Font-Julian, CI.; Ontalba Ruipérez, JA.; Orduña Malea, E. (2018). Hit count estimate variability for website-specific queries in search engines: The case for rare disease association websites. Aslib Journal of Information Management. 70(2):192-213. https://doi.org/10.1108/AJIM-10-2017-0226

Preprint of article accepted for publication in ASLIB Proceedings
https://doi.org/10.1108/AJIM-10-2017-0226

Hit Count Estimate Variability in Search Engines: the Case for Rare Disease Association Websites

Abstract

Purpose
The main objective of this paper is to determine the effect of the chosen Search Engine Results Page (SERP) on the page count indicator when performing website rankings based on webometric techniques.

Design/Methodology/Approach
A sample of 100 Spanish rare disease association websites at two moments in time (2016 and 2017) is analysed, obtaining the page count for the first and last SERP in two search engines (Google and Bing).

Findings
It has been empirically demonstrated that there are differences between the number of hits returned on the first and last SERP in both Google and Bing. These differences are significant when they exceed a threshold value on the first SERP.

Research Limitations/Implications
Future studies considering other samples, more SERPs, and generating different queries other than web size (<site>) would be desirable to draw more general conclusions on the nature of quantitative data provided by general search engines.

Practical implications
Selecting a wrong SERP to calculate some metrics (in this case, page count) might provide misleading results, comparisons and performance rankings. The empirical data suggest that the first SERP captures the differences between websites better because it has greater discriminating power and is more appropriate for webometric longitudinal studies.

Social implications
Findings allow improving future quantitative webometric analyses based on page count metrics in general search engines.

Originality/Value
The page count variability between SERPs has been empirically analysed, considering two different search engines (Google and Bing), a set of 100 websites focused on a similar market (Spanish rare diseases associations), and two annual samples, making this study the most exhaustive on this issue to date.

Keywords: Search Engines; Hit Count Estimates; Page Count; Rare Diseases; Google; Bing.

1. Introduction

The use of search engines to (automatically or semi-automatically) extract data on web size (number of files hosted by a web domain) and web visibility (number of links or mentions received by a web domain) has been one of the main applications of the instrumental branch of cybermetrics (Orduna-Malea & Aguillo, 2014). These web indicators may provide, in a way that complements other quantitative and qualitative procedures and techniques, evidence for the greater or lesser presence and impact of web content and, therefore, of the sites that generate it and of the individuals or legal entities that manage it.

Cybermetric techniques that evaluate the impact of content hosted by web domains have been applied to many types of websites, such as universities (Aguillo, Ortega, & Fernández, 2008), academic journals (Vaughan & Hysen, 2002; Vaughan & Thelwall, 2003; Thelwall, 2012), companies (Vaughan, 2004; Vaughan & Wu, 2004; Orduna-

Malea et al, 2015), the media (Gao & Vaughan, 2005), political parties (Park, Kim & Barnett, 2004; Romero-Frías & Vaughan, 2010), local government (Holmberg & Thelwall, 2008), museums (Espadas, Calero, & Piattini, 2008; Gouveia & Kurtenbach, 2009; Orduna-Malea, 2014) and even hospitals (Utrilla-Ramirez et al, 2009; Utrilla-Ramirez, Aguillo, & Ortega, 2011).

However, the precision of the analytical tool (search engines) has always been questioned (Snyder & Rosenbaum, 1999) due to both the internal functioning of search engines (not made for quantitative purposes) and the nature of web information itself (dynamic and volatile). Although the literature has studied and proposed various methods for collecting web data from a cybermetric perspective (Bar-Ilan, 2001; Thelwall, 2004; Thelwall, 2006; Thelwall, 2009; Thelwall & Sud, 2011) the limitations of search engines have restricted the expansion and evolution of cybermetrics as a discipline (Thelwall, 2010). The disappearance of specific search commands (in particular, the command for finding out the number of hyperlinks that a particular website receives) and of search engines and entire platforms, such as Altavista and Yahoo Site Explorer, that were equipped with certain essential tools for cybermetric analysis (Orduna-Malea & Aguillo, 2014), contributed to a gradual abandonment by researchers of search engines as data sources. At the same time, other specialised platforms emerged, such as Majestic, Open Site Explorer and Ahrefs, which, despite their undoubted benefits and features for cybermetrics, offer less coverage and limit their services to the gathering of large amounts of data.

Although these issues have greatly affected the use of general search engines as sources of hyperlinks, the use of commercial search engines (mainly Google, Bing and Yahoo!) to calculate web size has also been questioned from the outset. Given the impossibility of externally calculating the number of files hosted on a website (without webmaster access privileges), together with the added difficulty of quantifying dynamically generated content (no file associated), web size has traditionally been calculated by the number of URLs that a search engine has indexed on the corresponding website. In the case of Google and Bing there is a search command (<site>) that retrieves URLs from a particular website. The procedure is based on running a search query (e.g. <site:nasa.gov>) and noting the number of results provided by the search engine (hit count estimate, HCE).

However, the lack of precision in HCEs (Uyar, 2009; Satoh & Yamana, 2012), their variability over time (Bar-Ilan, 1999), biased search coverage (Thelwall, 2000; Vaughan & Thelwall, 2004; Lewandowski, 2015), the frequency of updates (Lewandowski, 2015), and certain functional limitations (such as only returning a maximum number of results regardless of the HCE value) have led to a certain rejection of the use of search engines (especially Google) as tools for obtaining web impact data (Lawrence, Pelkey, & Soares, 2010).

The fact that Google did not even offer an API (Application Programming Interface) to facilitate automated data collection led cybermetric studies to switch to Bing, which did offer an API (http://datamarket.azure.com/dataset/bing/search) although it was limited to a maximum of 5,000 free queries per month (this service was withdrawn from the market on 31 December 2016). Numerous cybermetric studies (Thelwall, 2008; Thelwall & Sud, 2012; Wilkinson & Thelwall, 2013) were performed using the Bing API, due mainly to the fact that cybermetric applications such as Webometric Analyst

(http://lexiurl.wlv.ac.uk) worked with its API. However, Bing's lower coverage in comparison to Google, and various limitations (e.g. inaccurate for queries with more than 1,000 hits) greatly restricted the use of this API for metric purposes.

One of the various limitations of search engines, summarised in a schematic but complete way by Wilkinson & Thelwall (2013), is the variation in HCEs on each individual SERP. For example, if we wanted to find out the number of pages indexed by Bing for the Library of Congress, we could make the following query: <site:loc.gov>. On the first SERP (configured to display 10 results), the search engine informs us that there are 2,300,000 results. However, when we skip to SERP 50, the search engine reports 626 results (Figure 1).

Fig. 1. Variability of HCEs according to Bing SERP.

There are, therefore, fluctuations (upward or downward) in the number of hits shown, depending on the SERP that we look at. This effect is not exclusive to Bing, and is also found in other commercial search engines like Google (although the fluctuations now seem to have lessened). The fact that the search engine provides a rounded hit estimate, coupled with the existence of quasi-duplicates (very similar results that the system automatically eliminates when it considers them to be duplicates), causes the search engine to return HCEs with high variability.

Thelwall (2008) was a pioneer in tackling this problem empirically, exploring the Bing API. He not only detected and quantified the variations in HCEs, but also found that for those queries with an initial hit count (the first SERP) ranging between 300 and 2,000 results, the fluctuation in results according to the SERP was higher than for queries with a higher or lower initial number of results on the first SERP. However, he did not focus on web size queries (<site> command) but rather on general queries, where variability is thought to be greater since specific search commands are not being used, which in principle should restrict and filter results and, therefore, offer greater stability.

Due to the above limitations, the web size obtained through cybermetric techniques that use search engines will depend not only on the search engine used (coverage) but also on the SERP used to note the number of hits obtained for a specific query. In the latter case, if significant differences were confirmed in the number of results according to the SERP, not only should it be indicated which SERP was used in a given study, but the most appropriate should be chosen and used for all elements of the population under study. This would raise new methodological questions when ranking websites according to their size: would the relative positions of the studied websites remain unchanged even though the raw size data have varied? Would the variation in size be constant for all elements of a sample or population of websites studied?

Moreover, web size variability could indirectly affect the calculation of other composite indicators, such as Web Impact Factor (Ingwersen, 1998), which is calculated as the number of links (or mentions) that a website receives, divided by the number of web pages (page count) that a particular search engine has indexed for this site. This indicator, which is analogous to Impact Factor (currently produced by Clarivate Analytics) on the web, has been used in numerous studies as it provides data on the average visibility of a site based on the amount of content generated (Li, 2003; Noruzzi, 2006). However, its reliance on search engine coverage (Thelwall, 2000), its low

discriminating power in certain contexts (Thelwall, 2002), and certain statistical artefacts (WIF may be the same for very small sites and very large sites if its value is not normalised) have limited its use of late. However, in some internal environments (subdomains of large domains) it has proven to be useful (Orduna-Malea, 2013).

The above reasons justify the need to know and understand web size variability in search engines according to the SERP. This would help improve cybermetric analyses that use this indicator both directly and indirectly.

In order to provide an empirical answer to the research problem posited in this paper, we decided to work on a specific case study: the websites of Spanish rare disease associations. These websites are of particular relevance because these associations need both presence and visibility on the web in order to effectively disseminate information both to patients in particular and to society in general.

Given the role of the web in general (as a primary platform for the dissemination of health-related information) and of commercial search engines in particular (as socially accepted gateways to information search and retrieval processes) (Halavais, 2013), websites about rare diseases would improve dissemination of these diseases and enhance their visibility in society. In this context, the occurrence of inconsistencies in results (due to the variability of results on the different SERPs) could make certain associations and diseases invisible to an applied cybermetric study that did not use the right SERP to extract the web size data, or did not interpret them properly. Hence the need to know the effects that the variability of results on each SERP has on the calculation of the size of a website.

Research questions

This paper therefore raises the following research questions:

(**RQ1**): Do web size data vary significantly from one SERP to another?

(**RQ2**): If the answer to the above question is affirmative, is the variation in web size constant for all the websites in a homogeneous set?

(**RQ3**): Are significant differences found for the above questions depending on the search engine used?

(**RQ4**): Does the chosen SERP affect related indicators?

(**RQ5**): Is the rate of the variability in results between SERPs constant over time?

No sector or area of activity is more appropriate a priori than any other as a case study for answering these research questions. We chose to tackle the problem by conducting a study of a sample of Spanish rare disease association websites as a particular case study.

2. Research Background

Rare diseases are those diseases that affect a small proportion of the population; their origin and volatility make it extremely difficult to treat them properly. Depending on

the country, there are subtle differences in the definition in terms of prevalence of a rare disease. For example, in Europe a disease is defined as rare when it affects 1 person in 2,000, in Japan 1 person in 2,500 (or fewer than 50,000 patients), whereas in the United States it affects 1 person in 1,500 (or fewer than 200,000) (Forman et al, 2012).

The *European Organisation for Rare Diseases* (EURORDIS) (http://www.eurordis.org) estimates that there are between 6,000 and 8,000 rare diseases, which are generally chronic, disabling and 80% of which have genetic origins (*EURORDIS*, 2012). Moreover, according to the *World Health Organisation* (Humphreys, 2012), about 8% of the world population is affected by rare diseases.

Various international bodies have developed European Reference Networks (ERN) (http://ec.europa.eu/health/ern/policy_en), which provide quality information to those affected and to their families, and help professionals and reference centres to exchange information. Similarly, there are associations, federations, research centres and institutes, both national and international, that provide information exclusively geared towards these groups. Due to the scarcity of information on rare diseases, the websites of these associations generally become the greatest source of specific, high-quality information.

Cybermetric studies on health and medicine to date have been somewhat scarce. It is worth mentioning the article by Bowler, Hong and He (2011), in which they analysed the web visibility of a set of websites related to child health, and Groselj's study (2014), which characterised websites that provide health information. Lastly, there is the study by Utrilla-Ramirez, Aguillo and Ortega (2011), which described the Ranking Web of World Hospitals (http://hospitals.webometrics.info), a tool developed by the Cybermetrics Lab of the Spanish National Research Council (CSIC), which ranks the hospital websites according to a range of cybermetric indicators, including web size.

In the specific case of the visibility of rare diseases on the web, mention should be made of the study by Castillo, López and Carretón (2015), who analysed the type of tools used for communication by 143 portals of Spanish organisations related to rare diseases. They found that 82% of the organisations had a portal, although only 58% of them had Web 2.0 features.

However, we have not found cybermetric studies that focus on the analysis of rare diseases in general, or on websites that provide information on these diseases in particular. These types of studies (applied cybermetrics) would allow us to gauge the effectiveness of websites about rare diseases as sources of information and to design solutions to improve them and good practices in the event that the results were unsatisfactory. However, appropriate and precise methods (instrumental cybermetrics) are required before such studies may be carried out; and it is for this reason that it is essential to ascertain the effects of web size variability according to the SERP (the goal of this paper).

3. Method

First, a homogeneous sample of websites was obtained. To this end, the rare diseases were first identified, in order to then search for and select the related associations and eventually to obtain their URLs. After that, the web indicators of various search engines

(in this case, Google, Bing, Alexa and Majestic) were gathered. Finally, a statistical analysis of the data was carried out in order to provide answers to the research questions posited above.

3.1. Phase 1: obtaining the sample

The first step consisted of downloading the list of rare diseases available at Orphadata (http://www.orphadata.org) in XML format. Since not all the diseases listed on this source are globally considered to be rare diseases, it was decided to restrict the analysis to Spain. For that reason, all those diseases that were not included in the database of the National Registry of Rare Diseases of the Carlos III Health Institute (https://registroraras.isciii.es) were eliminated from the list.

Given the high number of rare diseases, we chose to work with a sample of the 50 diseases with the greatest number of hits in Google. For this, an ad-hoc crawler was developed to automate the process of a text search in Google (each rare disease was searched for using its Spanish nomenclature) via https (<https://www.google.es/search?&q="name of the disease">) and to subsequently collect the number of results obtained for each disease. The 50 diseases selected, and the number of hits obtained for each, are listed in Appendix A.

The next step consisted of locating associations for the 50 selected rare diseases. To this end, the complete list of rare disease associations provided by three sources (Orpha.net, FEDER and EURORDIS) was compiled to obtain one single list, eliminating any duplicates (n= 438).

Once the full list of associations had been obtained, those associations related to the 50 chosen diseases were selected. We decided to include two associations per disease in order to obtain a final list of 100 associations. Appendix B contains the final list of the 100 selected associations, together with their official URLs.

3.2. Phase 2: obtaining the web data

Subsequently, data were collected for each website of each association. Table 1 shows the indicators selected, their description and the source from which they were obtained.

Table 1. Description of indicators and sources used

In the case of page count indicators, Google and Bing were selected because they are currently the general commercial search engines with the greatest coverage. In the case of visibility indicators, the data were taken from Majestic (https://majestic.com), the platform with the largest database of links at present. Additionally, Alexa (http://www.alexa.com) was used, in order to have an alternative source.

Data collection was carried out in two different time periods (August 2016 and August 2017), in order to have two annual samples to determine the temporal variability of the results.

3.3. Phase 3: Data analysis

Once all the data had been collected, an application was developed in Phyton to automatically extract and export them to a spreadsheet.

Subsequently, a statistical analysis of the results was performed, including a descriptive analysis of the distribution of the variables, a Mann-Whitney test to compare distributions and a regression analysis (potential trend). After this, we calculated the correlation between the different indicators used, according to both the source (search engine) and the SERP. Due to the non-normal distribution of the web data, the Spearman correlation was used ($\alpha < 0.1$).

Finally, different variants of the Web Impact Factor were calculated, taking into account the different sources for numerator (web visibility) and denominator (page count). In this way, a total of 12 WIF variants were obtained, the result of combining the web size data of the first and last SERP of Google and Bing with the visibility data of Majestic and Alexa.

4. Results

4.1. Volume of data according to the SERP

The descriptive analysis of the results reflects a higher data volume on the first SERP, although this difference seemed to be more pronounced in Google than in Bing (Table 2). While in Google the median of hits on the first SERP ($M_e = 313$) is twice that of the last SERP ($M_e = 145.5$), in the case of Bing the median values are very similar ($M_e = 146$ on $SERP_1$; $M_e = 135$ on $SERP_n$).

Table 2. Descriptive analysis of SERP samples

We also see how the standard deviation on Google $SERP_1$ is very high ($\sigma = 8{,}972.5$), which shows a high page count variability among the 100 associations analysed. However, taking the results of Google $SERP_n$, this variability is markedly reduced ($\sigma = 156.6$), lower even than that obtained for Bing $SERP_n$ ($\sigma = 278.6$).

The 10 associations with the greatest web size according to Google $SERP_1$ are shown in Table 3, along with the number of hits on $SERP_n$ and the percentage by which this number has reduced (a variable named *hit shrink* for the purposes of this paper). As can be seen, the $SERP_1$ results are on a much higher order of magnitude than the $SERP_n$ results. However, it should be noted that the overall volume of hits for Google is mainly due to coverage of two websites (<aecc.es> and <enfermedades-raras.org>), which obtained 79,800 and 42,700 hits respectively. The number of results for these two sites was drastically lower on $SERP_n$ (567 and 569, respectively), with more results shown for <enfermedades-raras.org>, although the difference in hits on $SERP_1$ was much more significant (difference of 37,100 hits).

Table 3. Top 10 web sites with the largest web size according to Google

If we compare the results for Google (Table 3) with Bing (Table 4) we may observe certain inconsistencies. For example, <cnio.es> obtained a page count on Bing $SERP_1$

(6,740) higher than Google $SERP_1$ (4,270). However, <enfermedades-raras.org> obtained 42,700 on Google $SERP_1$ compared to only 6,610 for Bing.

Table 4. Top 10 web sites with the largest web size according to Bing

With regard to the hit shrink rate, the partial results of Tables 3 and 4 appear to show random and contradictory data. In order to clarify this variability, the hit shrink average and median is broken down into ranges of values (Table 5), demonstrating a general trend for a greater percentage reduction for the queries that displayed the greatest number of hits on the first SERP.

Table 5. Hit shrink average and median in hit ranges ($SERP_1$)

Moreover, the existence of a strong correlation between the number of hits obtained and the hit shrink percentage (Table 6) is confirmed. That is, the higher the HCE values, the greater the number of hits "lost" between $SERP_1$ and $SERP_n$. This effect is more pronounced in Google (R_s= 0.9) than in Bing (R_s= 0.6).

Table 6. Hit shrink percentage correlation against SERP values

To better understand the differences between the data, a Mann-Whitney (two-tailed) test was also performed on the results obtained for $SERP_1$ and $SERP_n$ in both Google and Bing. In the case of Google (p-value: <0.0001; $\alpha < 0.1$), the test confirmed that the two samples are statistically independent, whereas in the case of Bing (p-value: 0.302; $\alpha < 0.1$), the results of the test indicate that there are no significant differences.

Although the Google $SERP_1$ results are statistically different from the $SERP_n$ results, the correlation between them is very high (r_s= 0.90), although slightly lower than that obtained for Bing between $SERP_1$ and $SERP_n$ (r_s= 0.96), as can be seen in Table 7. Likewise, the inter-search engine correlations are equally high. Data from Majestic (URLs indexed) were also included, and they show a complete absence of correlation with the other variables.

Table 7. Correlation (Spearman) between the number of hits according to SERP and Search Engine

The distribution of the data, especially for associations with a small web size, explains in part why the results of the first and last SERP in Google correlate, despite having statistically independent data distributions. Figure 2 shows how the results provided by $SERP_1$ and $SERP_n$ in Google are practically identical when they do not exceed an approximate value of 100, which is the case for 26% of the association websites. This effect is even more pronounced in the case of Bing (Figure 3), where results between SERPs are practically identical when they remain below the value of 1,000 (77% of the results).

Fig.2. Web size distribution for rare disease associations according to SERP (Google).

Fig.3. Web size distribution for rare disease associations according to SERP (Bing).

This distribution of the data, together with the previously obtained correlations (Table 7), enables $SERP_1$ results to be predicted according to the results obtained on $SERP_n$

(Figs. 4 and 5); high determination coefficients are obtained both in Google (R^2= 0.77) and Bing (R^2= 0.86).

Fig.4. Power trendline between $SERP_1$ (G) and $SERP_n$ (G).

Fig.5. Power trendline between $SERP_1$ (B) and $SERP_n$ (B).

4.2. Effects of the SERP on Impact Factor

Table 8 shows the correlation between the 12 Web Impact Factor variants considered. The results show WIF values that are strongly correlated between each other, both between $SERP_1$ and $SERP_n$ of the same search engine [higher in Majestic than in Alexa] (intra-search engine correlation), and between the same SERP in different search engines (inter-search engine correlation). However, the results also show an absence of correlation when, for the same SERP and search engine, the WIF values obtained through Majestic and Alexa are compared, showing the poor correlation between the two link sources.

Table 8. Correlation matrix between Web Impact Factor (WIF) variants

The lack of correlation between Majestic and Alexa is evident in the results in Table 9, which shows the correlation between the 12 WIF variants and the Citation Flow and Trust Flow values (from Majestic). In this case, slightly higher values are observed in the WIF variants calculated using Google and, more specifically, $SERP_n$, as the source. While Citation Flow reaches its highest correlation with WIF-3, WIF-8 and WIF-9 (R_s = 0.6; $\alpha < 0.1$), in the case of Trust Flow it is with the WIF-8 variant (number of sites according to Majestic divided by the number of hits according to Google $SERP_n$) (R_s = 0.6; $\alpha < 0.1$).

Table 9. Correlation between Web Impact Factor variants and Citation Flow and Trust Flow

4.3. Annual change in the hit shrink rate

If, lastly, we compare the results obtained in 2016 with those from 2017, we observe a strong correlation, especially between Google $SERP_n$ (R_s= 0.9) (Table 10).

Table 10. Correlation (Spearman) between SERPs (2016 – 2017)

Despite the strong correlation, there are notable differences in the observed hit shrink rate, both in Google (Fig. 6) and, especially, in Bing (Fig. 7). The annual comparison shows hit shrink rates that are similar in observations that had a high number of hits in $SERP_1$, but quite different in observations with fewer hits.

Fig.6. Variation of the hit shrink rate between 2016 and 2017 in Google.

Fig.7. Variation of the hit shrink rate between 2016 and 2017 in Bing.

5. Discussion

The web size value taken from the hit count estimates using a <site> query varies substantially between the first and last SERP. In the case of Google, this variability is statistically significant, while in Bing it is lower (RQ1).

In spite of the differences (significant or not) in the raw values of the number of hits obtained, the correlation between $SERP_1$ and $SERP_n$ (both in Google and in Bing) is positive and very high, which could indicate that, with some exceptions, websites that have a higher (or lower) number of results on $SERP_1$ also have a higher (or lower) number of hits on $SERP_n$. However, the results of the correlations are misleading in that they are biased due to the high number of observations with a low number of results.

In fact, the distribution of the results (Figures 2 and 3) shows a threshold above which the difference in hits between $SERP_1$ and $SERP_n$ is accentuated, whereas below this threshold the difference virtually disappears. Therefore, the percentage of observations that are above or below the threshold will largely determine whether the variability between SERPs is high or low. That is, variability depends on the order of magnitude of the values obtained, and is not constant (RQ2).

In the analysed sample (100 Spanish rare disease association websites), the percentage of queries that obtained fewer than 100 hits (where there are hardly any differences between the number of hits on $SERP_1$ and $SERP_n$) is so high (26% in Google, 77% in Bing) that not only is the correlation between SERPs strong, but the prediction of one value according to the other is also, within a tolerable range, accurate (Figures 4 and 5).

Moreover, these threshold values are different for each search engine (RQ3). In the specific case of Spanish rare disease associations, the threshold in Bing seems to be around 1,000 results, while in Google it is around 100. However, an analysis of other populations is required to establish whether these thresholds are independent of the samples or not.

These results partially concur with the results obtained by Thelwall (2008), who after analysing a set of 4,000 words in English from a sample of approximately 68,000 blogs, found that in queries that return less than 100 results on the first SERP, the variation of results in subsequent SERPs is lower. However, Thelwall detected greater variability in queries for which the first SERP returned between approximately 300 and 2,000 results, an effect not observed in our study. Nevertheless, the results obtained by Thelwall cannot be directly compared with the results of this study due to a number of methodological differences. While the Thelwall study analysed words in English (on all SERPs) with Bing, this study has focused on 100 specific <site:url.com> queries in both Bing and Google, taking only the values of the first and last SERPs. Moreover, this behavioural pattern of the Bing search engine technology may have been modified since Thelwall's study (carried out in 2006).

The strong correlation between $SERP_1$ y $SERP_n$ in both Google and Bing (a positive and significant correlation even between the SERPs of both search engines) demonstrates that the choice of SERP is not a critical factor when calculating the WIF, since when isolating the numerator of the WIF formula (visibility), the WIF values

obtained according to $SERP_1$ and $SERP_n$ correlate in a positive and significant way (see Table 8).

Although this paper has analysed the possible influence that the choice of a particular SERP may have when calculating the web size directly, and the WIF indirectly, it should be pointed out that it is not advisable to use this indicator exclusively to evaluate websites. Their use must be complemented by other webometric indicators (such as Citation Flow and Trust Flow) when gathering data on a website's impact.

In more specific terms, when associating WIF with quality indicators (particularly Trust Flow), it has been demonstrated that the WIF value taken from $SERP_n$ achieves a higher correlation with Trust Flow than the other variants; so the number of hits returned by $SERP_n$ could be considered a more accurate proxy of the impact of a website than that of $SERP_1$, although its discriminating power is lower.

Indeed, the effect of the discriminating power of $SERP_1$ was verified with the annual analysis. The search engine reduces the number of hits to a limited number (1,000 in Bing, 800 in Google). This fact affects longitudinal analyses because even if a website grows in a year, $SERP_n$ will always display a bounded value. This value should therefore not be used in longitudinal studies, unless the aim is to know simply which websites, within a wide margin, have a greater or smaller size within a specific set of websites.

However, the results should be treated with some caution, as there are other external variables (such as geolocation or desynchronisation between data centres) that could equally affect the number of hits displayed on each SERP. In addition, the <site> command is not exhaustive (it does not retrieve all URLs that are actually indexed by a search engine, it only offers an estimate). Despite these limitations, which in statistical terms affect all observations equally, we consider that the main conclusions of this paper (variability between SERPs for <site> queries) are not affected.

6. Conclusions

The main conclusions of this study are presented below:

(RQ1) It has been demonstrated that there are differences between the number of hits returned on the first and last SERP in both Google and Bing. These differences are significant when they exceed a threshold value on the first SERP of approximately 1,000 hits in Bing, and 100 hits in Google.

(RQ2) For those queries that obtain a number of hits below the threshold on $SERP_1$, the difference between SERPs is constant. However, for queries with hits above the threshold, the difference becomes unstable, following a power ratio.

(RQ3) The difference in the number of hits between $SERP_1$ and $SERP_n$ is, on average, significantly lower in Bing than in Google.

(RQ4) Although the raw final WIF values obviously vary according to the denominator value (HCE of a <site> query), a positive and strong correlation was obtained between all the WIF values calculated from all the SERPs. Therefore, the

choice of SERP is not critical, as long as it is analysed in relation to the values obtained for the other observations. That is, the value of the WIF of a website is not important in itself, but in relation to the value achieved by the other websites with which it is compared. Despite the fact that the choice of SERP is unimportant, we conclude that Google $SERP_n$ is a more accurate proxy of the impact of a website measured through Trust Flow.

(RQ5) The hit variability rate between $SERP_1$ and $SERP_n$ is not constant over time. This fact is determined by the hit shrink process, limited to a discrete margin on $SERP_n$ irrespective of the value reached on $SERP_1$. For this reason, $SERP_n$ (both in Google and Bing) is not recommended for use in longitudinal studies.

However, these conclusions are determined by the sample analysed (100 Spanish rare disease association websites). Other samples, considering more SERPs and generating different queries other than web size (<site>), would have to be studied to be able to draw more general conclusions regarding the variability of results according to the chosen SERP and the effect of this variability on the study of the web impact of a website.

In the case of the sample of rare disease associations, we may conclude that the web size of these associations in Bing may be studied regardless of the SERP used, especially with regard to relative size and not to absolute size (the number of hits is a mere approximation to the actual size, which cannot be determined by external methods). In the case of Google, the effect of the SERP is more pronounced, especially for sites with larger web sizes. However, the hit shrink effect could reduce the differences between the sites. Therefore, although $SERP_n$ is a more accurate proxy for web impact, $SERP_1$ might be more useful for determining relative web size differences.

References

Aguillo, I. F., Ortega, J. L.and Fernández, M. (2008), "Webometric ranking of world universities: Introduction, methodology, and future developments", *Higher education in Europe*, Vol. 33, No. 2-3, pp. 233-244.

Bar-Ilan, J. (1999), "Search engine results over time: A case study on search engine stability", *Cybermetrics*, Vol. 2/3, No. 1.

Bar-Ilan, J. (2001), "Data collection methods on the Web for infometric purposes—A review and analysis", *Scientometrics*, Vol. 50, No. 1), pp. 7-32.

Bowler, L., Hong, W. Y. and He, D. (2011), "The visibility of health web portals for teens: a hyperlink analysis", *Online Information Review*, Vol. 35, No. 3, pp. 443-470.

Castillo Esparcia, A., López Villafranca, P. and Carretón Ballester, M.C. (2015), "La comunicación en la red de pacientes con enfermedades raras en España", *Revista Latina de Comunicación Social*, Vol. 70, pp. 673-688.

Espadas, J., Calero, C. and Piattini, M. (2008), "Web site visibility evaluation", *Journal of the American Society for Information Science and Technology*, Vol. 59, No. 11), pp. 1727-1742.

EURORDIS (European Organization for Rare Diseases) (2012). *What is a Rare Disease?* Available at: http://www.eurordis.org/content/what-rare-disease (accessed 08 October 2017).

Forman, J., Taruscio, D., Llera, V. A., Barrera, L. A., Coté, T. R., Edfjäll, C., Gavhed, D., Haffner, M.E., Nishimura, Y., Posada, M., Tambuyzer, E., Groft, S.C. and Henter, J-I. (2012), "The need for worldwide policy and action plans for rare diseases", *Acta Paediatrica*, Vol. 101, No.8, pp. 805-807.

Gao, Y. and Vaughan, L. (2005), "Web hyperlink profiles of news sites: A comparison of newspapers of USA, Canada, and China", *Aslib proceedings*, Vol. 57, No. 5, pp. 398-411.

Gouveia, F. C. and Kurtenbach, E. (2009), "Mapping the web relations of science centres and museums from Latin America", *Scientometrics*, Vol. 79, No. 3, pp. 491-505.

Groselj, D. (2014), "A webometric analysis of online health information: sponsorship, platform type and link structures", *Online Information Review*, Vol. 38, No. 2, pp. 209-231.
Halavais, A. (2013), *Search engine society*, John Wiley & Sons.
Holmberg, K. and Thelwall, M. (2008), "Local government web sites in Finland: A geographic and webometric analysis", *Scientometrics*, Vol. 79, No. 1, pp. 157-169.
Humphreys, G. (2012), "Coming together to combat rare diseases", *Bulletin of the World Health Association*, Vol. 90, No. 6, pp. 401-476.
Ingwersen, P. (1998), "The calculation of web impact factors", *Journal of documentation*, Vol. 54, No. 2, pp. 236-243.
Lawrence, T., Pelkey, N. and Soares, S. (2010),"'Googleology': powerful tool or unreliable evidence?", *Bulletin of Zoological Nomenclature*, Vol. 67, No. 3.
Lewandowski, D. (2008), "A three-year study on the freshness of web search engine databases", *Journal of Information Science*, Vol. 34, No. 6, pp. 817-831.
Lewandowski, D. (2015), "Living in a world of biased search engines", *Online Information Review*, Vol. 39, No. 3. Available at: https://doi.org/10.1108/OIR-03-2015-0089 (accessed 12 November 2007).
Li, X. (2003), "A review of the development and application of the Web Impact Factor", *Online Information Review*, Vol. 27, No. 6, pp. 407-417.
Noruzi, A. (2006), "The web impact factor: a critical review", *The electronic library*, Vol. 24, No. 4, pp. 490-500.
Orduna-Malea, E. and Aguillo, I. F. (2015), *Cibermetría. Midiendo el espacio red*, UOC Publishing, Barcelona.
Orduna-Malea, E. (2013), "Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: The case of Spain", *Journal of the Association for Information Science and Technology*, Vol. 64, No. 10, pp. 2100-2114.
Orduna-Malea, E. (2014), "Caracterización y rendimiento del sistema museístico de la Comunidad Valenciana a través de un análisis cibermétrico", *In Gestión cultural: innovación y tendencias* (pp. 13-43), Tirant Lo Blanch, Valencia.
Orduna-Malea, E., Delgado López-Cózar, E., Serrano-Cobos, J. and Lloret-Romero, N. (2015), "Disclosing the network structure of private companies on the web: the case of Spanish IBEX 35 share index", *Online Information Review*, Vol. 39, No. 3, pp. 360-382.
Park, H. W., Kim, C. S. and Barnett, G. A. (2004), "Socio-communicational structure among political actors on the web in South Korea: The dynamics of digital presence in cyberspace", *New Media & Society*, Vol. 6, No. 3, pp. 403-423.
Romero-Frías, E. and Vaughan, L. (2010), "European political trends viewed through patterns of Web linking", *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 10, pp. 2109-2121.
Satoh, K. and Yamana, H. (2012, April), "Hit count reliability: how much can we trust hit counts?", In *Asia-Pacific Web Conference* (pp. 751-758). Springer, Berlin Heidelberg.
Snyder, H. and Rosenbaum, H. (1999), "Can search engines be used as tools for web-link analysis? A critical view", *Journal of documentation*, Vol. 55, No. 4, pp. 375-384.
Thelwall, M. (2000), "Web impact factors and search engine coverage", *Journal of documentation*, Vol. 56, No. 2, pp. 185-189.
Thelwall, M. (2002), "A comparison of sources of links for academic Web Impact Factor calculations", *Journal of Documentation*, Vol. 58, No. 1, pp. 66-78.
Thelwall, M. (2004), *Link analysis: An information science approach*, Academic Press, San Diego.
Thelwall, M. (2006), "Interpreting social science link analysis research: A theoretical framework", *Journal of the Association for Information Science and Technology*, Vol. 57, No. 1, pp. 60-68.
Thelwall, M. (2008), "Extracting accurate and complete results from search engines: Case study Windows Live", *Journal of the American Society for Information Science and Technology*, Vol. 59, No. 1, pp. 38-50.
Thelwall, M. (2009), *Introduction to webometrics: Quantitative web research for the social sciences*, Morgan & Claypool, San Rafael (CA).
Thelwall, M. (2010), "Webometrics: Emergent or Doomed?", *Information Research*, Vol. 15, No. 4.
Thelwall, M. (2012), "Journal impact evaluation: a webometric perspective", *Scientometrics*, Vol. 92, No. 2, pp. 429-441.
Thelwall, M. and Sud, P. (2011), "A comparison of methods for collecting web citation data for academic organizations", *Journal of the Association for Information Science and Technology*, Vol. 62, No. 8, pp. 1488-1497.
Thelwall, M. and Sud, P. (2012), "Webometric research with the Bing Search API 2.0", *Journal of Informetrics*, Vol. 6, No. 1, pp. 44-52.

Utrilla-Ramírez, A. M., Aguillo, I. F. and Ortega, J. L. (2011), "Visibilidad de la web hospitalaria iberoamericana. Perspectiva de su actividad científica en internet", *Medicina Clínica*, Vol. 137, No. 13, pp. 605-611.

Utrilla-Ramírez, A. M., Fernández, M., Ortega, J. L. and Aguillo, I. F. (2009), "Clasificación Web de hospitales del mundo: situación de los hospitales en la red", *Medicina clínica*, Vol. 132, No. 4, pp. 144-153.

Uyar, A. (2009), "Investigation of the accuracy of search engine hit counts", *Journal of Information Science*, Vol. 35, No. 4, pp. 469-480.

Vaughan, L. (2004), "Exploring website features for business information", *Scientometrics*, Vol. 61, No. 3, pp. 467-477.

Vaughan, L. and Hysen, K. (2002), "Relationship between links to journal Web sites and impact factors", *Aslib Proceedings*, Vol. 54, No. 6, pp. 356-361.

Vaughan, L. and Thelwall, M. (2003), "Scholarly use of the Web: What are the key inducers of links to journal Web sites?", *Journal of the Association for Information Science and Technology*, Vol. 54, No. 1, pp. 29-38.

Vaughan, L. and Thelwall, M. (2004), "Search engine coverage bias: evidence and possible causes", *Information processing & management*, Vol. 40, No. 4, pp. 693-707.

Vaughan, L. and Wu, G. (2004), "Links to commercial websites as a source of business information", *Scientometrics*, Vol. 60, No. 3, pp. 487-496.

Wilkinson, D. and Thelwall, M. (2013), "Search markets and search results: The case of Bing", *Library & Information Science Research*, Vol. 35, No. 4, pp. 318-325.

Preprint of article accepted for publication in ASLIB Proceedings
https://doi.org/10.1108/AJIM-10-2017-0226

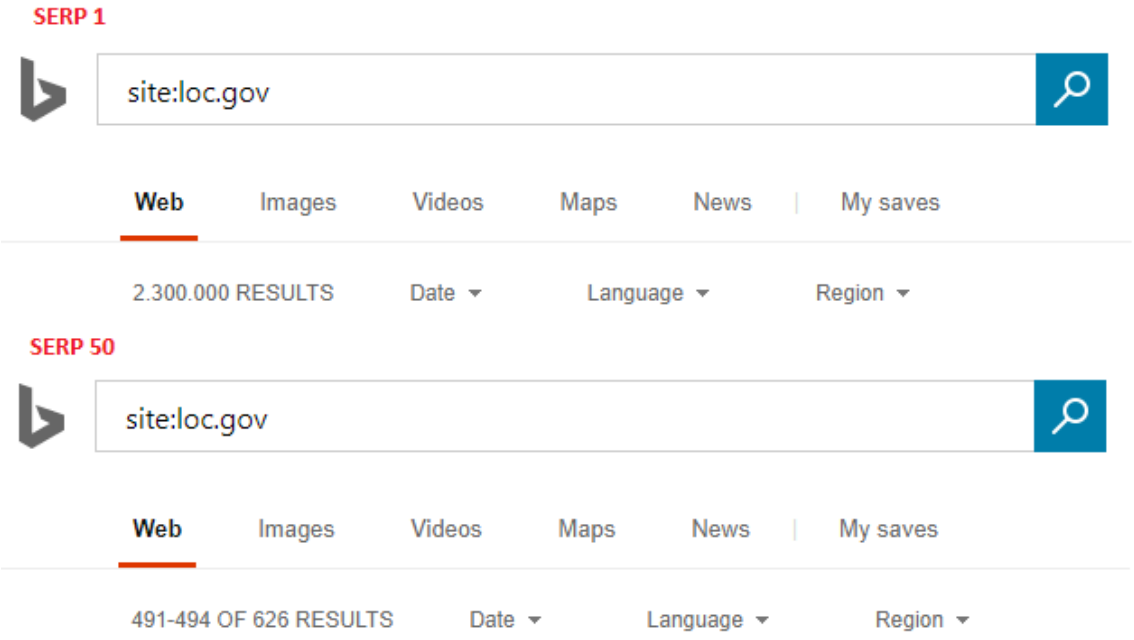

Fig. 1. Variability of HCEs according to Bing SERP.

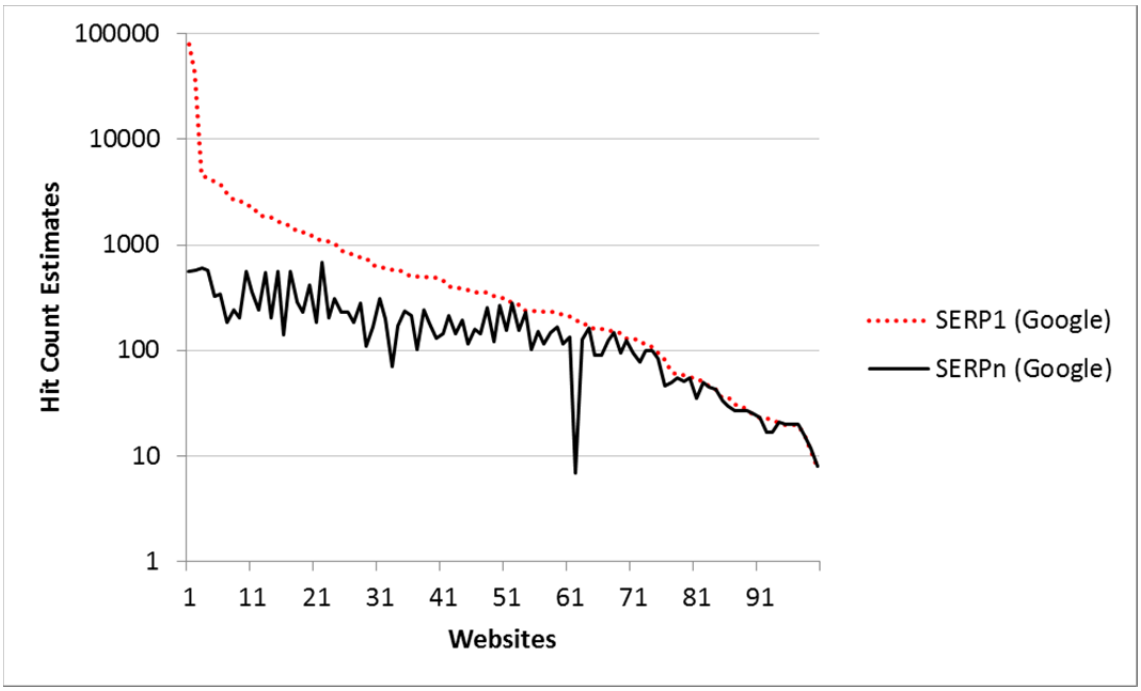

Fig.2. Web size distribution for rare disease associations according to SERP (Google).

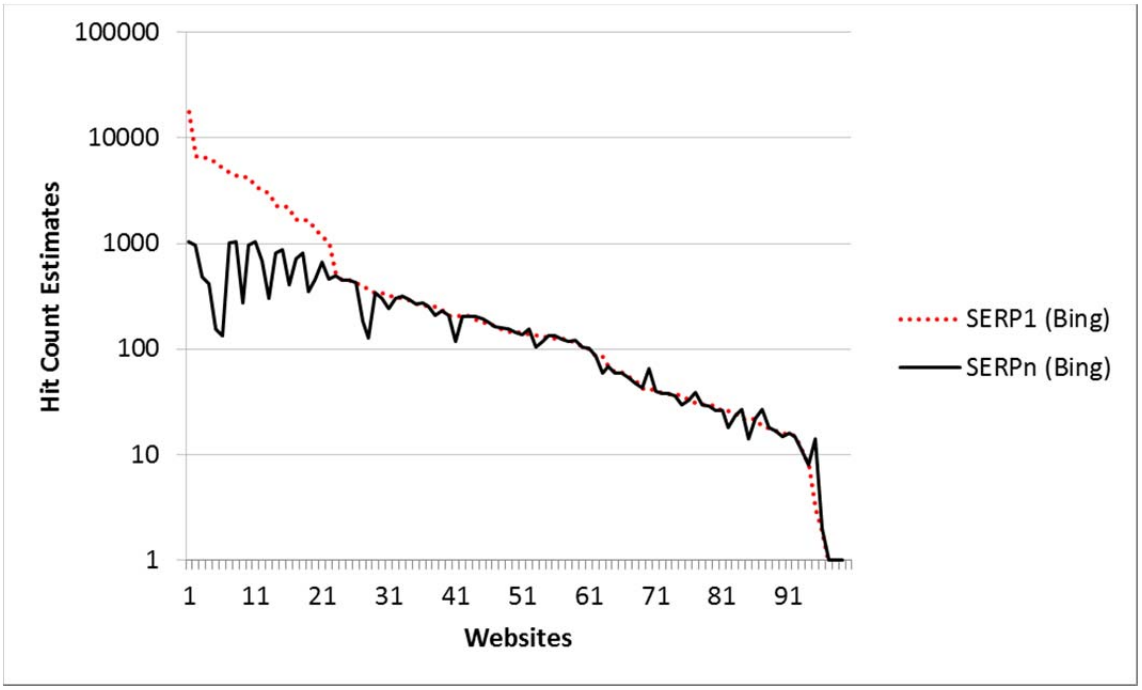

Fig.3. Web size distribution for rare disease associations according to SERP (Bing).

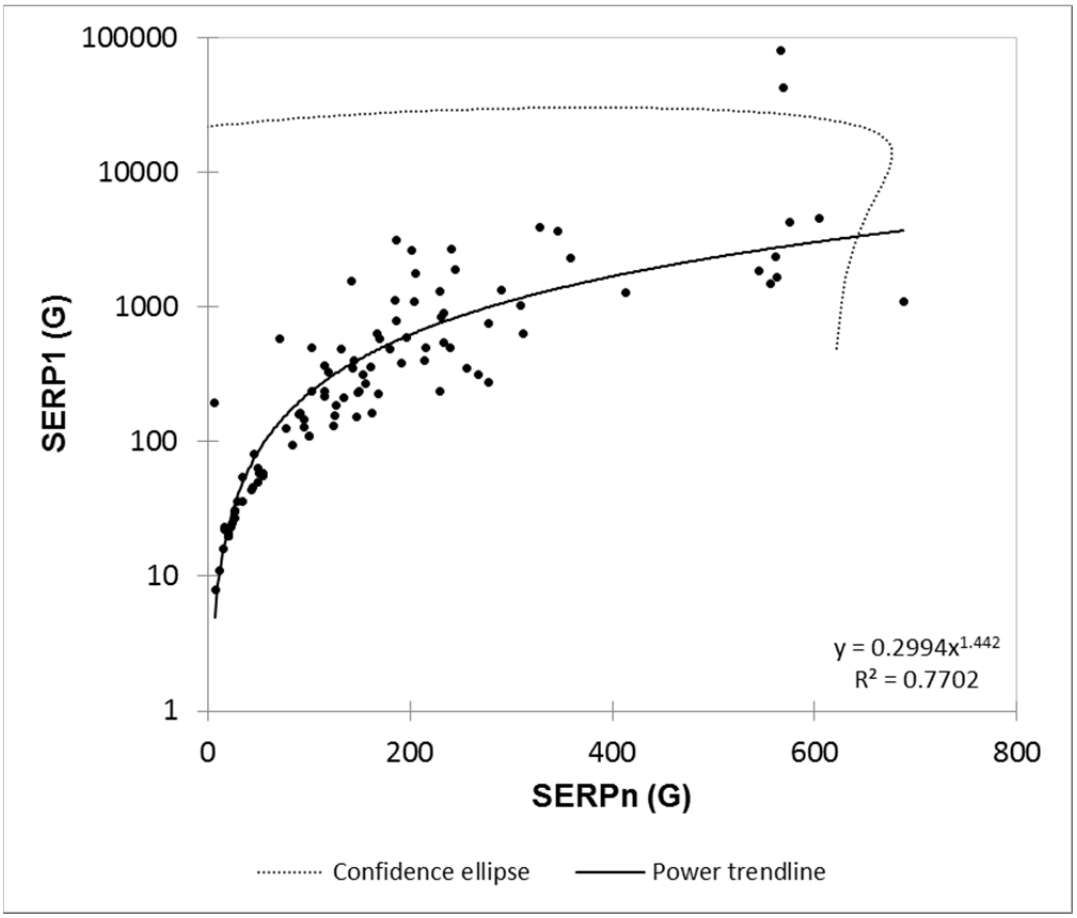

Fig.4. Power trendline between $SERP_1$ (G) and $SERP_n$ (G).

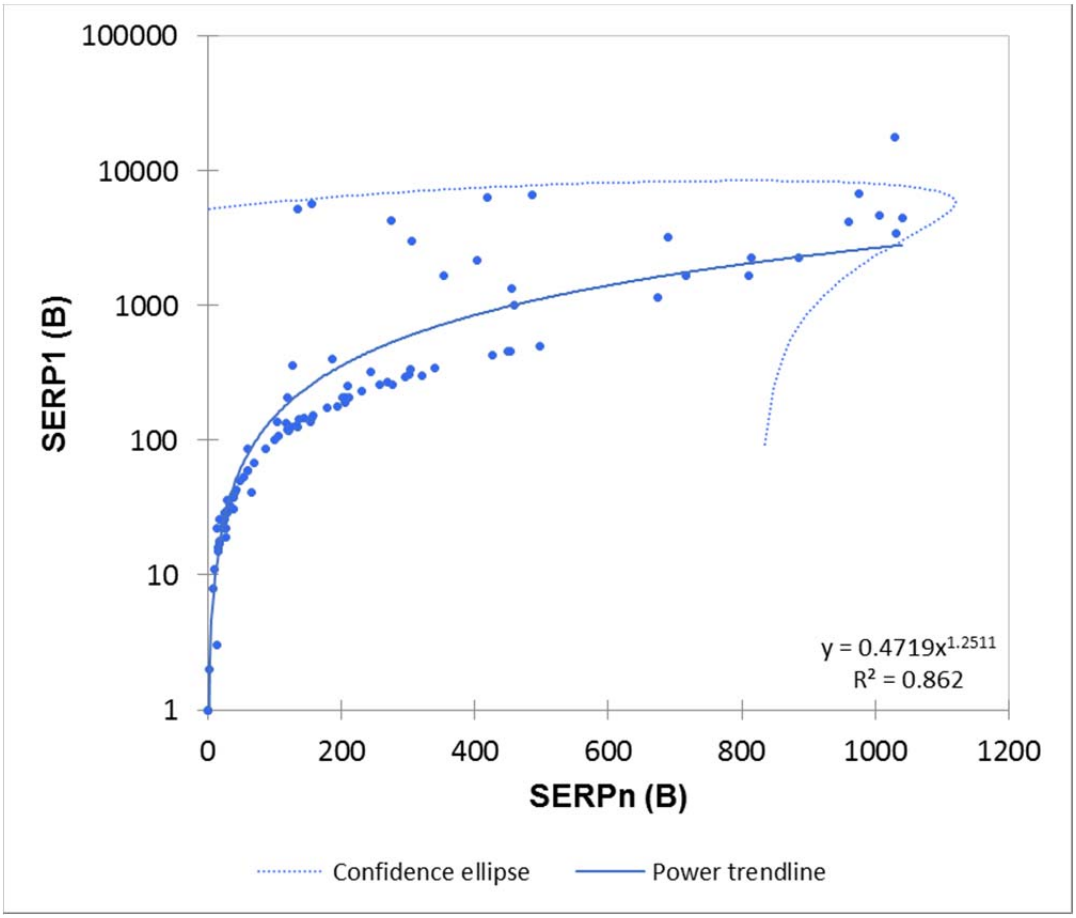

Fig.5. Power trendline between $SERP_1$ (B) and $SERP_n$ (B).

Fig.6. Variation of the hit shrink rate between 2016 and 2017 in Google.

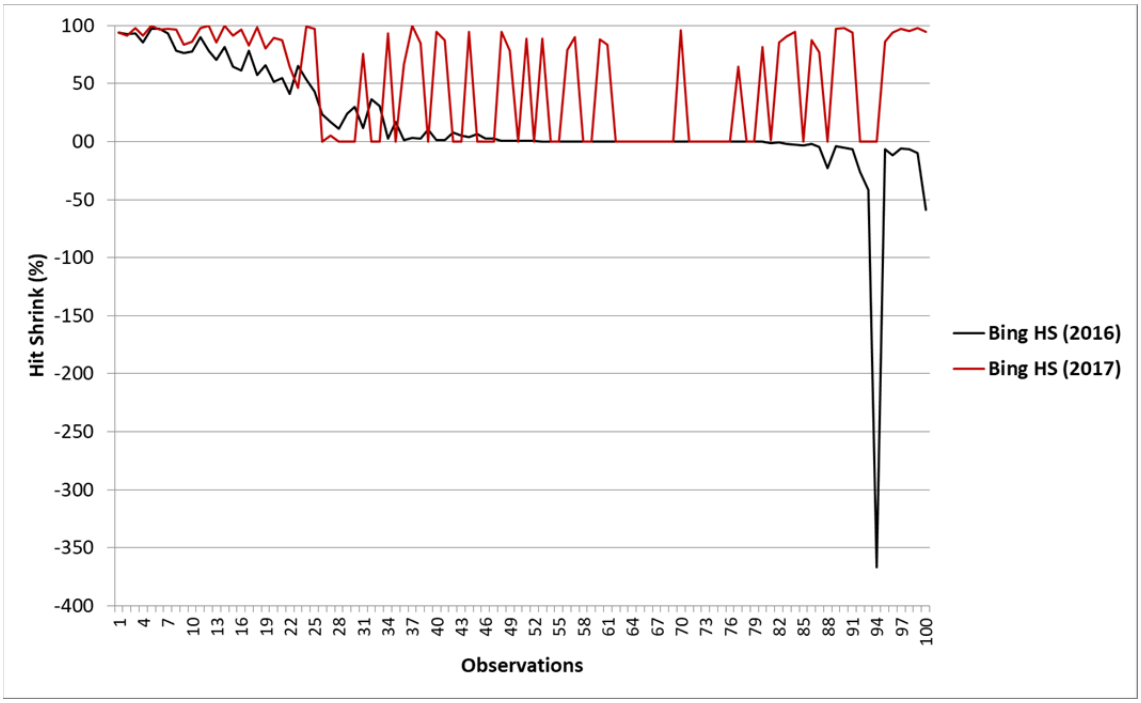

Fig.7. Variation of the hit shrink rate between 2016 and 2017 in Bing.

Table 1. Description of indicators and sources used.

Category	Indicator	Source	Scope	Query
Page Count	$SERP_1$	Google; Bing	Number of URLs indexed for a domain, considering the first search engine results page	site:website.com
	$SERP_n$	Google; Bing	Number of URLs indexed for a domain, considering the last search engine results page	site:website.com
	URLs indexed	Majestic	Number of URLs indexed for a domain	API
Visibility	External links	Majestic	Number of external URLs that link to reference URL	API
	Referral domains	Majestic; Alexa	Number of external web domains that link to reference URL	API
	Citation Flow	Majestic	Score between 0-100 which helps to measure the link equity or "power" the website carries	API
	Trust Flow	Majestic	Score between 0-100 which helps to measure the quality of a website, based on the number of links received from a seed of trusted sites, based on a manual review of the web	API
Impact	WIF		$WIF = \frac{Visibility}{Page\ count}$	Prepared by the authors

Note: More information about flow metrics (Citation Flow and Trust Flow) is available at Majestic website: https://es.majestic.com/support/glossary#FlowMetrics

Table 2. Descriptive analysis of SERP samples.

Statistics	GOOGLE		BING	
	$SERP_1$	$SERP_n$	$SERP_1$	$SERP_n$
Minimum	8	7	1	1
Maximum	79,800	688	17,800	1,040
1st Quartile	90.75	55	37	37
Median	313	145.5	146	135
3rd Quartile	862	231.5	439.5	304
Mean	1,901.6	178.96	1,023.44	239.6
Variance	80,504,990.4	24,534.3	5,532,931.4	77,588.4
Standard deviation	8,972.5	156.6	2,352.2	278.6

Table 3. Top 10 web sites with the largest web size according to Google.

URL	$SERP_1$ (G)	$SERP_n$ (G)	Hit shrink (%)
aecc.es	**79,800**	567	99.3
enfermedades-raras.org	**42,700**	569	98.7
asociaciondoce.com	**4,550**	604	86.7
cnio.es	**4,270**	576	86.5
asem-esp.org	**3,960**	329	91.7
mpsesp.org	**3,700**	346	90.6
ataxiasandalucia.org	**3,150**	187	94.1
fqmadrid.org	**2,690**	241	91
fegerec.es	**2,660**	202	92.4
fgcasal.org	**2,370**	562	76.3

Table 4. Top 10 web sites with the largest web size according to Bing.

URL	$SERP_1$ (B)	$SERP_n$ (B)	Hit shrink (%)
aecc.es	**17,800**	1,029	94.2
cnio.es	**6,740**	975	85.5
enfermedades-raras.org	**6,610**	487	92.6
acnefi.org	**6,340**	418	93.4
fibrosisquistica.org	**5,750**	155	97.3
corazonyvida.org	**5,190**	135	97.4
fesorcam.org	**4,700**	1,006	78.6
duchenne-spain.org	**4,450**	1,040	76.6
asociaciondoce.com	**4,280**	274	93.6
asem-esp.org	**4,230**	960	77.3

Table 5. Hit shrink average and median in hit ranges ($SERP_1$).

$SERP_1$	GOOGLE		BING	BING
	Average	Median	Average	Median
11 to 100	8.8	2.3	0.1	0
101 to 500	39.3	41.8	5	0.3
501 to 1000	70.4	72.9	N/A	N/A
1001 to 2000	74.4	78.4	58.2	55.9
2001 to 3000	86.1	87.8	69	64.4
3001 to 4000	92.1	91.7	79.5	78.4
4001 to 5000	86.6	86.6	81.5	78
> 5000	99	99	93.4	93.8

Table 6. Hit shrink percentage correlation against SERP values.

Variables	Hit shrink (%)
$SERP_1$ (Google)	0.898
$SERP_n$ (Google)	0.636
$SERP_1$ (Bing)	0.598
$SERP_n$ (Bing)	0.455

Table 7. Correlation (Spearman) between the number of hits according to SERP and Search Engine.

Variables	$SERP_1$ (G)	$SERP_1$ (B)	$SERP_n$ (G)	$SERP_n$ (B)	URLs (M)
$SERP_1$ (G)	**1**				
$SERP_1$ (B)	****0.81**	**1**			
$SERP_n$ (G)	****0.90**	****0.81**	**1**		
$SERP_n$ (B)	****0.80**	****0.96**	****0.81**	**1**	
URLs (M)	0.16	0.14	0.18	0.10	**1**

Note: The values in bold are different from 0 with a significance level $\alpha < 0.01$

Table 8. Correlation matrix between Web Impact Factor (WIF) variants.

	WIF-1	WIF-2	WIF-3	WIF-4	WIF-5	WIF-6	WIF-7	WIF-8	WIF-9	WIF-10	WIF-11	WIF-12
WIF-1	**1.0**	0.3	0.1	0.7	0.2	0.0	0.8	0.2	0.0	0.8	0.2	0.0
WIF-2	0.3	**1.0**	0.8	0.1	0.9	0.7	0.1	0.9	0.7	0.1	0.9	0.7
WIF-3	0.1	0.8	**1.0**	0.0	0.7	0.9	0.0	0.8	0.9	0.0	0.8	0.9
WIF-4	0.7	0.1	0.0	**1.0**	0.3	0.2	0.7	0.1	0.0	0.9	0.2	0.1
WIF-5	0.2	0.9	0.7	0.3	**1.0**	0.8	0.0	0.8	0.6	0.2	0.9	0.7
WIF-6	0.0	0.7	0.9	0.2	0.8	**1.0**	0.0	0.8	0.9	0.1	0.8	1.0
WIF-7	0.8	0.1	0.0	0.7	0.0	0.0	**1.0**	0.2	0.1	0.8	0.1	0.1
WIF-8	0.2	0.9	0.8	0.1	0.8	0.8	0.2	**1.0**	0.8	0.1	0.9	0.8
WIF-9	0.0	0.7	0.9	0.0	0.6	0.9	0.1	0.8	**1.0**	0.1	0.7	0.9
WIF-10	0.8	0.1	0.0	0.9	0.2	0.1	0.8	0.1	0.1	**1.0**	0.2	0.1
WIF-11	0.2	0.9	0.8	0.2	0.9	0.8	0.1	0.9	0.7	0.2	**1.0**	0.8
WIF-12	0.0	0.7	0.9	0.1	0.7	1.0	0.1	0.8	0.9	0.1	0.8	**1.0**

WIF-1: Sites (Alexa) / $SERP_1$ (Google); WIF-2: Sites (Majestic) / $SERP_1$ (Google); WIF-3: Ext. Links (Majestic) / $SERP_1$ (Google); WIF-4: Sites (Alexa) / $SERP_1$ (Bing); WIF-5: Sites (Majestic) / $SERP_1$ (Bing); WIF-6: Ext. Links (Majestic) / $SERP_1$ (Bing); WIF-7: Sites (Alexa) / SERPn (Google); WIF-8: Sites (Majestic) / SERPn (Google); WIF-9: Ext. Links (Majestic) / SERPn (Google); WIF-10: Sites (Alexa) / SERPn (Bing); WIF-11: Sites (Majestic) / SERPn (Bing); WIF-12: Ext. Links (Majestic) / SERPn (Bing)

Table 9. Correlation between Web Impact Factor variants and Citation Flow and Trust Flow.

Variables	Citation Flow	Trust Flow
WIF-1	0.1	0.1
WIF-2	**0.5	**0.4
WIF-3	**0.6	**0.5
WIF-4	-0.1	-0.1
WIF-5	**0.3	0.2
WIF-6	**0.4	**0.3
WIF-7	0.1	0.2
WIF-8	**0.6	**0.6
WIF-9	**0.6	**0.5
WIF-10	0.0	0.1
WIF-11	**0.4	**0.4
WIF-12	**0.5	**0.4

Note: The values in bold are different from 0 with a significance level $\alpha < 0.01$

WIF-1: Sites (Alexa) / Serp1 (Google); WIF-2: Sites (Majestic) / Serp 1 (Google); WIF-3: Ext. Links (Majestic) / Serp 1 (Google); WIF-4: Sites (Alexa) / Serp 1 (Bing); WIF-5: Sites (Majestic) / Serp 1 (Bing); WIF-6: Ext. Links (Majestic) / Serp (Bing); WIF-7: Sites (Alexa) / Serp n (Google); WIF-8: Sites (Majestic) / Serp n (Google); WIF-9: Ext. Links (Majestic) / Serp n (Google); WIF-10: Sites (Alexa) / Serp n (Bing); WIF-11: Sites (Majestic) / Serp n (Bing); WIF-12: Ext. Links (Majestic) / Serp n (Bing)

Preprint of article accepted for publication in ASLIB Proceedings
https://doi.org/10.1108/AJIM-10-2017-0226

Table 10. Correlation (Spearman) between SERPs (2016 – 2017).

	2017 $SERP_1$ (G)	**2017 $SERP_1$ (B)**	**2017 $SERP_n$ (G)**	**2017 $SERP_n$ (B)**
2016 $SERP_1$ (G)	0.877	0.728	0.839	0.742
2016 $SERP_1$ (B)	0.690	0.750	0.748	0.724
2016 $SERP_n$ (G)	0.809	0.727	0.904	0.786
2016 $SERP_n$ (B)	0.710	0.698	0.751	0.729