



Report on the loss of visibility of Catalan-language content in web search results

6th June 2023



fundació .cat



III institut
ramon llull



III ÒMNIUM
LENGUA CULTURA PAÍS



WACCAC



This license allows you to distribute, adapt and build on this study, even commercially, as long as the original creation is credited.

Table of contents

Introduction.....	4
Executive Summary	4
Background	12
Examples	15
Goals of this report.....	19
Methodology.....	20
General characterization of Contributors	20
Regarding the web search engines.....	21
Aggregated results	23
Site by site results.....	26
Language behavior of analyzed web sites	26
Meaning in the context of this report.....	27
Group 1. Sites where impact has been found.....	28
Group 2. Sites where no impact has been found	38
The outliers: impacted web sites that have applied countermeasures.....	39
Preliminary Conclusions.....	41
The impact is not general.....	41
The strength of the impact varies	41
No relation to domain (TLD).....	41
There is an inverse relation between Catalan and Spanish	41
Why is this happening? Some hypotheses.....	42
Next Steps	46
Further studies related to the hreflang parameter	48
Credits.....	48
Members of the Aliança per la Presència Digital del Català	49
Contributors.....	50
Annex 1. Technical Specifications.....	51
Which data is required?	51
Annex 2. Formal letter of request.....	57
Annex 3. Non-disclosure agreement with Contributors.....	59

Introduction

Catalan-speaking web users have been noticing since mid-2022 that Spanish content prevails in their organic web search results, even when the same or similar content is available in Catalan and they have configured their device/OS/browser/account to prioritize results in Catalan. Such **language preferences were respected in the past, but now they are ignored.**

Major web search providers have been approached about the issue, but they have refused to offer an explanation and, eventually, a solution. Instead, they have asked for the issue to be documented beyond users' subjective perception.

Thus, the Aliança per la Presència Digital del Català (Alliance for the Digital Presence of Catalan), at the request of the Government of the Generalitat de Catalunya (Regional Government of Catalonia), has undertaken the task of providing such documentation on the issue. Under the guidance of Fundació.cat, a member of the Aliança, it has analyzed web traffic data of over 600 multilingual websites in order to trace how traffic on their Catalan versions has evolved over time with regard to other languages.

Our main findings suggest that **66.5% (more than two thirds) of the websites have been (and still are) affected by the issue**, causing a loss of traffic on their Catalan versions. Moreover, there is a strong correlation (80% on average) between traffic in Catalan and Spanish, with the Spanish version accounting for almost one page view for each one the Catalan version loses. This correlation is much smaller (0.25) when comparing Catalan and English.

As the issue does not impact every website in the same way, this report studies common profiles of several websites, including some that have not been impacted at all.

Many of the impacted web sites are among the most relevant and heavily visited Catalan organizations, including the government, academia, media, and business sectors with a Catalan domain that publish their web content in Catalan.

This report is being made available to major web search providers for them to use it in their efforts to restore the visibility Catalan content has lost. It will also be available to the general public and the media, as well as to select members of the European Parliament who are active in matters related to minority EU languages for it to be used in their legislative initiatives.

This report comes at a critical time, when the emergence of AI chatbots is changing the way users search and interact with digital content, as such chatbots seem to draw mostly from content in major languages. Hence, it is crucial for the presence of original content in such non-major languages to be restored before chatbots prevail in regular web searches.

Executive Summary

Impact on the positioning of Catalan-language content in search engines

Since mid-2022, an issue has been identified regarding Catalan content and Internet search engines. **For multilingual websites, Catalan content disappears from the top positions of search engines, even when the search is performed in Catalan and the browsing environment is configured so as to give preference to Catalan.**

Before this happened, Catalan content on multilingual websites could be found in the top search positions, if the search engine itself interpreted (either because of the user's settings and preferences or because of the language used in the search itself) that this was the language in which the user made the query.

The exact start date of this behavior has not been set, nor is the reason known, despite inquiries made to search engine companies.

A data-based study

We at the Alliance for the Digital Presence of Catalan, as requested by the Generalitat de Catalunya (Regional Government of Catalonia) have tried to find out more about the issue, in order to quantify and observe it, to know how it impacts website traffic, and thus to have more and better arguments to claim with regard to the players involved (mostly Internet search companies) **so they can revert the situation and the preferences of the users when deciding in which language they want to receive search content can be honored again.** Moreover, we may be able to refute or confirm our hypotheses about what is happening and why.

In order to obtain data to carry out our study, **we have requested collaboration from several organizations and entities** that operate multi-language websites including Catalan-language content. They have been asked to **provide us with information on the web traffic** they have got during the years 2021 and 2022, up to the available data of 2023.

What data we have based our analysis on

We have analyzed the **organic web traffic from search engines of 639 multilingual websites** that include content in Catalan and one or more additional languages.

These websites are operated by organizations from the Catalan public, academic, media and business sectors that have agreed to collaborate with the study by the Alliance for the Digital Presence of Catalan, at the request—on their behalf— of the Fundació .cat, which collected the data and was in charge of drawing up the report, with the help of all Alliance members.

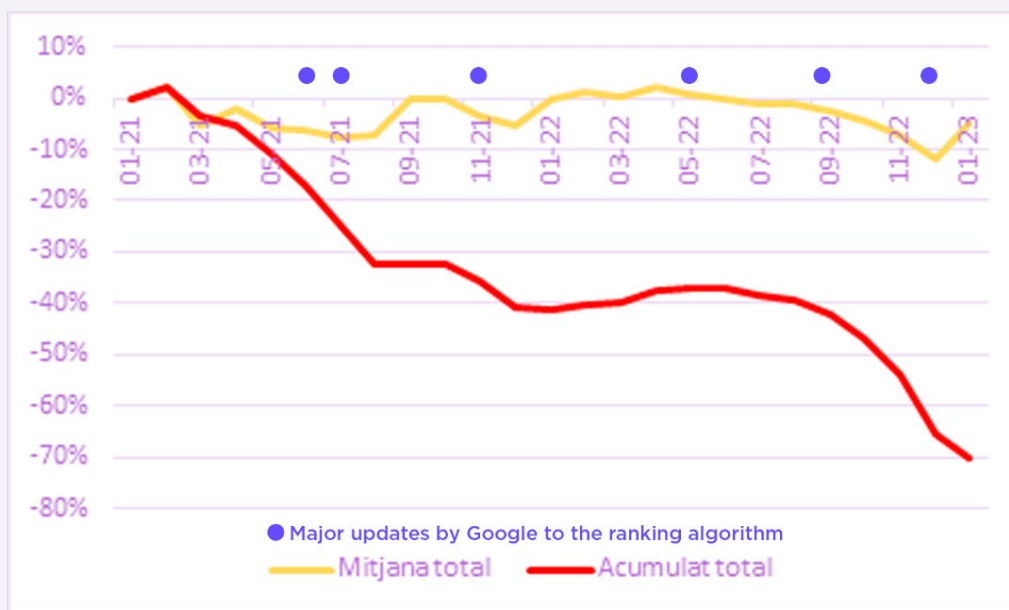
Some of these organisations do not wish to make their collaboration public, while others do. We appreciate the collaboration of all of them.

Main study results

Has Catalan-language content been losing visibility?

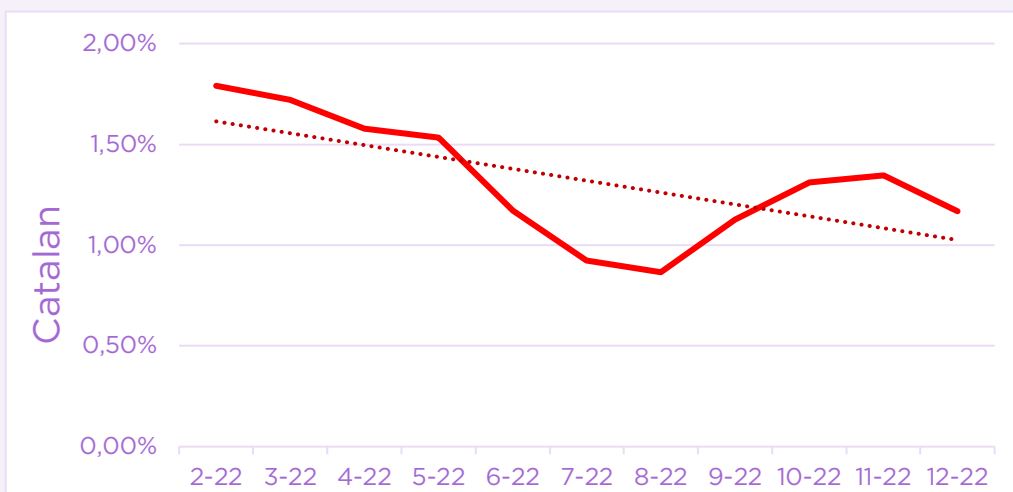
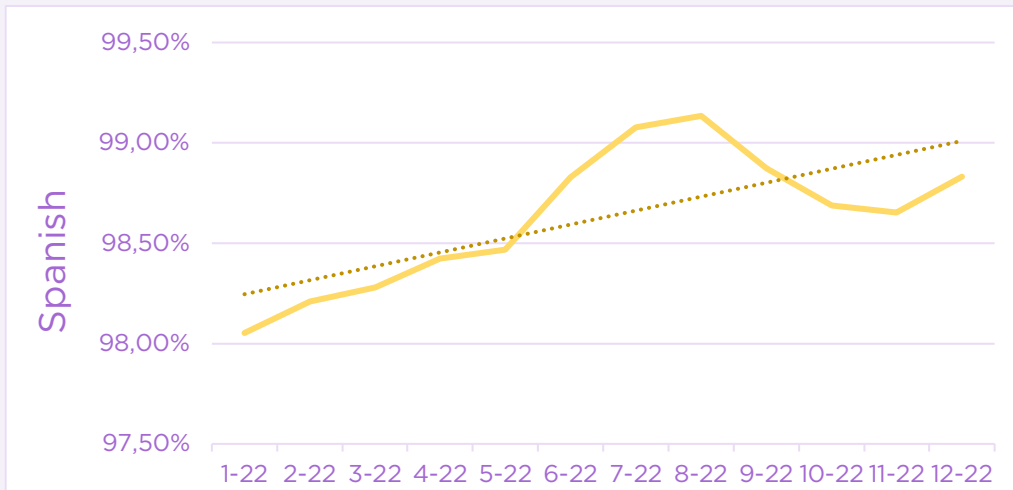
The change in the traffic trend from the Google search results (on which the current study focuses) is actually detected in a generalized way during the spring of 2022 and persists until today.

This becomes obvious in the following chart, which shows how the ratio of visits to Catalan- and Spanish-language content has evolved for the full set of analyzed web sites over the two-year period considered.



We have also focused on each of the impacted web sites, looking at the evolution of traffic to each language.

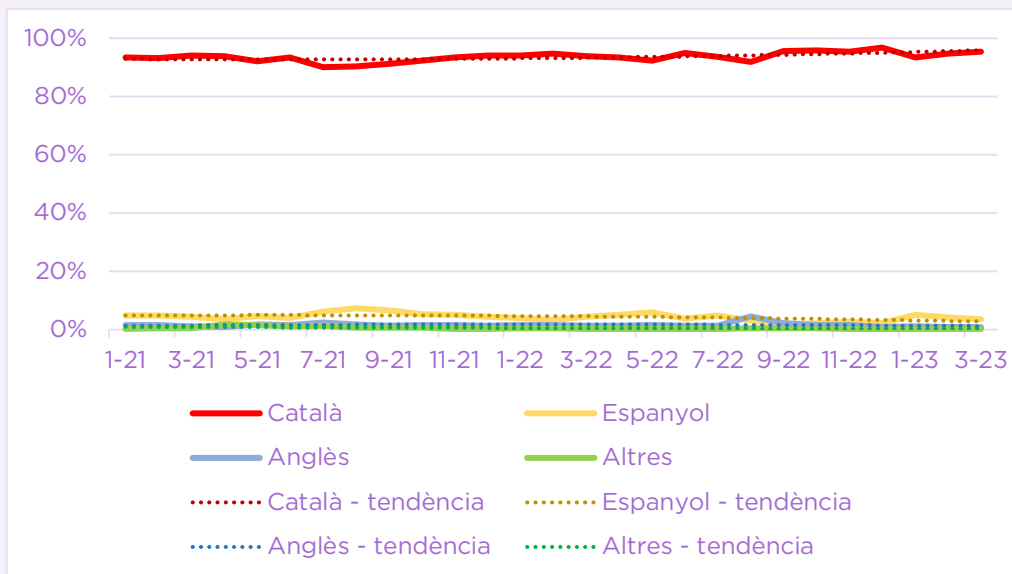
Example of data contributor no. 13



Is this impacting all web sites? How many of them are impacted?

Not all websites are impacted. According to the study, 66.5% of the analyzed websites have been impacted by the issue, thus losing traffic to their Catalan versions.

Example of data contributor no. 8 (unaffected)



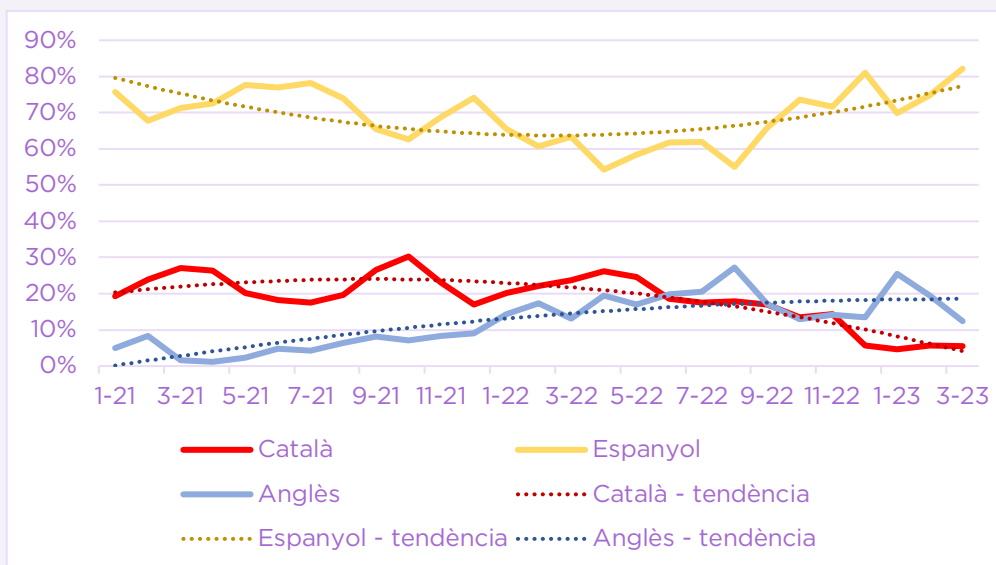
Is there any correlation between the loss of visibility of Catalan content and the increase in visits to pages in other languages, mainly Spanish?

There is a strong correlation of 80% between traffic in Catalan and traffic in Spanish, i.e. the Spanish version gains almost one pageview for every pageview the Catalan version loses.

It is not only that the visits in Catalan decrease, nor the total number of visits, but it is also observed that many of the visits that were previously in Catalan are now visits to the content in Spanish.

A specific site has managed to restore the traffic to its Catalan version only by forcibly de-indexing the Spanish version. As soon as the latter gets indexed again, it gets precedence over the former.

Example of data contributor no. 13



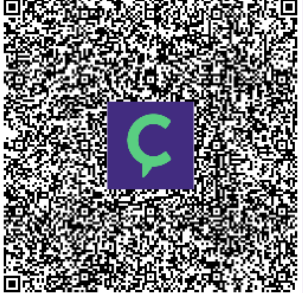
Do we have verified that the users' language preferences are being ignored?

Search engines no longer respect users' explicit language preferences. Catalan content has lost visibility in search results, regardless of the language settings of the device, browser, and user profile. In order to check this, we have set up our own controlled test environment:

Computer equipment used

OS / version	Windows 10 Pro 2H22 19045.2846		
Browser / version	Chrome 113.0.5672.126		
Browser context			
ACCEPT-LANGUAGE	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		
Cookie context	No general, nor Google cookies previously set		
Geolocation context			
Public IP	141.166.99.42	Google location	43470 La Selva del Camp

Query

Query literal	barcelona		
Query URL	Link Due to the URL's length, it's easier and more user-friendly to use a direct link.	URL's QR	

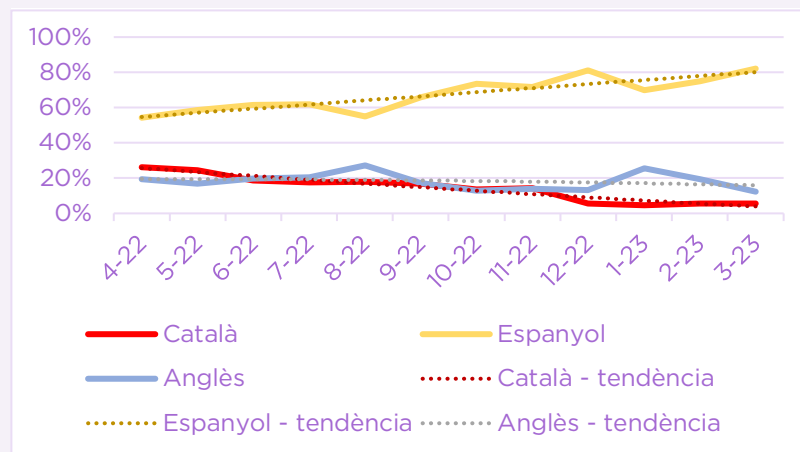
Search results

Results screenshot (top, not promoted content)			
# first result in Catalan	None in the first page of results.	# first result in Spanish / other language	1

When did the issue start?

The study shows that the change in trend of traffic contributed from the Google search engine can be identified in a generalized way during the spring of 2022 and persists until today.

Example of data contributor no. 13



- The effect can be seen on .com, .org, .cat or .es websites, hence the domain authority is not a relevant point in this sense.

Can the issue be caused by wrong labeling by the web sites?

The *hreflang* parameter has been raised as a possible cause of the issue. However, we have checked how *hreflang* is being used by all analyzed web sites, and verified that some of the impacted ones have the right configuration, while some of the not-impacted ones don't have it.

Background

Catalan is a Romance language with 10 million speakers across an area spanning four European countries (Andorra, Spain, France, and Italy). It is the 9th most widely spoken language in the European Union, alongside Greek, Czech, and Portuguese and ahead of Swedish and Danish. According to the World Language Barometer¹ by the French Ministry of Culture, Catalan is the 12th most influential language in the world. Moreover, it is the second language in Spain by number of speakers, well ahead of English.

Catalan is highly present on the internet: whereas it is the 75th language in the world by number of speakers, it is consistently rated between position 10 and 20 in terms of internet presence. Such reasonable satisfaction as to the presence of Catalan on the internet has given way in recent months to general frustration among Catalan-speaking internet users due to the marginalization of Catalan digital content in web search results: if a page is available in Catalan and Spanish, the Catalan version appears below the Spanish version in web search results, i.e. subordinated, and **this happens regardless of user preferences**. This phenomenon benefits clicks on Spanish websites and affects all other search engines as well, although it is more visible on Google, as it is the most widely used search engine. Everything suggests that the source of the problem, which is still undiagnosed, lies elsewhere.

The digital visibility of Catalan on the internet is threatened by this phenomenon. Content in Catalan continues to be published and can be accessed directly as usual, but it is losing visibility because many internet users are accessing it through web searches. What is more, a technical effect is causing both Google and other search engines (Bing, DuckDuckGo, Qwant, etc.) to show the results of the Spanish version of websites rather than those of their Catalan version.

For instance, when searching for “Merce Rodoreda” without an accent mark, the first result is an article about the writer on Wikipedia, but both Google and Bing lead to the Spanish version and only show the Wikipedia article in Catalan as a second option. Other searches, including those regarding corporate websites and official bodies, occur accordingly. This also happens when users explicitly set on their device, browser, and personal account (Google and Microsoft, respectively) that they prefer to get results in Catalan first.

However, the current situation is even more startling with regard to this marginalization of Catalan, which makes the problem even more incomprehensible. Both Google and Bing show so-called snippets, data summary boxes highlighted on the right (on computers) or at the top (on mobile devices),

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>

when searching for people, companies, and place names, among others, in the language we have configured on the browser. On the other hand, if this search is done with the correct Catalan spelling (Mercè, not Merce), the first organic result on the list is the Wikipedia article in Catalan. However, some searches for Catalan terms (xucrut, i.e. sauerkraut) actually return Spanish pages as the first results not containing the term we have entered, but its translation. To continue to draw on food, some searches for pollastre (i.e. chicken) provide the term's definition in the DRAE (the Royal Spanish Academy's Dictionary)!

The problem is rather complex and depends on a range of combinations, but the general situation is that Catalan pages are less visible than they used to be. Moreover, their Spanish equivalents are prioritized over Catalan websites, which means their lower visibility entails fewer clicks. This fact has **worsened with regard** to the recent past, when technically Catalan speakers received links in Catalan, if any, as the first suggestions. However, it is mainly a problem for the future, because with every click on search results we are training the search engine's algorithm which page of those suggested we are most interested in. If the latter is not in Catalan, this version will be increasingly discarded.

Content availability is not the problem

It is important to emphasize that we are not dealing with a supply problem. There has not been any decrease in Catalan content quantity or quality. The aforementioned satisfaction about the presence of Catalan on the internet is justified. For two decades, volunteers from the WICCAC association (Independent Webmasters in Catalan within Culture and Civic Areas, a member of the Aliança) have been compiling a monthly monthly barometer² that detailing the percentage to which Catalan is used on the websites of hundreds of companies, organizations, and institutions headquartered or operating in our language area. This figure, which has been increasing from 41% in August 2002 to 66% in December 2022, varies according to the business area and has a relative value, as it is not weighted with regard to each website's traffic: the website of a local real estate agency accounts for the same data as that of a generalist digital newspaper. Nevertheless, the barometer provides a comprehensive picture of the situation and allows us to concentrate improvement efforts.

Similarly, a recent study³ by Softcatalà—another group of volunteers, also a member of the Aliança—, shows that almost 470 of the half million most popular websites on the internet have a Catalan version. Six are among the first 1,000, another seven are among the first 5,000 and another 12 among the first 10,000 sites. Some are rather predictable, such as Booking, Google, Facebook, Outlook

² <http://wiccac.cat/webscat.html>

³ <https://github.com/jordimas/crux-top-lists-catalan>

and Twitter, but there are also less known websites with high traffic featuring a Catalan version, such as Chess.com and Bible.com. It should be noted that this analysis was based on an unexpected data source: the list of addresses Google Chrome users visit and save on the browser's cache memory, which Google accumulates and aggregates on a monthly basis for public use.

This study by Softcatalà concludes that Catalan boasts a particularly strong digital presence in the public and academic sectors. It also shows that the existence of the .cat domain is a hallmark: thanks to the puntCAT Foundation's 15 years of work it has become the second most used extension for Catalan content (141 of 470 websites), just behind .com (178 websites) on a global level and outpacing it as to Catalan websites, which is far above the .es domain (40 websites).

A serious problem calling for serious attention

Given that Catalan content has been losing ground in web search results with regard to Spanish content, that many multilingual websites have detected a sudden change in visits to their Catalan versions, that this change has happened without any significant amendment to the sites' architecture, content, or quality and that many websites have gone from traffic on their Catalan versions to this being channeled to their Spanish counterparts, we believe that Catalan has a serious visibility problem on the internet and swift measures are required from web search providers to restore the previous situation.

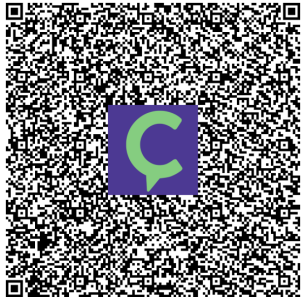
We must again note that search engine providers have started to ignore the specific language preferences each user sets in their device/operating system/browser/user account. In spite of the many elements involved in search engine rankings, we demand that search engines start taking into account user preferences, as they were doing prior to May 2022.

Examples



Apart from the situation already depicted on social networks by different users, the bias on search results must be shown in a neutral way. Thus, we have established a testing environment based on the Windows operating system, while choosing Catalan as the main preference for interfaces and browsing. The specifications of this testing environment are set out in the following table.

OS / version	Windows 10 Pro 2H22 19045.2846		
Browser / version	Chrome 113.0.5672.126		
Browser context			
ACCEPT-LANGUAGE	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		
Cookie context	No general, nor Google cookies previously set		
Geolocation context			
Public IP	141.166.99.42	Google location	43470 La Selva del Camp

Once it was ready, this environment was used to run several tests and assess their results. Three of the most relevant situations are described below:

Query literal	barcelona		
Query URL	Link Due to the URL's length, it's easier and more user-friendly to use a direct link.	URL's QR	

<p>Results screenshot (top, not promoted content)</p>			
<p># first result in Catalan</p>	<p>None in the first page of results.</p>	<p># first result in Spanish / other language</p>	<p>1</p>

<p>Query literal</p>	<p>sagrada familia</p>		
<p>Query URL</p>	<p>Link Due to the URL's length, it's easier and more user-friendly to use a direct link.</p>	<p>URL's QR</p>	
<p>Results screenshot (top, not promoted content)</p>			
<p># first result in Catalan</p>	<p>2 (sublevel)</p>	<p># first result in Spanish / other language</p>	<p>1</p>

The last situation differs slightly from the previous ones: in this case, the query forces the results in Catalan. This parameter was defined using the Google search in a specific language option from the public frontend.

Query literal	seat		
Query URL	Link Due to the URL's length, it's easier and more user-friendly to use a direct link.	URL's QR	
Results screenshot (top, not promoted content + news highlights)	 <p>The screenshot shows search results for 'seat' in Catalan. The top results are in Spanish: 'SEAT - M Automoción' and 'Encuentra los talleres de SEAT España SEAT'. Below these are news highlights from 'el Periódico' and 'EL NACIONAL.CAT', all in Spanish. A button at the bottom says 'Més notícies →'.</p>		
# first result in Catalan	10 (results) / 1 (news)	# first result in Spanish / other language	1 (results) / Not found (news)

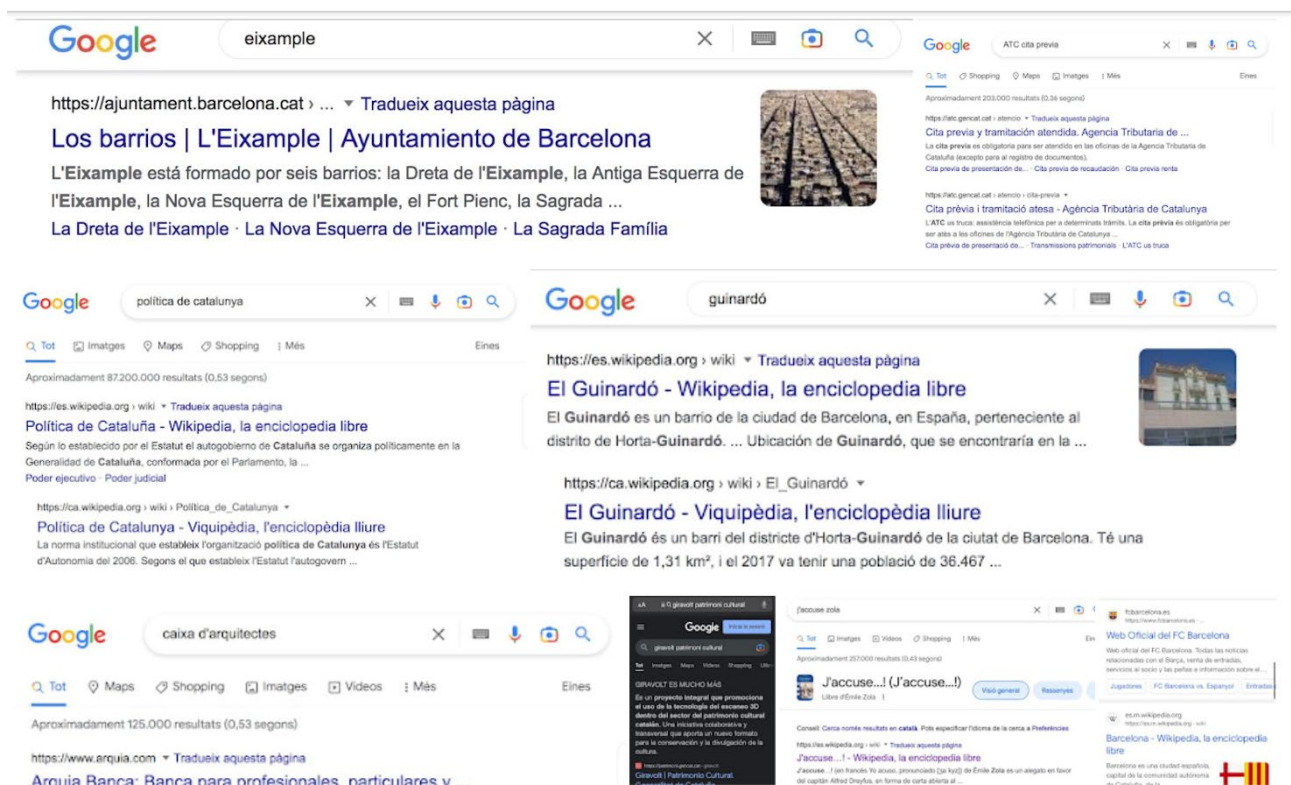
This first study stage concluded as shown in these examples: Google is ignoring user language preferences, at least as far as Catalan is concerned. Common search results prioritize contents in Spanish, even if Catalan is specified as the preferred language for results.

Tests also show that Google can tell Catalan from Spanish, as the suggested news highlights were exclusively in the preferred language, when specifically defined.

User-provided examples

As soon as the issue with the visibility of Catalan content was noticed by the public, many users posted screenshots of their search results on social networks. We have collected some of them, in no particular order, in this Google Photos album:

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



Goals of this report

The main civic organizations promoting and advocating for Catalan have formed⁴ the Aliança per la Presència Digital del Català (APDC, Alliance for the Digital Presence of Catalan). Acció Cultural del País Valencià, Amical Wikimedia, Fundació.cat, Institut d'Estudis Catalans, Institut Ramon Llull, Obra Cultural Balear, Òmnium Cultural, Plataforma per la Llengua, Softcatalà, and WICCAC have pooled their technical resources, knowledge, and rallying capacity in order to tackle the issue regarding web search results.

Through informal contacts with major web search providers, these companies have acknowledged the issue and have requested factual data on its extent in order to solve it. Members of the APDC, led by Fundació.CAT, have prepared this report, which tracks and details the loss of visibility of Catalan content in web search results over time for the involved companies to better diagnose the issue based on facts instead of perceptions.

We have listed dozens of organizations from the government, academia, media, and business sectors in the Catalan realm that publish their web content in Catalan, in order to provide us with historical traffic data from their 600+ websites (main domains and subdomains), based on organic search results (thus, excluding sponsored results). Next, we have compiled and analyzed such data to be able to quantify the growing demotion of Catalan content in web results as well as to identify and reverse any milestones related to technical and infrastructure changes.

This report will be made available to the general public and the media via the APDC website (<https://aliançadigital.cat>) and will also be provided to the major search providers, either directly or via the Regional Government of Catalonia, which has expressed its interest in helping to solve the issue and has even contacted one of the companies. Furthermore, this report is being handed to select members of the European Parliament, who are active in minority EU language matters, for it to be used in their legislative initiatives.

We believe this report comes at a critical time, as the emergence of AI chatbots (such as ChatGPT, Bing Chat and Google Bard) are poised to change the way users search and interact with digital content, whereas such chatbots seem to draw mostly on content in major languages (as recognized by OpenAI), even if they are more than able to interact with users in many languages, including Catalan. Thus, the appropriate presence of original content in such non-major languages must be restored before chatbots take hold of regular web search.

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

Methodology

In order to diagnose the problem with Catalan results in web search results and help to solve it, we have applied a two-level approach. On the one hand, we have tried to assess the actual demotion in Catalan positioning as compared to the previous situation. On the other hand, we have asked several experts to provide their opinion regarding the likely causes of such demotion, in order for any involved partners to not have to start from scratch and explore these possibilities themselves.

Annex 1 includes a copy of the technical specifications provided to all contributors that were willing to supply their web traffic data for the purposes of this report.

General characterization of Contributors

This report sets out the analysis of the web traffic data provided by 13 contributors, hence covering an aggregated set of 639 domains and subdomains, for which data has been provided to us either individually or in an aggregated manner.

Such contributors have been selected based not only on technically required criteria (site multilingualism, comparison, and so forth), but also on relevance. These are important websites in terms of number of visits, some of them operated by Tier 1 public institutions.

These contributors have been classified into three groups:

- **Group 1.** Web sites affected by the change of behavior of Google's ranking algorithm. This group includes a set of 8 Contributors, spanning 425 domains/subdomains.
- **Group 2.** Web sites not affected by the above change. Sites in this group have not been affected by the change in ranking algorithm, so they keep their previous behavior and trends without any noticeable change. This group includes 3 Contributors with 212 domains/subdomains.
- **Group 3.** This group contains web sites whose data cannot be analyzed, due to their behavior or technical structure. It is made of 2 Contributors, representing 2 domains/subdomains.

All data provided by contributors has been normalized to a format with relative weight by language and time period (monthly), as this is the preferred format by contributors choosing the most restrictive structure.

What is more, the contributors' websites are using different TLDs (Top-Level Domains), so the analyzed websites may end in .com, .cat, .org and .es, among others.

Regarding web search engines

All data below refers to site visits provided specifically by organic (unpaid) results of Google's web search engine to the contributors' websites. We have focused our research on this case, as it is the one that has caused the current issue due to a change in its usual behavior. Google has been—and still is—the dominant web search engine in the Spanish market, accounting for a market share of 95% or above during the period under review.

We have also confirmed with each contributor that no major changes in the website architecture and/or content have been made which could have directly affected the site indexing and positioning on search engines during the period under review. Most of them have been actively working in these areas with no major changes in strategy or resources.

We might repeat this very process with other search engines in the coming months.

Aggregated results

As we will see later, traffic provided by Google organic search results to each web site participating in this study has been analyzed separately.

However, in order to have an overall view of the evolution of visits landing on Catalan- or Spanish-language web content coming from Google searches, we have also calculated the ratio between those arriving to each language, month by month and for each participating web site. Then we have normalized such ratio based on the starting value of the time series, which we set at January 2021. Finally, we have averaged the ratios of the different web sites (yellow line) and we have accumulated such changes over time (red line).

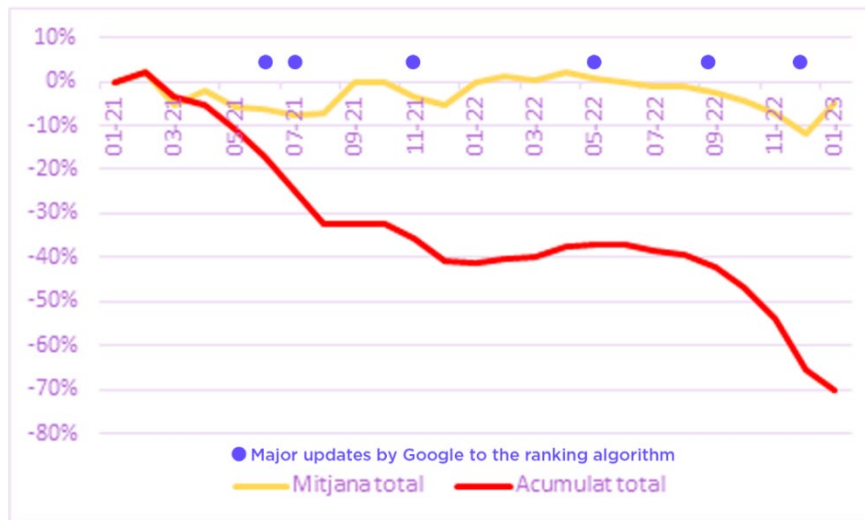
We have done this only with web sites impacted by the loss of traffic to their Catalan-language version, which is the only relevant situation in the context of this study.

Given that not all participating web sites have provided their absolute traffic data, we have been unable to weigh the results based on their volume. Due to this, we have differentiated two situations. The first one considers that each participating web site has the same weight. The second one considers just the web sites that we have confirmed to get more than 500.000 sessions over the whole time series. In both charts we have noted major changes in Google Search ranking algorithm, as disclosed by Google itself⁵

Loss of traffic by Catalan content in favor of Spanish content in the full set of analyzed web sites

The following chart shows how the ratio of visits to Catalan- and Spanish-language content has evolved over time in our full set of web sites.

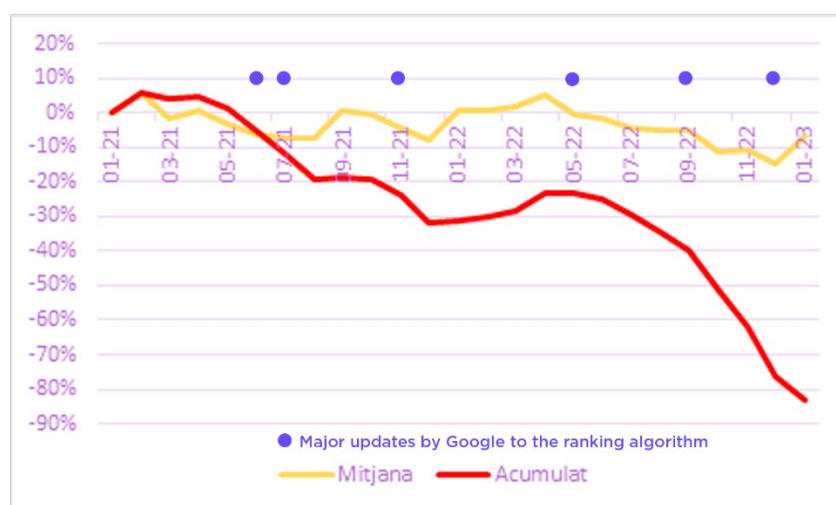
⁵ <https://status.search.google.com/products/rGHU1u87FJnkP6W2GwMi/history>



After a short initial period apparently showing growth in the traffic to the Catalan version of the impacted web sites, the situation suddenly turns into a downward trend, which for 12 months remains mostly steady around 40%. From then on, between the end of Summer 2022 and the beginning of Autumn 2022, it starts falling again strongly until reaching a 70% loss at the beginning of 2023.

Loss of traffic by Catalan content in favor of Spanish content in the most visited web sites

On the other hand, when discarding the less visited web sites in our data set and considering only the ones getting more traffic (more than 40 million sessions in one case), the ratio between the Catalan- and the Spanish-language content evolves as shown in the following chart:



The loss of visits coming from Google searches by Catalan-language content in the most-visited web sites is even more noticeable than in the previous view (full set of analyzed web sites): the initial reinforcement of Catalan-language content remains for a few more months, but from May 2021 there is an initial fall to around 20% of loss of visits to Catalan content. This situation remains steady, with minor variations, until April - May 2022, when a sustained fall starts until exceeding 80% of lost visits on an accumulated basis.

Even leaving out the first part of the time series and focusing on the last half of the year 2022, month-to-month losses are sustained in both views, which causes a critical loss of traffic to Catalan-language content.

Site-specific results

Language behavior of analyzed websites

Every single website covered by this analysis is multilingual, hence **all of them provide content in more than one language** to their visitors. This means that these websites feature their user interface and at least part of their content in Catalan and Spanish, whereas some of them may include other languages, such as Occitan, English, or French, among others.

Given that many multilingual websites are not symmetrical, i.e. they do not offer the same content in every language included in their interface, in case of doubt we have chosen those that display the closest correspondence in terms of content.

While focusing on the relative behavior of Catalan and Spanish content, we found that **the set of cases show a negative correlation**, which is more or less strong, between both languages:

	ca – es Correlation	ca - en* Correlation
Contributor 1	-0,98	N/A
Contributor 2	No valorable	No valorable
Contributor 3	-0,52	-0,41
Contributor 4	-0,41	-0,02
Contributor 5	-0,99	-0,67
Contributor 6	-0,87	-0,17
Contributor 7	No valorable	No valorable
Contributor 8	-0,85	-0,42
Contributor 9	-0,96	-0,23
Contributor 10	-0,71	-0,46
Contributor 11	-0,97	-0,24
Contributor 12	-0,55	0,29
Contributor 13	-1,00	N/A
Average	-0,80	-0,25
Max	-1,00	-0,67
Min	-0,41	0,29

* Not all Contributors' websites are available in English.

This strong negative correlation means that an increase in visits to the website in Spanish entails a loss of visits to the website in Catalan. This validates the perception already stated by several webmasters. Thus, not only are visits to Catalan content decreasing, but many visits that were previously aimed at Catalan are now directed towards Spanish content.

Furthermore, this ratio tends to -0,80, which is a very strong correlation: for each visit to the Spanish version, the Catalan version loses almost one visit.

On the other hand, when compared to the correlation between Catalan and English, the average ratio is just -0,25, which even becomes neutral or positive in some cases (contributors 4 and 12, respectively). This average ratio means that a single visit to the Catalan version of the website is lost for every four visits to the English version.

Meaning in the context of this report

On a practical level, these results show that search traffic coming from Google is based on a nearly exclusive relation between Catalan and Spanish: every visit to the Spanish section of the website equals a lost visit to the Catalan section, so a loss of traffic in the Catalan content means a growth of traffic in the Spanish one.

Therefore, by prioritizing Spanish content over Catalan content in search results on Google, even when the keywords used are Catalan common terms, has caused the websites behind the links to experience a change in their visits' profile, with subsequent changes in the usage trend of the languages offered.

We have thus confirmed that changes made by Google in the search results have had a significant impact on many analyzed websites.

We need to emphasize that **66,5% of the analyzed websites have been impacted**, so the effect is extremely relevant given this volume of websites. We are a bit surprised, though, that not every website has been affected.

Finally, the charts below are displaying a clear coincidence in time of such incidents in the form of a trend interruption/break. This would lead us to think that the issue occurred due to some sudden change in Google's indexing. All trends would not happen at the same time if they were caused by internal website changes.

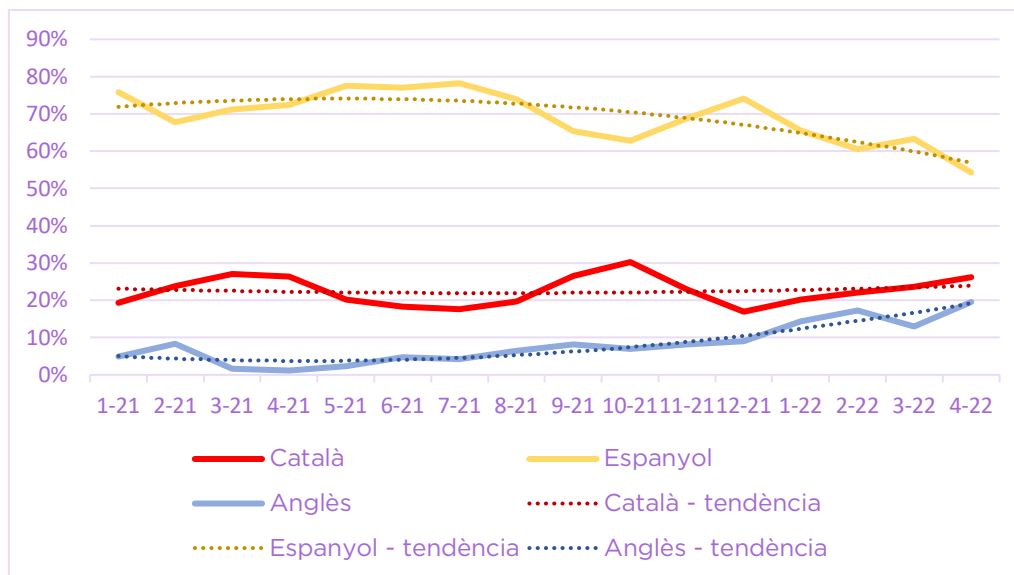
Group 1. Sites that were impacted

In order to illustrate this issue, below we are reviewing four of the analyzed situations.

Contributor 3

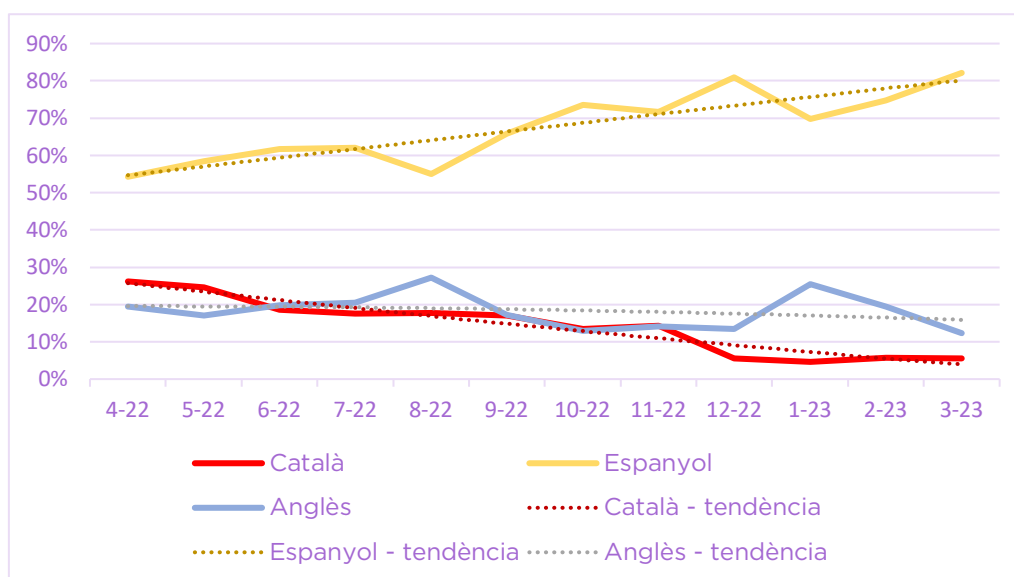
Contributor Card			
Description	This first case is an organization with a dual focus, local and international, communicating actively in several languages as an indispensable element of its activity.		
Ca - es Correlation	-0,52	Ca - en Correlation	-0,41
Has hreflang?	Yes, no errors. Possible improvement: adding x-default.		
Full Series	<p>The chart displays percentage trends for three languages: Català (red), Espanyol (yellow), and Anglès (blue). It compares current data (solid lines) with trends (dotted lines) from January 2021 to March 2023. The y-axis ranges from 0% to 90%. Espanyol consistently has the highest percentage, starting around 75% and ending near 80%. Català and Anglès have lower percentages, generally between 10% and 30%. The legend indicates: Català (solid red), Espanyol (solid yellow), Anglès (solid blue), Català - tendència (dotted red), Espanyol - tendència (dotted yellow), and Anglès - tendència (dotted blue).</p>		

The first chart shows the situation *before* the traffic issue on Google was detected:



As can be easily seen, there is a clear 16-month trend, where English grows, while Catalan remains stable, with brief variations around 10%. Meanwhile, Spanish shows a clearly downward trend, from 75% of visits to 55% at the end of the period.

During this period, we also see a peak of visits to the Catalan version in October 2021, as well as the minimum difference between the Catalan and Spanish versions: 28% in April 2022.

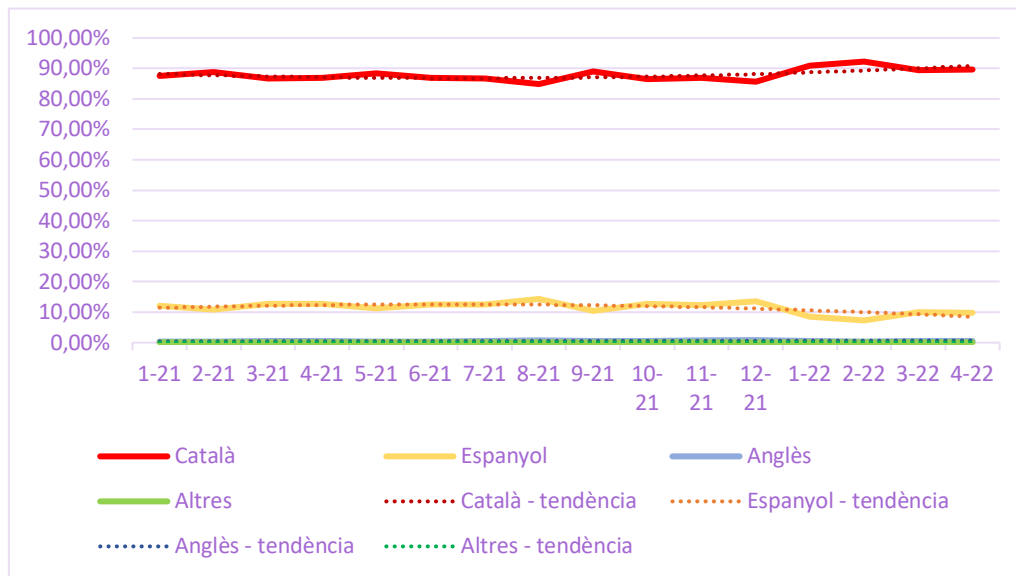


From then on, the series continues as of April 2022, where we can see a strong change in trend: Spanish starts to grow quickly until reaching its peak: above 80% at the end of the series. During the same period, Catalan is no longer stable and reaches its minimum, somewhat below 5% of traffic, in January 2023. This is also the moment of the biggest difference with Spanish, more than 75% in December 2022. Actually, a change in trend also happens with English, from an upsurge to a slow loss.

Contributor 5

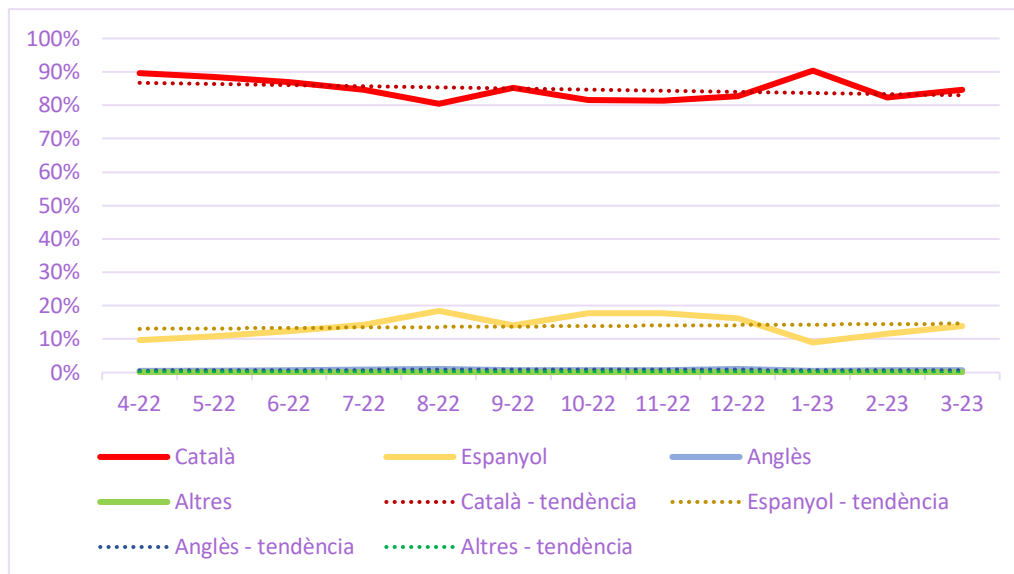
Contributor Card			
Description	Our second reviewed case is a specific website of a public administration focused mainly on the Catalan-speaking audience. However, it also provides content in Spanish and English, as well as a specific regional case.		
Ca - es Correlation	-0,99	Ca - en Correlation	-0,67
Has hreflang?	Yes, but the reference between languages is not well configured. Possible improvements: add reference and x-default.		
Full Series			

The chart below displays the year 2021 and the first part of 2022, up to the moment when a change in trend is detected:



As expected from this website, Catalan features a volume of visits that is much higher than that of other languages, hence being the main one by far. During this period, Catalan shows slow but sustained growth, probably reaching its maximum threshold, while the other languages remain stable or decrease slowly, as is the case with Spanish.

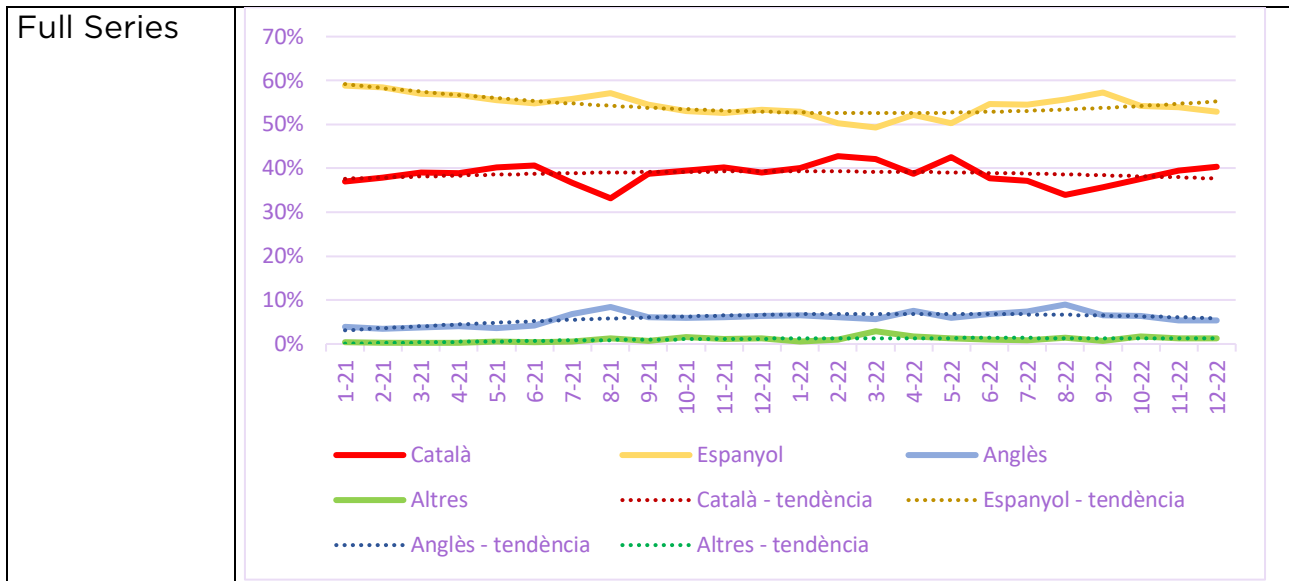
Maximum visits to Catalan content coming from search engines occur in February 2022, with more than 92% of total traffic, accounting for the maximum difference between Catalan and Spanish: almost 85%. Spanish reaches its peak level during this first period in August 2021, with slightly over 14% of visitors.



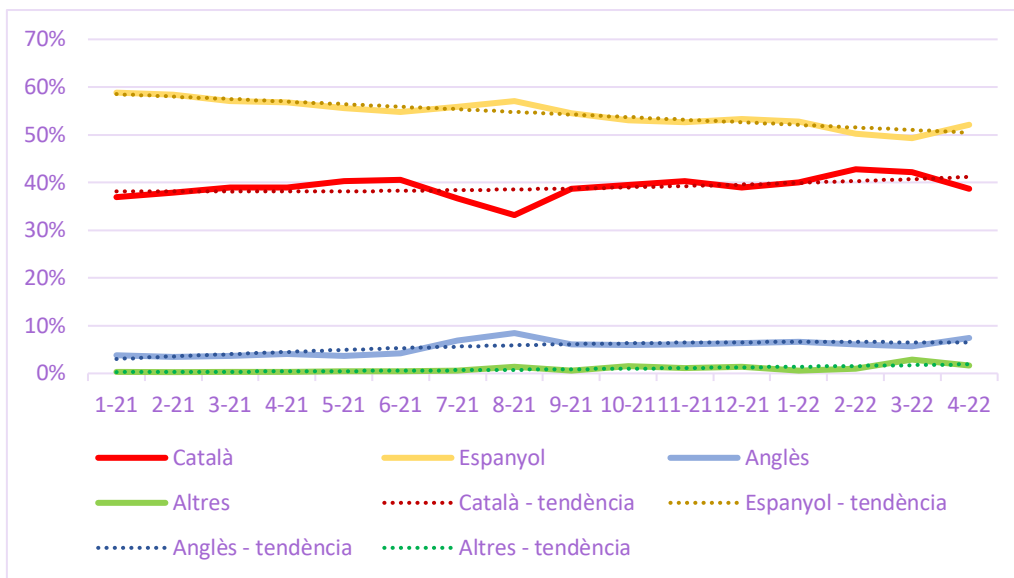
From April 2022 on the trend changes, especially with regard to the Catalan to Spanish ratio. Visits to content in Spanish increase until exceeding 18% in August 2022, its peak level within the series. At the same time, we notice a minimum distance with regard to Catalan, which stands at just 62%. This is not a one-off increase, as Spanish remains stable at this peak level and the growth trend persists in the year 2023, with a single decrease in January 2023.

Contributor 10

Contributor Card			
Description	This is a public institution with a large number of websites that have been collected as an aggregate set. The multilingual aspect is required both by the diversity of the target population and, to a lesser extent, by tourism.		
Ca - es Correlation	-0,71	Ca - en Correlation	-0,46
Has hreflang?	25 percent of sites have it and well formed. The rest varies but we are considering they don't have it.		

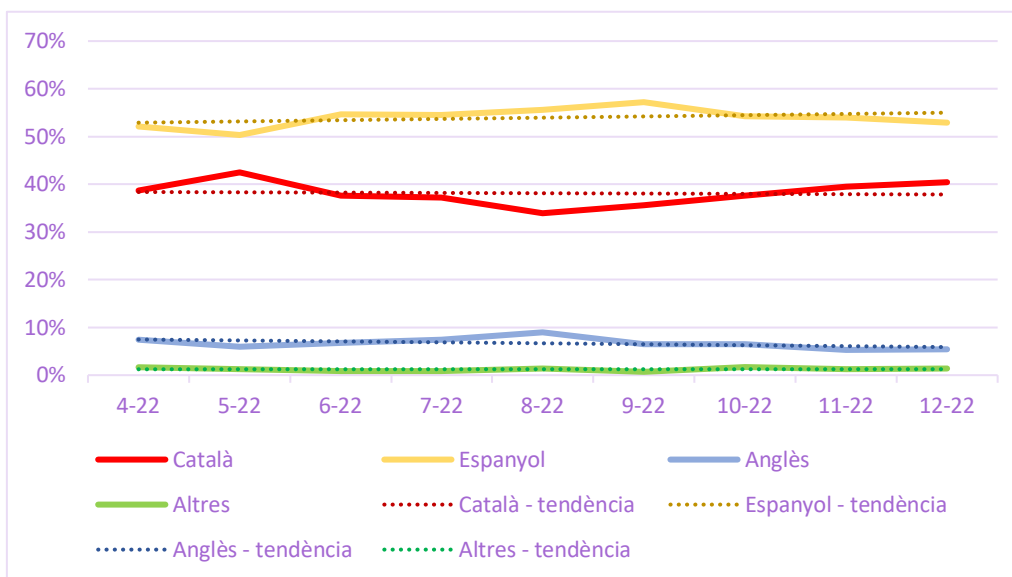


Following the same model that in the previous case, our first chart displays how Google-originated traffic has evolved across the set of websites provided to us:



In this case, Spanish content also shows a decreasing trend, while Catalan features the strongest growth. However, English, and to a lesser extent the other languages available, also display a slower growth trend. We must emphasize that both Spanish and English show a sudden growth in the months of July and August, with a decrease of Catalan.

The maximum level of traffic towards Catalan pages happens within this period: February 2022, almost reaching 43% of the visits. The minimum difference between visits to Catalan and Spanish pages happens also within this period: somewhat around 7% in March 2022.



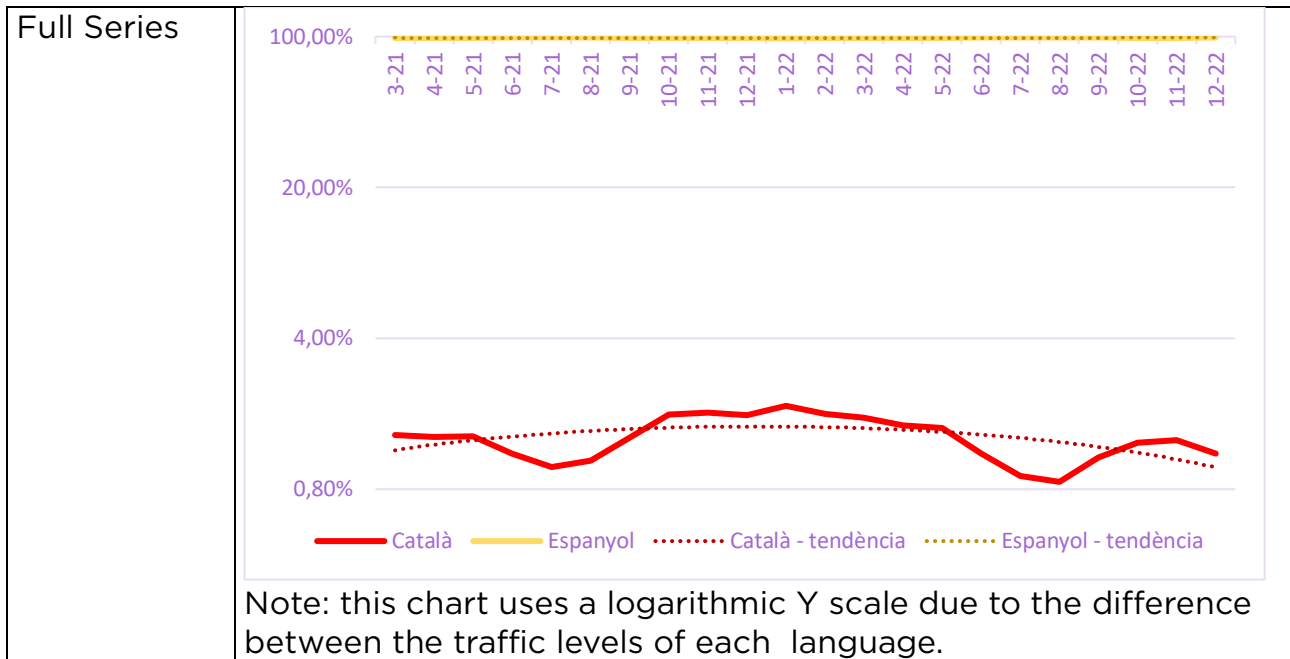
Now on to the next period. Once again, trends change: the decrease of Spanish slows down, while all other languages stop growing and start decreasing. In terms of Catalan with regard to Spanish, the same drastic change that happened in the previous summer now becomes a slump from which Catalan does not recover even at the end of the year.

In figures, Catalan starts from a peak level of almost 43% of traffic in February and goes down to under 34% in August. On the other hand, Spanish grows more than 57% in September 2022, thus almost reaching its own peak level in the full series of 59%.

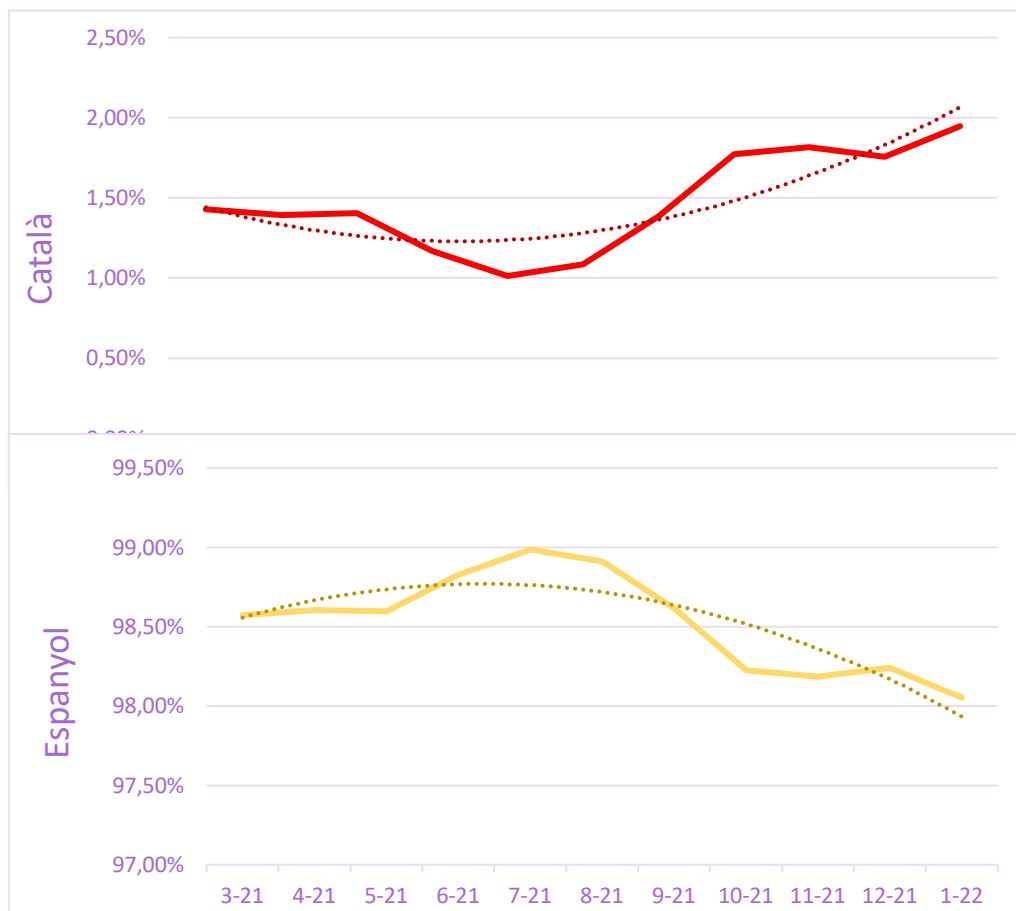
This Contributor has told us that the trend during 2023 has been unstable so far, and we look forward to getting updated data from them in order to confirm this.

Contributor 13

Contributor Card			
Description	Worldwide organization that provides its traffic data for Catalan and Spanish. It operates one of the most visited sites on the internet, serving all kinds of user profiles. Several search engines are quoting its content directly.		
Ca - es Correlation	-1,00	Ca - en Correlation	N/A
Has hreflang?	No, but it has a specific treatment for being indexed correctly.		

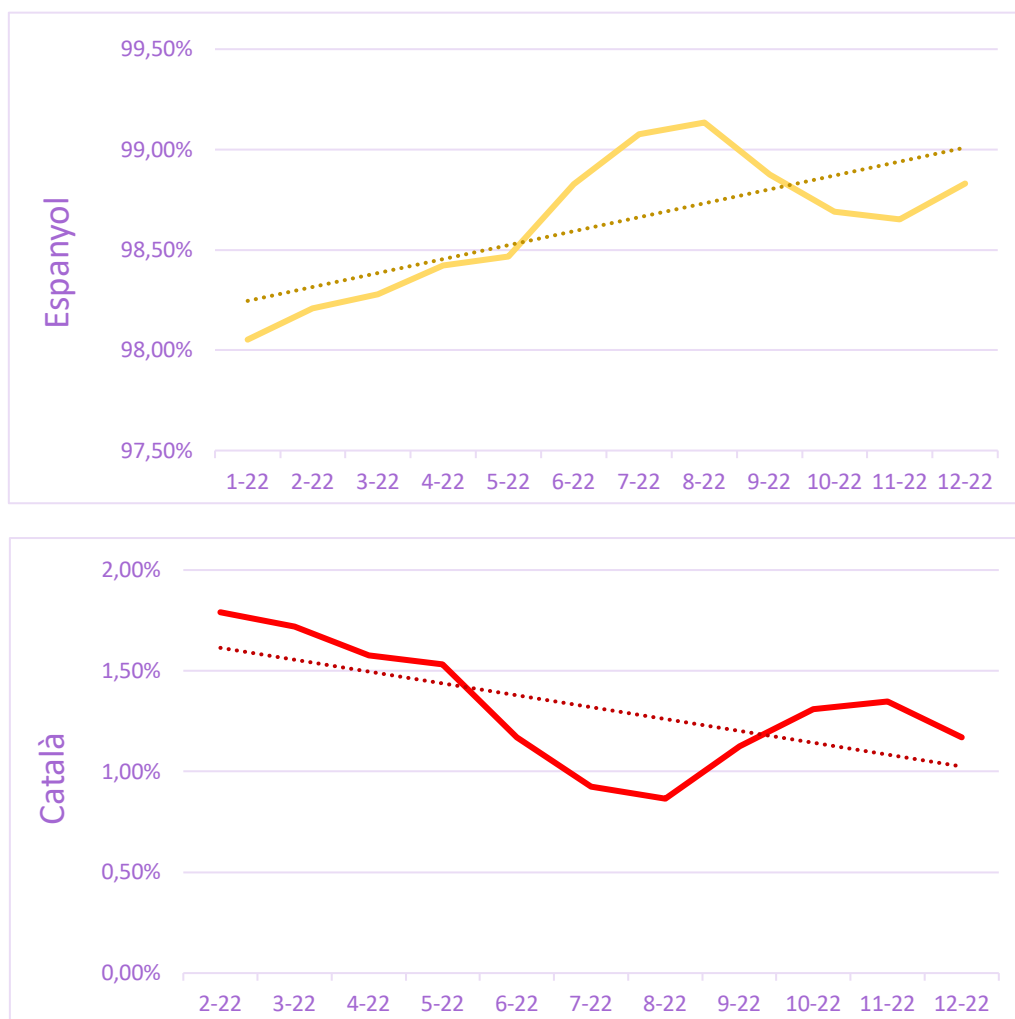


This case describes one of the strongest impacts within those analyzed. As a global website, there is a huge difference between the number of users in Catalan and in Spanish, so we are providing separate charts for each language, hence covering the first part of the time series.



We start by emphasizing that, probably due to the its singularity, this site seems to be getting a special treatment by Google. Thus, the variations in search-originated traffic start a bit earlier, between January and February 2022, instead of March and April 2022 as in the other cases.

During the first period (above charts), Catalan reaches a maximum close to 2 percent in January 2022. At the same time, we see the smallest difference between the two languages, a bit above 96 percent. There is a clear growth trend of Catalan and a decrease of Spanish, both of them slow. Keep in mind that these charts display a significant enlargement to the series.



On the other hand, during the second part of the series, Catalan content gets its minimum traffic, under 0.9%, while Spanish content achieves its peak level, above 99%. Even if we discard the discernible seasonality cycle, the minimum level of Catalan in August 2022, usually a low point in time, is still 0,2% below the

one in 2021. Moreover, in the same month of 2022, we see the greatest distance between Spanish and Catalan.

In terms of trends, even considering oscillations, Catalan content takes a steeper downward path than the previous increase, while Spanish surges, reaching its peak level.

Group 2. Sites that were not impacted

Contributor 8

Contributor Card			
Description	Organization focused on promoting and advocating for Catalan culture and language. Its website has a dual purpose: promoting the organization's own activities in Catalan and spreading the daily reality abroad by providing content also in other languages.		
Ca - es Correlation	-0,85	Ca - en Correlation	-0,42
Has hreflang?	Partially. Languages and variables are defined, but return links are missing.		
Full Series			

As shown in the above full-series chart, the peak level of visits to Catalan content coming from search engines happens when Google has been applying a language bias for months. In December 2022, almost 97% of visits coming from this search engine are to Catalan sections. Meanwhile, Spanish reaches its peak in August 2021, whereas English does so in August 2022. Thus, we believe this website has not been affected by variations in visits coming from Google search results.

The outliers: impacted websites that have applied countermeasures

The issue, which has already lasted over a year since its onset, has caused the impacted sites to notice sustained changes in the profile of their visits from Google.

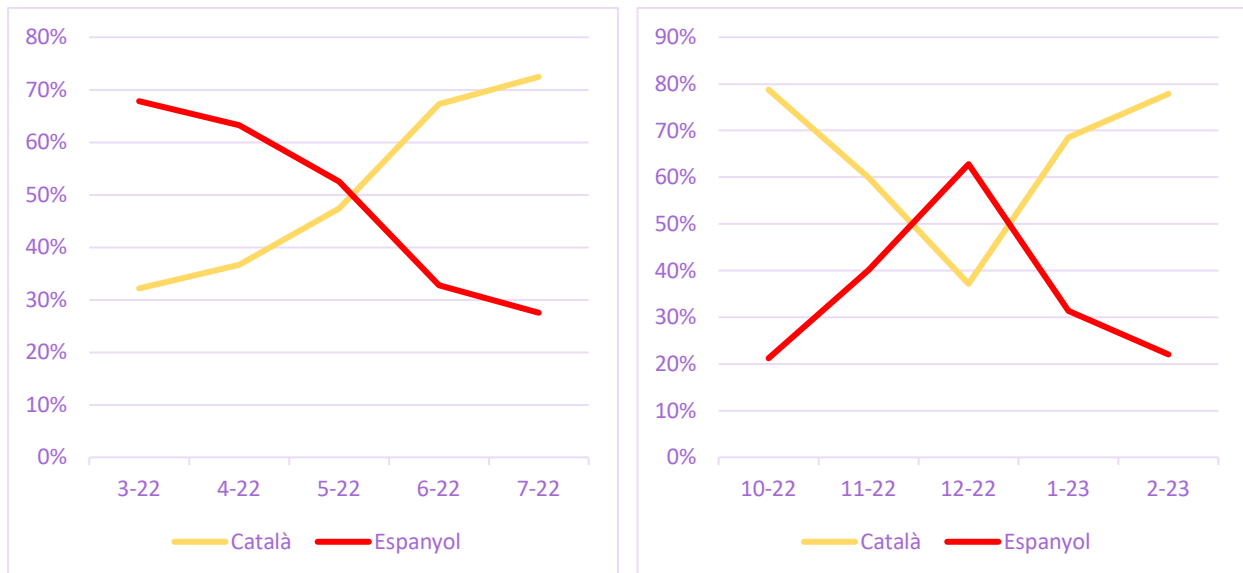
Beyond criticism regarding some information websites by users who now retrieve Spanish content, where they previously found Catalan content, in some cases the issue could harm the business model or a major brand component.

Owing to this, we have encountered situations where the website owners have taken countermeasures in order to prevent the issue from causing even more trouble. Below we describe one of these cases.

Contributor 1

Contributor Card			
Description	50-year old Catalan business operating an e-commerce platform. They have only sites in Catalan and Spanish because they do not sell overseas. At the same time, this makes them more sensitive to any changes in behavior by their target audience.		
Ca - es Correlation	-0,98	Ca - en Correlation	N/A
Has hreflang?	Yes, using the right format and links.		
Full Series	No specific value apart from the effect of the countermeasures.		

In this case, there is a historical trend to content being indexed while giving priority to Spanish, even if the company is actively prioritizing Catalan content on the website. Based on this, the situation became untenable in May 2022, so they decided to de-index all Spanish content on the website, thus immediately encountering a change in behavior.



After a first de-indexing operation, Catalan traffic seemed to stabilize for a few months, until it suddenly started to go down in November 2022. This required the previous action to be repeated by forcibly de-indexing the Spanish content in December 2022.

Preliminary Conclusions

The impact is not general

Our website analysis shows that not all sites are impacted in the same way, regardless of the main language they use for their target audience, as shown in the previous examples. However, such impact is present in most cases, i.e. around two thirds of the websites we have been able to analyze.

The strength of the impact varies

In the same way that the impact is not general, there are cases where variation in traffic volume coming from search engines is lower than in other cases where such variation ends up changing the user profile via this path.

No link to domain (TLD)

TLD refers to the top level domain used by the website in its main name. The impact happens in .com, .org, .cat and .es websites, thus we believe the domain authority is not a relevant factor in this issue.

There is an inverse relationship between Catalan and Spanish

As suspected when the Aliança Digital became public, the increase in traffic of Spanish content coming from Google entails a decrease in traffic of Catalan. This trend has been confirmed in all analyzed cases.

One of the contributors helps to confirm the above: as soon as they forcibly de-index their Spanish-language content (an extreme countermeasure which they apply for business reasons), the traffic to their Catalan-language content recovers its full visibility in web searches. When the Spanish version gets indexed again, the Catalan one drops again.

Why is this happening? Some hypotheses

In addition to the analysis of search-originated web traffic data from contributors, we have asked several experts to brainstorm on the reasons that could be affecting the visibility of Catalan content in search results. This exercise is meant to help web search providers to discard some possible lines in their research of the issue. Most of them refer to Google, because it is the prevailing search engine (95,9% market share in Spain in April 2023, according to StatCounter GlobalStats).

Political reasons

When the loss of visibility of Catalan in web searches became evident, many users on social media attributed this to some deliberate intent by Google to penalize any Catalan content. This theory has been driven by the lack of public or institutional acknowledgement or support by any representative of Google Spain.

However, we discard any animosity of Google against Catalan as many of its consumer products and services are already available in Catalan.

Click feedback

Catalan-speaking web users generally click on Catalan and Spanish results indistinctly. Thus, web content in Spanish tends to get more clicks than web content in Catalan, hence Google decides that Spanish content is more relevant because it gets more clicks.

The problem here is that Google is not complying with the users' preferences ("I'd rather like to see pages in Catalan"), so when a content is available in several languages, as in multilingual sites, Google discards the language preference in the browser or user profile and gives more relevance to the site with more visits, which ends up getting even more visits.

Language encoding

It has been suggested that the internet and the WWW powers-that-be (ICANN, IETF, W3C) have applied some change in content language encoding that has caught search engines off-guard.

But, apparently, such powers-that-be are not applying this sort of changes. They made recommendations that search engines and web browsers can follow, but those don't cover in which order a search engine must display its results. For

instance, the W3C defines labels with the hreflang and lang attributes that are used to link pages in several languages on the same website. Google is taking these labels into account, but it prefers to detect the language by itself, not considering the ISO-code specified in the labels:

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

Other than that, we are not sure that the issue is also affecting other languages. Actually a few complaints about Ukrainian have surfaced on Twitter.

Google mistakes Catalan for Spanish

It has been said that Google could mistakenly identify Catalan content as Spanish, maybe due to some change in the AI engine that detects languages.

We discard this as Google is not confusing Catalan and Spanish, It is even able to provide results in Catalan if forced to do so. What is more, Google offers the option (which requires an additional click) of searching specifically for Catalan content in the top toolbar.

Discrepancy between organic results and snippets

The same search produces a different output in the organic results and the right (top on mobile) content snippet: organic results penalize Catalan, whereas snippets respect user preferences. Sometimes snippets even seem to be delayed: at the time of testing, while searching for *GSMA*, the snippet showed Stéphane Richard as the Association's president, whereas the linked Wikipedia article already mentions José María Álvarez Pallete as the current president.

This seems to be a dead end, because Wikidata is not the only source for snippets, which may also draw on official websites and other repositories, such as IMDb, FilmAffinity and others.

The important thing here is that many snippets are displaying text in Catalan, but not because they take the information from a Catalan source: they are actually taking it from an English site and translating it into Catalan, thus giving a false impression of normality. This impression is even stronger on mobile devices, because the snippet/infobox displays on top of the actual search results.

AdWords issue

A SEO/SEM marketing specialist emphasizes the weird, as yet unexplained case (<https://twitter.com/EvaOlivaresb/status/1618646946713055232>), where an advertiser buys keywords in Catalan, but users are presented with search results in Spanish.

With regard to the above, some have raised the hypothesis that Google is focusing on major languages to help boost its advertising business, while applying a combination of positioning and scale to prioritize Spanish in an area that is a relevant market for the company.

Core Update

Google updated its search core in May 2022, right when the issue started. As the company says, these kinds of updates change the search behavior substantially. According to several SEO websites, this core update included changes in the algorithm's guiding principles, hence promoting the EEAT (Experience, Expertise, Authoritativeness, and Trustworthiness) strategy. The main change happens in the second 'E' (Experience). How this EEAT affects searches can be seen in Google's Guidelines for Quality Raters (<https://services.google.com/fh/files/misc/hsw-sqrg.pdf>).

Several possible situations have been identified:

- The May 2022 update modified something related to “regional” languages that went unnoticed.
- Quality Raters are doing something wrong with Catalan content
- Some other variable we have not detected

Actually, the above Guidelines mention specifically Catalan (page 137) in the section about labeling content in “foreign languages”:

For example, most Catalan-speaking users in Spain also speak Spanish. Therefore, for rating tasks in Catalan (ES), the Foreign Language flag should NOT be assigned to landing pages in Catalan, Spanish, or English

Could this be a hint as to is a wrong understanding of Catalan/Spanish as a foreign language, so the update does not work well in areas with more than one vehicular language?

Geolocation

A user with its environment configured in Catalan makes a search while in France. The first organic result comes up in French and the second in Catalan: <https://twitter.com/jordimash/status/1633522801968570394?s=20>

This could point to a prevalence of the state (due to advertising reasons?). However, we discard this option because the same search, also in France but with misspelled keywords gets the first result in Spanish, just as if the same search is made in Catalonia:

<https://twitter.com/jordimash/status/1633525602174001160?s=20>

Next Steps

Adding more Contributors

The above analysis is based on the data provided by the 13 contributors that managed to meet our deadline. However, several other organizations have provided their valuable data after our deadline, so those could not be included in this report. We plan to study the latter and add it to an updated report, should it be relevant. In any case, such additional contributors will be involved in further analyses.

Continued monitoring

We expect this report to be forwarded to the companies operating web search engines, so they can use our findings to diagnose and solve the problem as soon as possible. Meanwhile, we will continue monitoring the behavior of major search engines regarding Catalan content, while adding new data sources and reporting back to involved parties in case no improvement is seen within a reasonable timeframe.

To do so, the APDC will commission the development and deployment of an ongoing generic, geographically distributed monitoring service that is able to provide early warnings of any future worsening of the visibility of Catalan content in web search results. Softcatalà, a member of the Aliança, has already developed a prototype of said system, which might be moved on to production (see partial screenshot below).

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	ca	es	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	ca	es	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	ca	es	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	ca	es	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	ca	es	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	ca	es	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	ca	es	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	ca	es	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	ca	es	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

Moreover, a plugin for major web browsers will be created and made available to the public, so internet users can automatically provide anonymous data about the search results they are getting in their daily browsing with regard to the language configuration of their OS/browser/profile/device and the keywords they entered. This crowdsourced data will then be used to enhance the above dashboard.

Further studies related to the hreflang parameter

One hypothesis that has arisen several times, even within discussion threads on social media, links the issue to Google's handling of the *hreflang* parameter.

hreflang is a HTML attribute which is used to specify in which languages a website is available. It is not a selector for visitors to choose in which language they want to display the website, but an information for the indexing algorithms crawling the website to be aware of the existence of those languages.

The analyzed web sites do not display a clear influence of the *hreflang* parameter: some of the impacted sites include *hreflang* with the required configuration, while some of the non-impacted sites are not using *hreflang* well.

However, based on this hypothesis, we are commissioned a SEO and web positioning consultancy to investigate if the *hreflang* parameter can have influenced the issue with Catalan and to what extent. The study will use ten websites—half of them with the *hreflang* parameter set and the other half without it—, in order to track the evolution of traffic coming from Google over the last 16 months.

The resulting data will be correlated to different items to assess the level of influence of each one. The base data are extracted from the contributors' Google Analytics and Google Search Console.

Moreover, another experiment will be carried out with another five websites that have not used *hreflang* yet, but will rely on our support to apply the parameter, so we can assess how their ranking evolves in search results.

We are currently assessing which contributors may be most helpful in this study. Only a few among the ones involved at the first stage meet the new requirements, so we will need to enroll new ones.

Credits

- Report Authors: Albert Cuesta (albertcuesta@fundacio.cat) , Pep Masoliver (jmasoliver@fundacio.cat)
- Technical Management, Data Processing: Pep Masoliver
- Data Acquisition: Griselda Casadellà (gcasadella@fundacio.cat)

Members of the Aliança per la Presència Digital del Català

- Acció Cultural del País Valencià: Anna Oliver
- Amical Wikimedia: Robert Garrigós, Xavier Dengra
- Fundació .cat: Genís Roca, 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, Pere Orga
- WICCAC: Joan Soler

Contributors

The Aliança per la Presència Digital del Català and the report authors wish to thank each and every organization that answered our call to provide traffic data from their websites in order to process it for this report.

A few of them have expressly agreed to be mentioned here, namely:

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



We are not publishing any other names here in order to preserve their privacy, as stated in the Non-Disclosure Agreement (NDA) shared with each of them. However, such information could be disclosed to one or more web search providers, should they require it to analyze a specific case more in -depth. This disclosure will only be made on a case-by-case basis and only if the Contributor has agreed to it, as is also stated in the NDA.

Annex 1. Technical Specifications

Which data is required?

Data collection summary file

Type of website Multilingual websites

Observable traffic Organic search traffic

Frequency Daily

Format CSV

Data to be provided Visits by search engine and target site (direct from Google Analytics).
Visits by search engine and language relating to 100.

Period 1/1/2021 - present

Multilingual websites

The incident detected in search engines primarily affects multilingual websites, which means that the content version in a language other than that preferred by a user is positioned higher. This is why data collection focuses on this type of website.

Furthermore, a control group of Catalan monolingual websites is monitored to discover the trend of internet users with regard to this language preference.

Organic search traffic

In terms of traffic the website receives, our interest is focused on organic search traffic, i.e. traffic resulting from the natural site indexing, while ruling out all sponsored links or promotions of any kind.

It is important to take into account and include all search engines (Google, Bing, DuckDuckGo, etc.) when exporting data. Furthermore, where the website offers its content in more than two languages, it is important to bear this in mind: apart from the visits to the Spanish versions of these sites, this situation might also have caused visits to be transferred to other languages to a lesser extent.

Daily

Apart from certifying the linguistic bias that may exist in search engine results, our positioning study seeks to detect the time when negative changes were applied to search engines. Thus, we hope to receive the data on a daily or weekly basis at most, in order to be able to attach a time to the changes in behavior.

CSV

The preferred format for the export is CSV, i.e. comma separated values, as this is a standard that is available in most software (e.g. In Google Analytics), while providing a structure that is very easy to work with.

Where it is not possible to provide the data in this format, please use an editable alternative that is easy to import. Data provided as an image will be ruled out, due to difficulties in importing.

In order to simplify the extraction task, the possibility of creating a view of the software used to monitor web traffic (such as Google Analytics) and sharing access with the inbox posicionament@fundacio.cat is also offered. This will be used to download the data automatically whenever necessary.

2 possibilities from which to choose:

Given the sensitive nature of processing traffic data, two possibilities for collaboration are offered. In both cases, the data will be processed confidentially, as described below, so that the metrics cannot be associated with the corresponding data subject without their explicit consent.

One of the following three possibilities must be chosen:

ABSOLUTE VALUES

A Visits by search engine and target/landing site, i.e. the absolute value of the organic visits received by a specific site from a specific search engine (the language does not have to be specified as it is detected automatically). In this case, a description of how to export it for Google Analytics users is included as an annex to this document.

RELATIVE VALUES

Relative visits by search engine and language, which consists of preparing the data on the visits and converting them into relative to hide the absolute values. For example, by taking this table of absolute data:

	Search engine	Catalan	Spanish	Other	Total
01/01/2021	Google	70	240	20	330
	Bing	20	35	15	70
	DuckDuckGo	5	10	2	17
02/01/2021	Google	100	250	35	385
	Bing	30	55	20	105
	DuckDuckGo	8	13	4	25

B

The equivalent relative version therefore corresponds to:

	Search engine	Catalan	Spanish	Other	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
02/01/2021	Google	26.0%	64.9%	9.1%	100
	Bing	28.6%	52.4%	19.0%	100
	DuckDuckGo	32.0%	52.0%	16.0%	100

The refining of this data depends on the languages included in each case and, therefore, a generalised calculation guide cannot be made.

RELATIVE VALUES WITH TRAFFIC VARIATIONS

C Visits by search engine and language, giving the total number of visits on the first day a reference value of 100. The total relative value varies upwards or downwards, depending on the evolution of traffic for each day of the period in relation to the first.

For example, by taking this table of absolute data:

	Search engine	Catalan	Spanish	Other	Total
01/01/2021	Google	70	240	20	330
	Bing	20	35	15	70
	DuckDuckGo	5	10	2	17
02/01/2021	Google	100	250	35	385
	Bing	30	55	20	105
	DuckDuckGo	8	13	4	25

The equivalent relative version, considering the variations in total traffic, would therefore be as follows:

	Search engine	Catalan	Spanish	Other	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
02/01/2021	Google	30.4	76.0	10.6	117
	Bing	42.9	78.6	28.5	150
	DuckDuckGo	47.0	76.5	23.5	147

In the case of the third day, the total number of visits would again be compared with the first, and so on.

The refining of this data depends on the languages included in each case and, therefore, a generalised calculation guide cannot be made.

Period

The series of data requested ranges from 1st January 2021 to present. This is due to the need to have a full series of one year with no remarkable incidents prior to 2022.

This monitoring is to be ongoing, in order to be able to quickly detect problems that may arise with search engines in the future, as described below. In any event, this is a voluntary option that is offered separately from the initial data provision.

From whom is data requested?

The data is requested based on two forms of collaboration:

- Direct request for data, to form part of or become an organization of special interest or that may have suffered bias by the search engines.
- In the future, a public call for the provision of data will be considered, which will be open to all multilingual website managers wanting to contribute. In this case, the contributions are sent to a website from where the form of collaboration can be chosen.

Furthermore, obtaining data directly is not ruled out, whether this be, for example, through direct enquiries to the search engines on the evolution of the positioning of specific words or other alternative methods of collection that do not depend on the aforementioned.

Under which conditions?

Collaborators providing traffic data decide on which the of recognition they would like to receive and the way in which this data is to be processed. Two levels of collaboration are therefore established:

- Totally anonymous: the logo of the collaborator is not included and their name is not given anywhere.
- Collaborator with anonymous data: the name and logo is included in the list of collaborators, but their data is never given publicly.

The website traffic data will be processed confidentially and will not be published individually under any circumstances without the explicit consent of the corresponding data subject, in line with the foregoing.

Where the data of a specific website, by way of example and always privately, is to be used to show the situation with regard Google or those responsible for any search engine, the data subject will be made aware of this and no specific approval will be required, and under no circumstances may this data be used or published by third parties.

As the holder of the traffic data, the puntCAT Foundation undertakes to sign non-disclosure agreements (NDA) with all parties, specifying the limitations of

use chosen by the collaborator, according to the profiles described above, and being held responsible for its safekeeping.

Where data is used for general publications, it is used anonymously, grouped into categories or into one single total, so that ownership cannot be recognised.

Despite the fact that the initial analysis focuses on the impairment of the service occurring throughout 2022, the aim is to be able to maintain these metrics constantly in order to agilely detect new variations. To this end, the same non-disclosure agreements include the possibility of extending this collaboration further by indicating that it is automatically renewed.

How is it processed?

Once collected, the data on each website will be reviewed to detect the evolutionary trends of organic search traffic: if a negative trend for the Catalan language is seen, i.e. the number of visits from the search engines drop in favour of Spanish or other languages, it will be included in the evidence file. Where this trend is not observed, a statistic record will be made to count the number of cases that have not been subject to service bias.

The data forming part of the evidence file will be analysed further to quantify the loss of users of the Catalan version in favour of the others. This will provide an accumulated indicator of the total number of visits that the Catalan language has lost due to the search engines over the year. Likewise, this same process will record the number of cases with more noticeable drops so that they can be used as an example for the search engines, where required.

Annex 2. Formal letter of request

Barcelona, 1st March 2023

Dear Sirs,

We are writing to ask for the collaboration of your organization in diagnosing the problem that is negatively affecting the visibility of web content in Catalan. You are probably aware that websites in Catalan have become less present in search engine results over recent months in favor of the versions of the same content in other languages, even if the user has Catalan set as a preference in their browsing environment.

This was first observed in 2022, but the precise date was not ascertained. The reason for this is unknown, despite the many different informal enquiries made to the search engine companies by interested individuals and companies. The main organisations promoting and defending the Catalan language (Amical Wikimedia, .cat Foundation, Institute of Catalan Studies, Ramon Llull Institute, Òmnium Cultural, Plataforma per la Llengua, Softcatalà, WICCAC) have therefore joined forces to allocate technical resources, knowledge and mobilisation capacity to diagnose the problem and help reverse it.

The Government of the Generalitat (Regional Government of Catalonia) is [actively involved](#) and has commissioned our alliance to create a device that quantifies the current loss of visibility of the Catalan language in web searches in comparison with the previous situation, and to ask the digital companies for solutions to this. It will also be capable of detecting similar incidents in the future.

The .cat Foundation, with the support of the other organisations, has agreed to draw up an initial report of the situation based on objective traffic data from search engines on a great many websites offering content in Catalan and in other languages.

We believe that your organisation's website or group of websites could provide very valuable information on this. We would therefore ask that, at your earliest convenience, you provide the figures of visits from search engines in any of the languages you offer on your website, in line with the technical specifications document we are attaching.

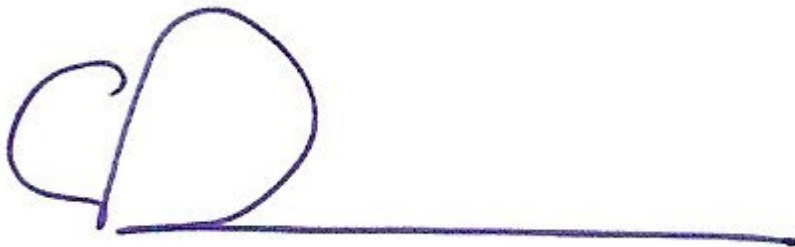
As you will see, we are asking for daily data from 1st January 2021 to 31 January 2023 in order to detect the time when the anomaly occurred in the inter-annual comparison and to quantify the extent of the problem.

In addition, we should be grateful if you would continue to provide updated data on a monthly basis in order to continue monitoring the presence of Catalan content in internet search engines.

In any event, we will keep the data you provide confidential and shall only disclose it on an aggregate basis, making it impossible to identify the individual source. This is included in the non-disclosure agreement that we are also attaching.

We hope you will agree to the collaboration of your organisation, and thank you in advance for your kind assistance. Please do not hesitate to contact us at suport@fundacio.cat or on 93.675.03.54 should you require any further information.

Kindest regards,



Genis Roca
Chairman of the .cat Foundation



Albert Cuesta
Coordinator of the
Alliance for the digital
presence of the Catalan
language

Annex 3. Non-disclosure agreement with Contributors

The Alliance for the Digital Presence of Catalan (hereinafter, "the Alliance") takes on the following confidentiality commitment with regard to the collaborating entity in studying the digital presence of Catalan.

- All the information the Foundation (Fundació .cat) receives or accesses from the collaborating entity in relation to its participation in the Alliance is, by default, fully confidential, regardless of its support, the time or the method by which this information is provided.
- The information received by Fundació .cat may only be used by employees and/or collaborators who must have access to it in order to carry out the tasks which are necessary for the study.
- All the employees and collaborators of Fundació .cat are subject to maintaining the confidentiality of the information to which they may have access while performing their functions through specific contractual clauses.
- Fundació .cat will not communicate the information provided by the collaborating entity to third parties, except in cases where it is displayed in search engines to show the results of the research and only if both parties expressly agree to this.
- The exchange of e-mails, chat sessions, SMS, calls or any other type of electronic communication between both parties will remain strictly confidential with no expiration date and subject to professional secrecy.
- Fundació .cat guarantees to the collaborating entity its full compliance with the legal provisions established by the data protection regulations and guarantees that it has established all the material, technical, human and legal means necessary to ensure compliance with this set of confidentiality measures.

Fundació .cat will issue a final report that it will share individually with all collaborating entities, as well as a global report with the aggregate data it will make public. This global report will not include individualized references to any of the collaborating entities nor will it allow the identification of specific cases.



Aliança per la
presència digital
del català

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

info@aliançadigital.cat



fundació .cat



institut
ramon llull



ÒMNIUM
LENGUA CULTURA PAÍS



SOFTCATALÀ

WACCAC