialment el comportament de les cerques. Segons diversos llocs web de SEO, aquesta actualització principal va incloure canvis en els principis rectors de l'algoritme per promoure l'estratègia EEAT (per les seves sigles en anglès, expertesa, experiència, autoritat i fiabilitat). El principal canvi es troba a la segona «E», (experiència). Es pot consultar com afecta l'EEAT a les cerques a les directrius de Google per als avaluadors de qualitat (https://services.google.com/fh/files/misc/hsw-sqrg.pdf).

S'han identificat diverses situacions possibles:

- L'actualització de maig del 2022 va modificar algun aspecte relacionat amb els idiomes «regionals» que va passar desapercebut.
- Els avaluadors de qualitat estan fent quelcom malament amb el contingut en català.
- Alguna altra variable que no hem detectat.

De fet, les directrius anteriors esmenten específicament el català (pàg. 137) a l'apartat d'etiquetatge de continguts en «llengües estrangeres»:

Per exemple, la majoria dels usuaris de parla catalana a Espanya també parlen espanyol. Per tant, per a les tasques de qualificació en català (ES), la bandera de llengua estrangera NO s'ha d'assignar a les pàgines d'aterratge en català, espanyol o anglès.

Això pot ser un indici que hi ha una mala comprensió del català/espanyol com a llengua estrangera, motiu pel qual l'actualització no funciona correctament als territoris amb més d'una llengua vehicular?

Geolocalització

Un usuari amb l'entorn configurat en català realitza una cerca des de França. El primer resultat orgànic apareix en francès i el segon en català: https://twitter.com/jordimash/status/1633522801968570394?s=20

Podria això apuntar a una prevalença de l'Estat (per motius publicitaris?). Tanmateix, ho descartem perquè la mateixa cerca realitzada a França però amb paraules clau mal escrites mostra el primer resultat en espanyol, igual que si es

realitza la mateixa cerca a Catalunya:
https://twitter.com/jordimash/status/1633525602174001160?s=20

Propers passos

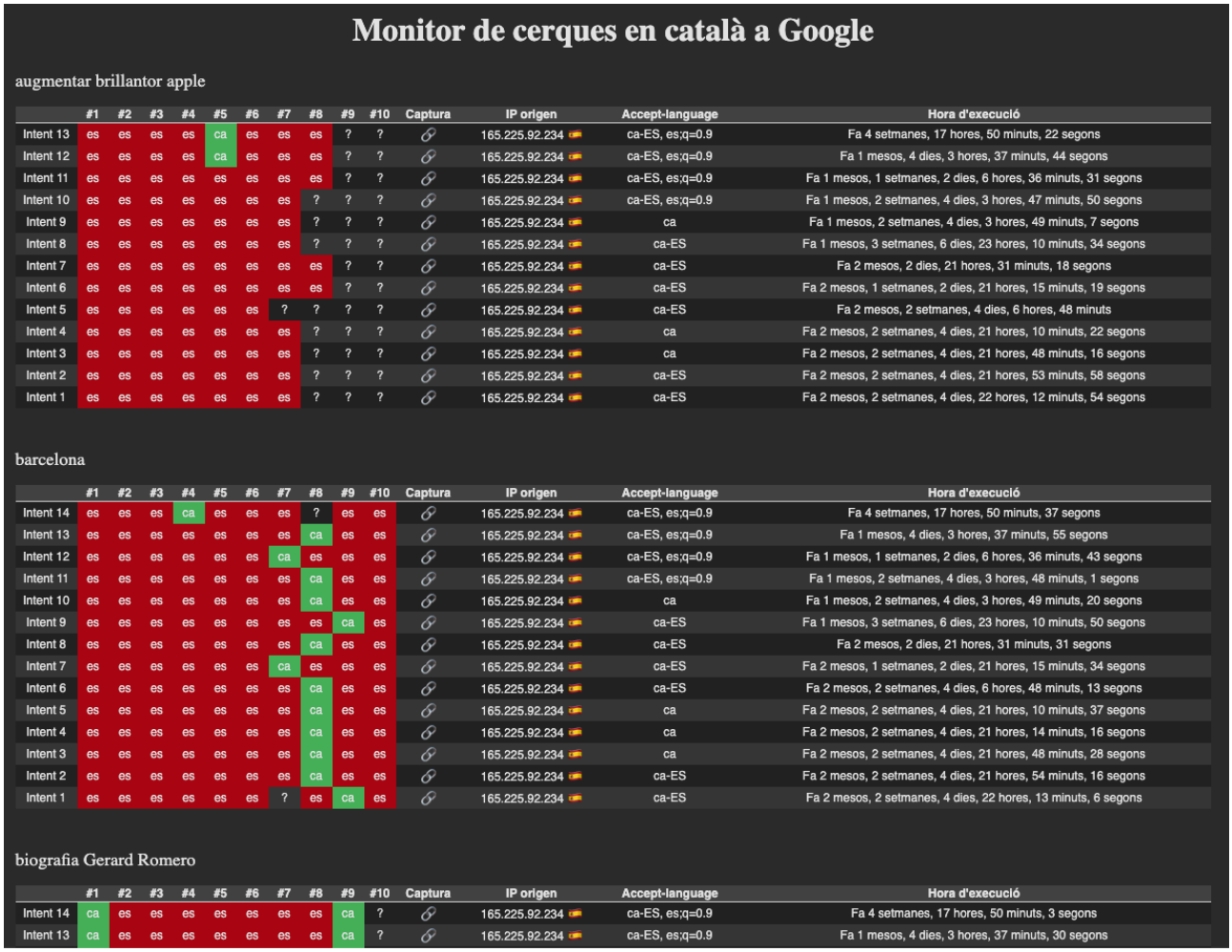
Afegir més col·laboradors

L'anàlisi anterior es basa en les dades facilitades pels tretze col·laboradors que han arribat a temps de poder entrar a l'anàlisi. No obstant això, altres organitzacions també han proporcionat dades molt valuoses després d'aquest moment límit, de manera que aquestes no s'han pogut incloure en aquesta versió de l'informe. Tenim previst examinar aquestes dades i, si són rellevants, les afegirem a un informe actualitzat. En qualsevol cas, aquests col·laboradors addicionals participaran en altres anàlisis posteriors.

Seguiment continuat

Esperem que aquest informe s'enviï a les empreses que operen motors de cerca web perquè puguin utilitzar els nostres resultats per diagnosticar i resoldre el problema el més aviat possible. Mentrestant, continuarem monitoritzant el comportament dels principals cercadors pel que fa als continguts en català; amb aquest objectiu, afegirem noves fonts de dades i informarem a les parts implicades en cas que no s'observi cap millora en un període de temps raonable.

Per fer-ho, l'APDC encarregarà el desenvolupament i el desplegament d'un servei de seguiment permanent genèric, distribuït geogràficament, que notifiqui amb antelació qualsevol empitjorament futur de la visibilitat dels continguts en català als resultats de cerca web. Softcatalà, entitat membre de l'Aliança, ja ha desenvolupat un prototip d'aquest sistema, que s'ha de traslladar a producció (vegeu la captura de pantalla parcial a continuació).



Monitor de cerques en català a Google

augmentar brillantor apple

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Captura	IP origen	Accept-language	Hora d'execució
Intent 13	es	es	es	es	ca	es	es	es	?	?	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 4 setmanes, 17 hores, 50 minuts, 22 segons
Intent 12	es	es	es	es	ca	es	es	es	?	?	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 4 dies, 3 hores, 37 minuts, 44 segons
Intent 11	es	es	es	es	es	es	es	es	?	?	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 1 setmanes, 2 dies, 6 hores, 36 minuts, 31 segons
Intent 10	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 2 setmanes, 4 dies, 3 hores, 47 minuts, 50 segons
Intent 9	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca	Fa 1 mesos, 2 setmanes, 4 dies, 3 hores, 49 minuts, 7 segons
Intent 8	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca-ES	Fa 1 mesos, 3 setmanes, 6 dies, 23 hores, 10 minuts, 34 segons
Intent 7	es	es	es	es	es	es	es	es	?	?	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 dies, 21 hores, 31 minuts, 18 segons
Intent 6	es	es	es	es	es	es	es	es	?	?	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 1 setmanes, 2 dies, 21 hores, 15 minuts, 19 segons
Intent 5	es	es	es	es	es	es	?	?	?	?	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 setmanes, 4 dies, 6 hores, 48 minuts
Intent 4	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 10 minuts, 22 segons
Intent 3	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 48 minuts, 16 segons
Intent 2	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 53 minuts, 58 segons
Intent 1	es	es	es	es	es	es	es	?	?	?	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 setmanes, 4 dies, 22 hores, 12 minuts, 54 segons

barcelona

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Captura	IP origen	Accept-language	Hora d'execució
Intent 14	es	es	es	ca	es	es	es	?	es	es	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 4 setmanes, 17 hores, 50 minuts, 37 segons
Intent 13	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 4 dies, 3 hores, 37 minuts, 55 segons
Intent 12	es	es	es	es	es	es	ca	es	es	es	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 1 setmanes, 2 dies, 6 hores, 36 minuts, 43 segons
Intent 11	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 2 setmanes, 4 dies, 3 hores, 48 minuts, 1 segons
Intent 10	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca	Fa 1 mesos, 2 setmanes, 4 dies, 3 hores, 49 minuts, 20 segons
Intent 9	es	es	es	es	es	es	es	es	ca	es	🔗	165.225.92.234	ca-ES	Fa 1 mesos, 3 setmanes, 6 dies, 23 hores, 10 minuts, 50 segons
Intent 8	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 dies, 21 hores, 31 minuts, 31 segons
Intent 7	es	es	es	es	es	es	ca	es	es	es	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 1 setmanes, 2 dies, 21 hores, 15 minuts, 34 segons
Intent 6	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 setmanes, 4 dies, 6 hores, 48 minuts, 13 segons
Intent 5	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 10 minuts, 37 segons
Intent 4	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 14 minuts, 16 segons
Intent 3	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 48 minuts, 28 segons
Intent 2	es	es	es	es	es	es	es	ca	es	es	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 setmanes, 4 dies, 21 hores, 54 minuts, 16 segons
Intent 1	es	es	es	es	es	es	?	es	ca	es	🔗	165.225.92.234	ca-ES	Fa 2 mesos, 2 setmanes, 4 dies, 22 hores, 13 minuts, 6 segons

biografia Gerard Romero

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Captura	IP origen	Accept-language	Hora d'execució
Intent 14	ca	es	es	es	es	es	es	es	ca	?	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 4 setmanes, 17 hores, 50 minuts, 3 segons
Intent 13	ca	es	es	es	es	es	es	es	ca	?	🔗	165.225.92.234	ca-ES, es;q=0.9	Fa 1 mesos, 4 dies, 3 hores, 37 minuts, 30 segons

A més, es crearà i es publicarà un connector per als principals navegadors web perquè els internautes puguin proporcionar automàticament dades anònimes sobre els resultats de la cerca que obtenen mentre naveguen, en relació amb la configuració d'idioma del seu sistema operatiu, navegador, perfil o dispositiu i les paraules clau que han introduït. Així, aquestes dades col·lectives s'utilitzaran per millorar els resultats de la taula anterior.

Estudis addicionals sobre el paràmetre «hreflang»

Una hipòtesi que ha sorgit en diferents ocasions, fins i tot en diversos fils de debat a les xarxes socials, vincula el problema amb la gestió que fa Google del paràmetre «hreflang».

«hreflang» és un atribut HTML que s'utilitza per especificar en quins idiomes està disponible un lloc web. No es tracta d'un selector perquè els visitants escullin en quin idioma volen que es visualitzi el lloc, sinó d'informació per als algorismes d'indexació que rastregen la xarxa perquè siguin conscients de l'existència d'aquests idiomes.

Partint d'aquesta hipòtesi, estem treballant amb Human Level, una consultora especialitzada en SEO i posicionament web, que investigarà si el paràmetre «hreflang» pot haver influït en el problema amb el català i fins quin punt. L'estudi analitzarà deu llocs web (la meitat dels quals disposa del paràmetre «hreflang» establert i l'altra meitat que no el té), per tal de fer un seguiment de l'evolució del trànsit procedent de Google durant els darrers setze mesos.

Les dades que s'obtinguin es correlacionaran amb diversos elements per avaluar el nivell d'influència de cadascun. Les dades base s'extreuen de Google Analytics i Google Search Console dels col·laboradors.

A més, es realitzarà un altre experiment amb cinc webs més que encara no utilitzen el paràmetre «hreflang», però se'ls ajudarà a aplicar el paràmetre perquè puguem avaluar com evoluciona el seu posicionament als resultats de la cerca.

Actualment estem avaluant quins col·laboradors poden contribuir més en aquest estudi. Pocs dels que participen en la primera fase compleixen els nous requisits, motiu pel qual haurem d'incloure col·laboradors nous.

Crèdits

- Autors de l'informe: Albert Cuesta (albertcuesta@fundacio.cat), Pep Masoliver (jmasoliver@fundacio.cat)
- Gestió tècnica i tractament de dades: Pep Masoliver
- Adquisició de dades: Griselda Casadellà (gcasadella@fundacio.cat)

Membres de l'Aliança per la presència digital del català

- Acció Cultural del País Valencià: Anna Oliver
- Amical Wikimedia: Robert Garrigós i Xavier Dengra
- Fundació .cat: Genís Roca i Roger Serra
- Institut d'Estudis Catalans: Àngel Messeguer
- Institut Ramon Llull: Àlex Hinojo
- Obra Cultural Balear: Llorenç Garcia
- Òmnium Cultural: Iker de Luz
- Plataforma per la Llengua: Marc Biosca
- Softcatalà: Joan Montané i Pere Orga
- WICCAC: Joan Soler

Col·laboradors

L'Aliança per la presència digital del català i els autors de l'informe volen agrair a totes i cadascuna de les entitats que han donat resposta a la nostra crida per facilitar les dades de trànsit dels seus llocs web perquè les puguem tractar per elaborar aquest informe.

Algunes d'aquestes entitats han acceptat expressament que les esmentem aquí:

Ajuntament de Barcelona

Amical Wikimedia

Eurecat, Centre Tecnològic de Catalunya

Generalitat de Catalunya

Institut Ramon Llull

Meteocat, Servei Meteorològic de Catalunya

Òmnium Cultural

Universitat Pompeu Fabra

Universitat de Barcelona

Universitat de Girona

Fundació Mobile World Capital Barcelona

No publiquem cap altre nom aquí per tal de preservar-ne la privadesa, tal com s'indica a l'Acord de confidencialitat compartit amb cadascuna de les entitats. No obstant això, aquesta informació es podria revelar a un o més proveïdors de cerca web, si ho requereixen per analitzar un cas concret més a fons. Aquesta divulgació només es realitzarà en casos específics i únicament amb l'acceptació prèvia del col·laborador, ja que també està coberta per l'Acord de confidencialitat.

Annex 1. Especificacions tècniques

Quines dades es requereixen?

Resum del recull de dades – arxiu

Tipus de lloc web	Webs multilingües
Trànsit observable	Trànsit orgànic procedent de cercadors
Freqüència	Diària
Format	CSV
Dades a aportar	Visites per cercador i pàgina de destí (directe de Google Analytics). Visites per cercador i idioma relatives a 100
Període	1/1/2021 - actualitat

Webs multilingües

La incidència detectada als cercadors afecta principalment als webs multilingües, fent que posicioni millor la versió dels continguts en un idioma diferent al preferit per l'usuari. És per això que la recollida de dades es centra en aquest tipus de llocs web.

En paral·lel, també es fa seguiment d'un grup de llocs web de control monolingües en català, per a conèixer de la tendència de navegants amb aquesta preferència d'idioma.

Trànsit orgànic procedent de cercadors
A nivell del trànsit que rep el lloc web, l'interès es centra en el trànsit provinent de cercadors de forma orgànica, és a dir, el que arriba resultat de la indexació natural del lloc, descartant els enllaços patrocinats o les promocions de qualsevol tipus.

És important considerar i incloure tots els cercadors (Google, Bing, DuckDuckGo, ...) en el moment d'exportar les dades. A més a més, en cas que el lloc web ofereixi els continguts en més de 2 llengües, és important tenir-les en compte: a banda de les visites cap a les versions espanyoles d'aquests llocs, aquesta situació també en pot haver provocat transvasament de visites cap als altres idiomes en menor mesura.

Diària
Més enllà de certificar el biaix lingüístic que hi pot haver en els resultats de les cerques, el mateix estudi de posicionament pretén detectar en quin moment s'han aplicat els possibles canvis negatius als motors de cerca. És per això que s'espera rebre les dades amb una freqüència diària o, com a molt, setmanal, per a així poder situar temporalment les variacions de comportament.

CSV
El format preferit per a l'exportació és el CSV, és a dir, el de valors separats per comes, per ser un estàndard disponible a la gran majoria de programari (per exemple, a Google Analytics i, al mateix temps, oferir una estructura molt fàcil de treballar-hi.

En cas que no sigui possible aportar les dades en aquest format, cal que es faci en un d'alternatiu editable i que sigui de fàcil importació. Es descartaran les dades aportades com a imatge, per la complicació en la importació.

De cara a simplificar la tasca d'extracció, també s'ofereix la possibilitat de crear una vista al programari utilitzat per monitoritzar el trànsit al web (com ara Google Analytics) i compartir-ne l'accés amb la bústia posicionament@fundacio.cat. Aquesta servirà per a descarregar les dades de forma automàtica quan sigui necessari.

Dues possibilitats per triar:

Degut a la sensibilitat que pot generar el fet de tractar amb dades de trànsit, s'ofereixen dues possibilitats de col·laboració. En tots dos casos les dades es tractaran de forma confidencial, com es descriu més endavant, de manera que no es pugui associar les mètriques al corresponent titular sense permís explícit d'aquest.

Cal triar una de les tres possibilitats següents:

A

VALORS ABSOLUTS

Visites per cercador i pàgina de destí/aterratge, és a dir, valor absolut de les visites rebudes per trànsit orgànic d'un cercador concret a una pàgina concreta (no cal concretar-ne la llengua, es detecta automàticament). Per a aquest cas, s'inclou una descripció de com exportar-les pels usuaris de Google Analytics com a annex a aquest document.

VALORS RELATIUS

Visites per cercador i llengua relatives, que consisteix en elaborar les dades de les visites i convertir-les en relatives per ocultar els valors absoluts. Per exemple, agafant aquesta taula de dades absolutes:

	Cercador	Català	Espanyol	Altres	Total
01/01/2021	Google	70	240	20	330
	Bing	20	35	15	70
	DuckDuckGo	5	10	2	17
2/1/2021	Google	100	250	35	385
	Bing	30	55	20	105
	DuckDuckGo	8	13	4	25

B

Aleshores, la versió relativa equivalent correspon a:

	Cercador	Català	Espanyol	Altres	Total
01/01/2021	Google	21,2 %	72,7 %	6,1 %	100
	Bing	28,6 %	50,0 %	21,4 %	100
	DuckDuckGo	29,4 %	58,8 %	11,8 %	100
2/1/2021	Google	26,0 %	64,9 %	9,1 %	100
	Bing	28,6 %	52,4 %	19,0 %	100
	DuckDuckGo	32,0 %	52,0 %	16,0 %	100

El refinament d'aquestes dades depèn dels idiomes inclosos en cada cas, de manera que no se'n pot fer una guia de càlcul generalitzable.

VALORS RELATIUS AMB VARIACIONS DE TRÀNSIT
Visites per cercador i llengua agafant el total de visites del primer dia com a valor de referència 100. El valor relatiu total oscil·larà a l'alça o a la baixa, depenent de l'evolució del trànsit de cada dia del període respecte del primer.

Per exemple, agafant aquesta taula de dades absolutes:

	Cercador	Català	Espanyol	Altres	Total
01/01/2021	Google	70	240	20	330
	Bing	20	35	15	70
	DuckDuckGo	5	10	2	17
2/1/2021	Google	100	250	35	385
	Bing	30	55	20	105
	DuckDuckGo	8	13	4	25

C

Aleshores, la versió relativa equivalent, tenint en compte les variacions en el total de trànsit, quedaria com:

	Cercador	Català	Espanyol	Altres	Total
01/01/2021	Google	21,2	72,7	6,1	100
	Bing	28,6	50,0	21,4	100
	DuckDuckGo	29,4	58,8	11,8	100
2/1/2021	Google	30,4	76,0	10,6	117
	Bing	42,9	78,6	28,5	150
	DuckDuckGo	47,0	76,5	23,5	147

En el cas del tercer dia, el total de visites es tornaria a comparar amb el primer, i així successivament.

El refinament d'aquestes dades depèn dels idiomes inclosos en cada cas, de manera que no se'n pot fer una guia de càlcul generalitzable.

Període

La sèrie de dades sol·licitada va des de l'1 de gener de 2021 fins a l'actualitat. Això és degut a la necessitat de poder disposar d'una sèrie completa d'un any sense incidències remarcables, abans de 2022.

Aquesta monitorització té vocació de continuïtat, per poder fer una detecció ràpida de problemes que hi pugui haver amb els cercadors en un futur, tal com es descriu més endavant. En qualsevol cas, es tracta d'una opció voluntària i que s'ofereix a banda de l'aportació de dades inicial.

A qui es sol·liciten dades?

Les dades es sol·liciten a partir de dues vies de col·laboració:

- Sol·licitud directe de les dades, per formar part o ser en si mateix una organització d'especial interès o que pugui haver sofert el biaix per part dels cercadors.
- En un futur, es plantejarà una convocatòria pública d'aportació de dades, oberta a tots els gestors de llocs web multilingües que hi vulguin contribuir. En cas, les aportacions es condueixen a partir d'un lloc web des d'on es pugui escollir el marc de col·laboració.

A més a més, no es descarta cap via d'obtenció de dades directe, ja sigui, per exemple, fent consultes directes als cercadors sobre l'evolució del posicionament de paraules concretes o altres mètodes alternatius de captació que no depenguin dels anteriors.

En quines condicions?

Els col·laboradors que aporten dades de trànsit decideixen quin tipus de reconeixement volen rebre i quin tractament s'ha d'aplicar a aquestes dades. Per això s'estableixen 2 nivells de col·laboració:

- Totalment anònim: no s'incorpora el logotip del col·laborador, ni es cita el seu nom enlloc.
- Col·laborador amb dades anònimes: s'incorpora el nom i el logotip a la llista de col·laboradors, però les seves dades mai es citen de forma pública.

Les dades de trànsit dels llocs web seran tractades de forma confidencial i no es publicaran en cap cas de forma individual sense disposar de permís explícit per part del corresponent titular, d'acord amb els supòsits anteriors.

En cas que les dades d'un lloc web concret s'hagin d'utilitzar a tall d'exemple i sempre de forma privada, per mostrar la situació de cara a Google o els responsables de qualsevol motor de cerca, se n'informarà el titular perquè en tingui coneixement, sense necessitat d'aprovació específica i en cap cas podran ser utilitzades ni publicades per part de tercers.

La Fundació .cat, com a dipositari de les dades de trànsit, es compromet a signar documents de cessió de dades (NDA) amb totes les parts, especificant-hi les limitacions d'ús escollides pel col·laborador, segons els perfils descrits amb anterioritat, i responsabilitzant-se de la custodia.

En cas d'utilitzar dades de cara a publicacions generals, aquestes es treballaran de forma anònima, agrupant-les per categories o en un sol conjunt, de manera que no sigui possible reconèixer-ne la titularitat.

Tot i que l'anàlisi inicial posa el focus en la degradació del servei soferta durant l'any 2022, la voluntat és poder mantenir aquestes mètriques de forma constant per poder detectar noves variacions amb agilitat. En aquest sentit, els mateixos documents de cessió de dades ja preveuen la possibilitat d'allargar aquesta col·laboració més enllà, fent que es renovi de forma automàtica.

Com es processaran?

Un cop recollides, les dades de cada lloc web es revisaran per detectar la tendència evolutiva del trànsit orgànic provinent dels cercadors: si mostren una tendència negativa pel català, és a dir, que les visites aportades des de cercadors disminueixen a favor de l'espanyol o altres idiomes, s'incorporaran a l'arxiu d'evidències. En cas que no mostrin aquesta tendència, se'n farà un registre estadístic per comptabilitzar els casos que no han sofert el biaix de servei.

Les dades que formin part de l'arxiu d'evidències s'analitzaran més a fons amb l'objectiu de quantificar la pèrdua d'usuaris de la versió catalana a favor de les altres. D'aquesta manera es pot elaborar un indicador acumulat del total de visitants que ha perdut el català per culpa dels cercadors durant el darrer any. Al mateix temps, el mateix procés registrarà els casos amb les baixades més acusades, per poder-los utilitzar d'exemple de cara als motors de cerca, si és necessari.

Annex 2. Carta formal de sol·licitud

Barcelona, 1 de març del 2023

Benvolgut, benvolguda,

Us escrivim per demanar-vos la col·laboració de la vostra organització en el diagnòstic del problema que afecta negativament la visibilitat del contingut web en llengua catalana.

Probablement esteu al cas que, des de fa uns mesos, les pàgines web en català han perdut presència en els resultats dels cercadors en favor de les versions del mateix contingut en altres idiomes, fins i tot si l'usuari té configurat el seu entorn de navegació per donar preferència al català.

El fenomen es va començar a observar durant l'any 2022, però no s'ha determinat la data exacta. Tampoc se'n coneix el motiu, malgrat les nombroses consultes informals realitzades a les empreses de cerca per part de particulars i entitats interessades.

És per això que les principals entitats de promoció i defensa del català (Amical Wikimedia, Fundació .cat, Institut d'Estudis Catalans, Institut Ramon Llull, Òmnium Cultural, Plataforma per la Llengua, Softcatalà, WICCAC) ens hem aliat per dedicar recursos tècnics, de coneixement i capacitat de mobilització a diagnosticar el problema i contribuir a revertir-lo.

El Govern de la Generalitat s'hi ha implicat activament i ha encomanat a la nostra aliança la creació d'un dispositiu que quantifiqui la pèrdua actual de visibilitat del català en les cerques web respecte a la situació anterior, i serveixi per abordar les empreses digitals reclamant solucions, essent capaç de detectar incidències similars en el futur.

La Fundació .cat ha assumit, amb el suport de la resta d'entitats, la confecció de l'informe inicial de situació, a partir de dades objectives de trànsit procedents de cercadors en un gran nombre de webs que ofereixen contingut en català i en altres llengües.

Considerem que el web o el conjunt de webs de la vostra organització ens pot proporcionar informació molt valuosa en aquest sentit. Per aquest motiu us demanem que ens proporcioneu com més aviat millor les xifres de visites procedents de cercadors a cadascun dels idiomes que ofereix el vostre web, d'acord amb el document d'especificacions tècniques que adjuntem.

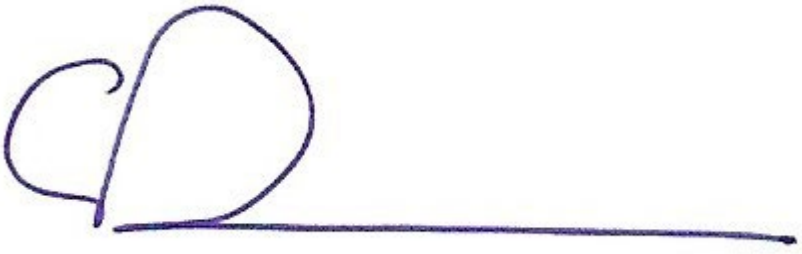
Com veureu, us demanem les dades dia a dia, entre l'1 de gener de 2021 i el 31 de gener de 2023, per tal de detectar en la comparació interanual el moment en què va sorgir l'anomalia i quantificar la magnitud de l'afectació.

Addicionalment, us convidem a continuar proporcionant mensualment les dades actualitzades, amb l'objectiu de mantenir la monitorització continuada de la presència del contingut en català als cercadors web.

En qualsevol cas, mantindrem la confidencialitat de les dades que ens proporcioneu, que només es donaran a conèixer de manera agregada, sense possibilitat d'identificar-ne l'origen individual. Així ho recollim en el compromís de confidencialitat que també adjuntem.

Confiem poder comptar amb la col·laboració de la vostra organització, que us agraïm per avançat. Per a qualsevol aclariment, podeu escriure'ns a suport@fundacio.cat o bé trucar-nos al 93.675.03.54.

Ben cordialment,



Genís Roca
President de la Fundació .cat

Albert Cuesta
Coordinador de l'Aliança per la presència digital de la llengua catalana

Annex 3. Acord de confidencialitat amb els col·laboradors

L'Aliança per la presència digital del català (d'ara endavant, "l'Aliança") adquireix el present compromís de confidencialitat amb l'entitat col·laboradora en l'estudi de la presència digital del català

- Tota la informació que la Fundació (Fundació .cat) rebi o a la que hi accedeixi que provingui de l'entitat col·laboradora en relació a la seva participació a l'Aliança és, per defecte, totalment confidencial, sense importar-ne el suport, el moment, ni el mètode pel qual es facilita aquesta informació.
- Que la informació que rebi la Fundació només podrà ser usada pels empleats i/o col·laboradors que imperativament hagin de tenir-hi accés a fi de desenvolupar les tasques necessàries per l'estudi.
- Tot el persona laboral i col·laborador de la Fundació .cat està subjecte al manteniment de la confidencialitat de la informació a la que poden tenir accés en l'exercici de les seves funcions mitjançant clàusules contractuals específiques.
- La Fundació .cat no transmetrà la informació proporcionada per l'entitat col·laboradora a terceres parts, excepte en els casos en què es mostri als cercadors per mostrar els resultats de la investigació, i només si ambdues parts acorden expressament que es faci.
- L'intercanvi de correus electrònics, sessions de xat, SMS, trucades, o qualsevol altre tipus de comunicacions electròniques entre les dues parts romandran estrictament confidencials sense data de caducitat i sotmeses al secret professional.
- La Fundació .cat garanteix a l'entitat col·laboradora la seva plena adequació a les disposicions legals establertes per la normativa de protecció de dades, i garanteix que ha establert tots els mitjans materials tècnics. humans i legals necessaris per garantir el compliment d'aquest conjunt de mesures de confidencialitat.

La Fundació .cat emetrà un informe final que compartirà de manera individualitzada amb totes les entitats col·laboradores, i un informe global amb les dades agregades que farà públic. Aquest informe global no contindrà referències individualitzades a cap de les entitats col·laboradores, ni permetrà la identificació de casos concrets.

Aliança per la presència digital del català

Plaça Nova 5, 7a planta,
08002 Barcelona
936 750 354

info@aliançadigital.cat