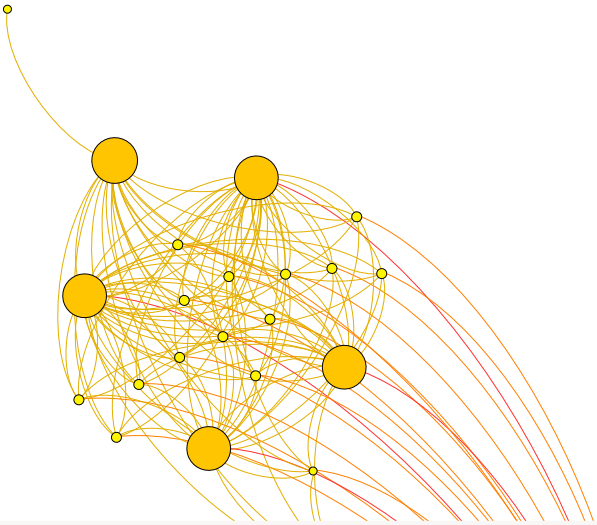
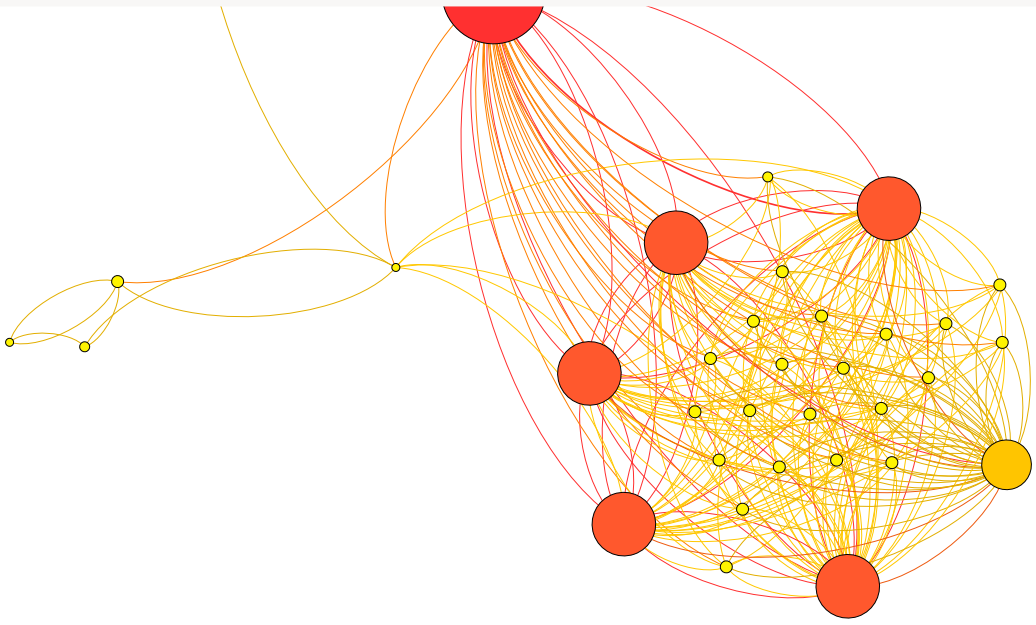
Technical SEO for web developers



A reference guide of technical SEO for software developers

Rubén Martínez



{ paradigma

SEO for Web Developers

Table of Contents

Introduction

The goal of this document is to be a quick reference handbook for web developers, either back and front-end ones. It should also help demystify the some of the many stereotypes about SEO.

This book will help you and your clients speak the same language with each other and with in-house or consulting specialists.

This eBook is a work in progress. This is the 2nd version with a deep revision of the first edition published in 2013. The most recent version of this book is available to download at paradig.ma/ebook-SEO

This work is distributed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. You must give appropriate credit to Paradigma and the author. You should not use the material or part of it for commercial purposes.

We are looking for translators of this eBook. Please contact us for translations to other languages.

Disclaimer

The examples on this document are provided for illustration purposes only and in good faith. The author does not endorse or otherwise the merits or lack thereof the websites and tools mentioned on this eBook.

About the author

Rubén Martínez is a marketer with a vast experience in international and multilingual SEO. Rubén learned the basics of online marketing while launching his start-up in London, United Kingdom.

Later on, as a team member of another start-up, Lokku, Rubén contributed to the growth of Nestoria, a smart property search engine launched from London to 9 countries in 6 languages.

Rubén deals with all things inbound marketing, analytics, SEM and SEO at Paradigma in Madrid, Spain. Paradigma is a 150 strong Big Data and web software development company offering innovative technology for business world-wide.

Special thanks

Many people contributed to this eBook in a way or another, not the least by asking great questions or by patiently answering mine.

Oscar Méndez at Paradigma accepted my proposal to write this eBook in the first place. María Arana, Mike Astle, Juan Cantero, Marc Tobias Metten and Gonzalo Alamar and a few others helped to turn my notes into this eBook.

What is SEO?

SEO stands for Search Engine Optimization. **SEO is everything that helps a website generate more revenue by converting traffic from search engines into leads or purchases**.

SEO is traditionally identified with the techniques that help improve the rankings of web pages on Google – this was just one of the visible effects of SEO.

The basics of SEO can be applied not only to generic search engines like Google, Baidu or Yandex but also to vertical ones like Indeed or Yelp, social networking services like Facebook or LinkedIn and to virtually all repositories of content with search engine functionalities.

This eBook focuses on SEO for Google because most of users have a strong preference for Google, not only as a generic search engine, but also as their gateway to the Internet.

E.g. when a user thinks about checking a movie on the Internet Movie Database website, he or she will often just write "*imdb"* on Google and click on the first result rather than directly typing the domain name and extension "*imdb.com*" in the address bar of their browser.

Which are the differences between technical and off-page SEO?

Two SEO approaches are required to drive users from search engines to websites: technical SEO and off-page SEO.

Technical SEO is everything related to a page and a website that is under the direct and usually immediate control of web developers and webmasters. This document is focused on technical SEO almost exclusively.

Off-page SEO is everything external to the development of a website like content marketing, link building and social sharing, which are not under the direct control of developers and webmasters.

In the early days of Internet, search engines simply did not use links as a ranking factor. Websites managed to show up on search engines' results

just by minding basic on-page SEO guidelines, like inserting titles on their HTML documents.

In addition to on-page and off-page SEO, content marketing is the other essential component for a sustainable and profitable online business.

Why is SEO important?

Being visible *and* ranking high on Google results not only in inbound traffic, but also in trustworthiness, authority or empowerment of prescription for websites and businesses of all sizes, markets and languages.

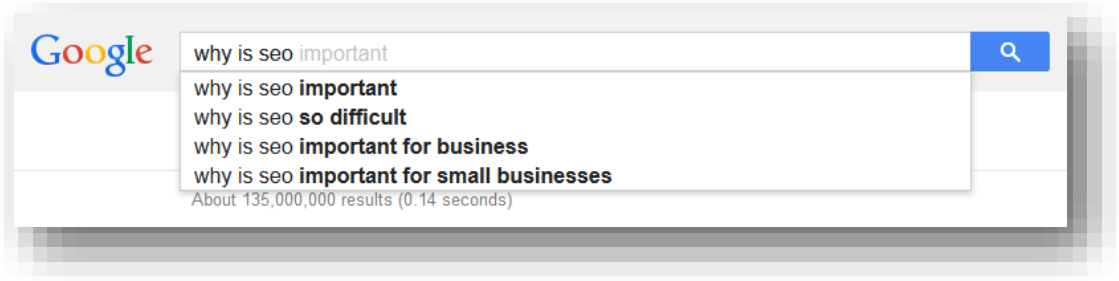


Figure. Screenshot of Google search box for the query "why is SEO". The search engine suggests auto-completions of related queries

Washingtonpost.com is a news website. It enjoys the massive awareness and brand reputation of its offline precursor, The Washington Post. You might think that the newspaper does not "need" Google for its business. The website however actively helps Google find all of the content by posting a comprehensive and updated sitemap.

The file http://www.washingtonpost.com/robots.txt includes the line:

```
Sitemap: http://www.washingtonpost.com/web-sitemap-
index.xml
```

By pointing the crawling bots to its sitemap, The Washington Post is investing in their SEO for profit.

Is SEO *free*?

SEO is definitely not free. It is however hugely cost-effective in comparison to any other investment in the marketing mix.

In addition to getting good training and reading books like this one, we generally advise to work with professional SEOs and digital strategists to save time and resources later in the lifetime your project.

Long term success in search marketing requires best practises, experience and intuition. SEO has a reputation of being a trade restricted to a few in the know. This stereotype is rooted in the fact that, so far, there is no tool, automated method or machine learning approach that manages to squeeze all the value of all the SEO developments. That is why you will not find "SEO tricks" on this eBook.

Organic traffic flows in when great content meets good SEO. **Content is King** - as long as it is fresh, relevant or engaging. SEO just makes it easy for bots to find, index and rank websites with the right content. However even the best content needs to be published efficiently so that search engines find it and deal with it.

The conclusion is that good SEO requires experts and content, neither of which come cheap, but it generates potentially massive amounts of traffic with high rates of conversion over the long term.

Google's official stance on SEO

While many affiliates and some SEOs are known for trying to systematically out-smart search engines with short term tactics, the best-practice SEO requires patience, experience, good relations within the search industry, great communication skills and, above all, an avid curiosity.

Google recently claimed that "*Many SEOs and other agencies and consultants provide useful services for website owners*"[i].

The relationship between Google and the marketing industry is rich and complex. Google communicates regularly with the SEO industry and

provide tools, posts on forums by Google employees, videos[ii], etc to webmasters and marketers.

Jargon buster

These terms will help you understand some of the concepts used in this eBook. We list them in alphabetical order and some of them are a bit abstract but please soldier on:

- **AJAX** is a number of techniques to create client-side asynchronous web applications brought together by JavaScript. Rich features using AJAX are popular because they usually improve the user experience by efficiently refreshing content. AJAX is however an issue for Google's crawler because it cannot read its content. Google expects that developers carry out some extensive hacking to deal with AJAX (see section "Make AJAX crawl-able" below).
- **Backlinks** or inbound links are links from external websites pointing to another website as opposed to internal links from pages on one website to other pages on the same site. Backlinks are not to be confused with the links to search results on Google SERPs.
- **Corpus** is a collection of documents in a machine-readable format, usually text. Examples of corpora (plural of corpus) are dumps of databases of any nature and format, the scraping of a website or any number of websites, etc.
- **Crawler** or web spider is a bot programmed to browse systematically websites for the purpose of indexing, like Googlebot
- **CTR** stands for Click Through Rate or clicks on a search result divided by the number of impressions (or how many times it showed on any SERP)
- **Document** is a piece of text or rich media that can be accessed and stored individually. An example of a document is a webpage or a downloadable pdf file.
- **Graph** is the interconnection between documents (vertices) by edges (links). An example of a graph is the network of websites linked with each other.

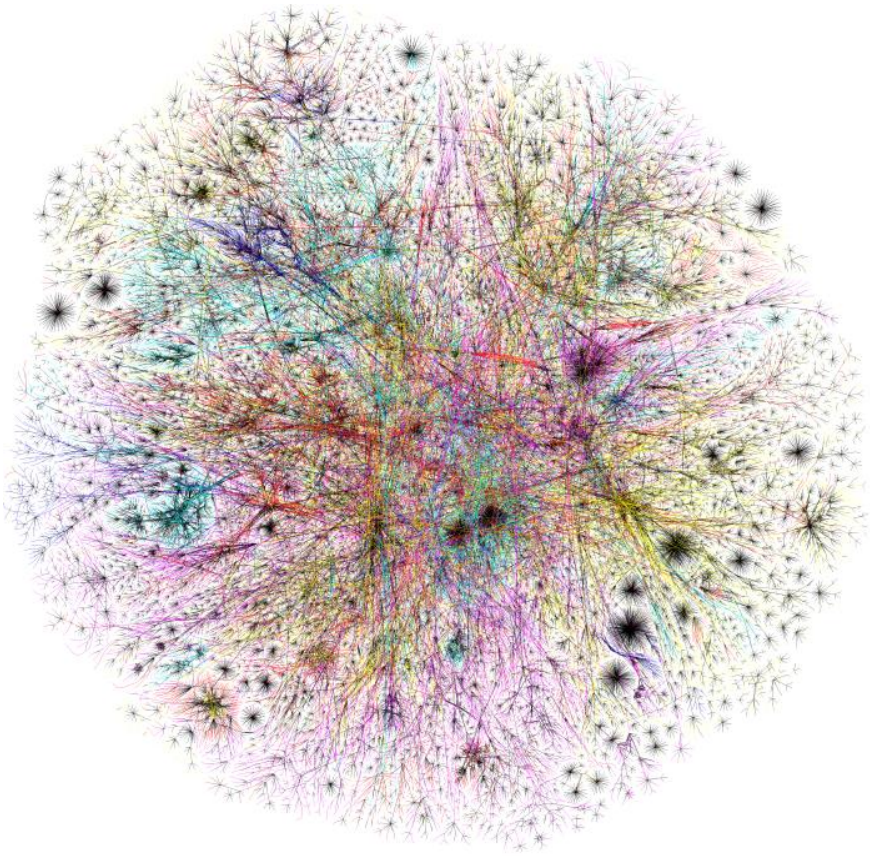


Figure 1 of a graph of 5 million edges and an estimated 50 million hop count by the Opte project modelling the Internet in 2003. Colours modified from the original representing different regions of IP addresses like Asia Pacific, Europe, North and South America, etc.

The concept of graph is a key concept in SEO. Many projects of new websites usually start their life as the output of a number of functionalities and ad-hoc extensions, rather than a body of interconnections. Search engines and SEOs however think of websites and actually the entire Internet as a graph.

- **Information Architecture** (abbreviated as IA) is the organization of documents and their connections. Websites are, in terms of IA, dynamic and connected structures of bot-readable content. Technical SEO is mostly applied IA for search.
- **PageRank** is a metric used by Google to determine the importance of an element (e.g. document, graphs or parts of them). It is one of more than 200 factors used to determine rankings of search results on Google. SEOs tend to prefer concepts like link juice or authority and new metrics to Google's PageRank.

- **SEOs** or search marketers are professional practitioners of SEO.
- **SERP** is the acronym of Search Engine Results Page or the list of links to results that search engines return in response to a user's query, e.g. http://www.google.com/#output=search&q=serp
- **Silos** are groups of subject-specific content on websites, e.g. categories separated as tree or sub-categories and detail pages. Vertical silos are frequent in tree structures where category pages are linking down to sub-category pages. The webpages under silos are hardly linked with webpages of other silos, i.e. there are few or no transversal or cross links across silos.

SEO deals with the bottlenecks in the search flow

SEO exists because people think and write in a different way from how search engines work.

There are a number of components in the search process that result in bottlenecks in the flow of information. The bottlenecks are inefficiencies that may result in a poor match between the search intent of the user and the purpose of the author or publisher.

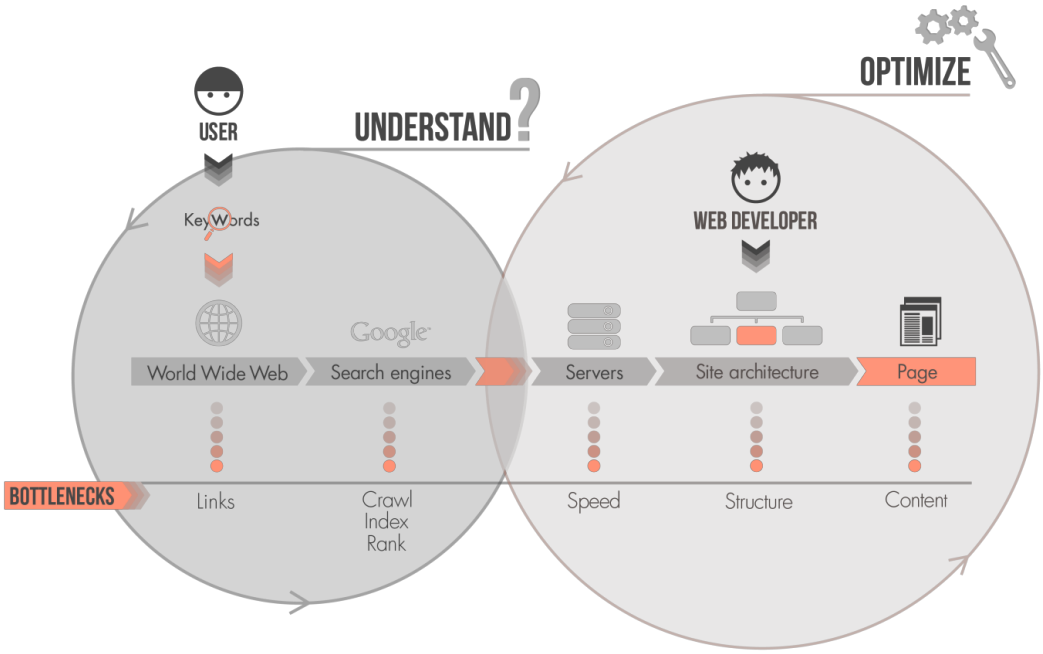


Figure 2 the flow of search is represented on the diagram above from the left (users) to the right (content). There are a few bottlenecks that affect the efficiency of the search.

SEOs can only try to **understand**, but not influence, the systemic bottlenecks: from the true meaning of keywords and search intent to Google's limitations.

Web developers and SEOs can **optimize** or have a direct control over the rest of bottlenecks down the flow: mostly speed, structure and the content itself.

Bottleneck 1: Limitations of keywords

Meaning of keywords

Many words convey different meanings. This is a challenge both for search engines and for SEOs.

E.g. "*Metro*" is a word that, when combined with the name of a location as in "*metro {location}",* might mean different searches in Europe, Canada and US:

- a local railway transport
- a local section of a newspaper or local news paper
- a brand of food stores

Search intent

There are a few possible search intents:

- **Informational:** wikis, news, blogs and publishing sites
- **Navigational:** video hosts, social networks
- **Commercial:** informational search with future transactional implications, e.g. vertical search engines, classified aggregators, price comparison sites
- **Transactional:** retail, e-commerce sites

Google tries to estimate search intent from users' previous activity and context but it is a formidable challenge for a generic search engine.

Changes in business models are often followed by SEO adjusting to different search intent than the one previously targeted.

E.g. When paywalls are introduced in newspapers, their SEO adjusts from informational intent (advertising intermingled with content) to transactional intent (content only with restricted access).

Google search features

The average number of words per query by users increased steadily from 2+ to 4+ search terms (excluding stop words) in the last few years. We users educated ourselves to write longer queries.

Google recently started to alter the search behaviour of users by limiting the long tail of search queries by tapping into the search data of users:

- **Auto-complete search suggestions:** Google displays suggestions that might be related to the one you are typing. This influences your query by modifying it or by accepting the one Google suggests instead of the one you originally intended to write.
  E.g. when you type just “bank”, Google suggests
  Bank of America
  Bank
  Bank of Scotland

- **Google Instant:** Google updates its SERP with different results as you type your query in the search box. This deters many users from typing queries longer than three or four search terms because they are presented with results earlier on.

Search engines’ bias

Brands, media titles, universities and institutions are said to be over-represented on Google SERPs. Google probably deals with entities in their graph and training dataset that might be equivalent to what we humans refer to as brands.

Google discretionary classifies reputation and authority of those entities according to their undisclosed criteria. Google adjusts their criteria logarithmically or with manual actions.

When Google releases multiple or significant adjustments at the same time, a significant amount of websites might be affected. These situations are known as updates. The best known recent updates are called Google Panda and Penguin.

Bottleneck 2: The World Wide Web & Search Engines

The World Wide Web (WWW) is an unstructured, de-centralized and ever changing ecosystem. Google tries to make sense of it with different types of software:

- **Crawling software**: bots are fast and greedy but blind to whatever is not text. Search engines try to cope with users creating incomplete, biased or inadequate connections between pieces of information. Google in particular is trying to cope with the surge of social networks.
  The quality of links as indicators of relevancy or popularity is also evolving: it is more convenient to link to content via social shares than with traditional hyperlinks on web pages.
- **Indexing software:** the main current challenges are identifying unique content and attributing authorship. Valuable content is often very duplicated across the WWW.
- **Ranking software:** Google claims to fight web spam with training data built for machine learning algorithms allegedly used for user-specific and session-specific rankings.

Bottleneck 3: Web servers, websites and code

The performance of websites in terms of access to findable content and downloading speed depends on two main factors:

- Site architecture: the distribution of the information in a structure of categories or sections organised by topic and other criteria
- Page speed: from the point of view of the crawlers, the code engineered for simplicity and speed is key for downloading the content from the server fast

When the speed of the web servers or the availability of networks is poor, the flow of information gets disrupted.

Bottleneck 4: The content itself

The publication of the content can itself be a bottleneck. SEO optimizes the IA of the content to match the requirements of search engines:

Mark-up: Poor or missing tagging frequently leaves search engines to classify content on their own without the help from its publishers.

Format: Search engines are good at interpreting text only. They only get the mark-up of images and videos – if it exists at all. All the client-side interaction with AJAX, Flash content, etc. are totally lost to search engines.

Duplication: If there are duplicated versions of the same content in multiple documents, search engines have trouble identifying the original document or even the first to be detected.

Attribution: Attribution of authorship is, like uniqueness, a hard issue for search engines to deal with.

SEO for new websites

Web developers that pay attention to on-page SEO in the early stages of new projects usually save lots time once in production. Critically, a well thought-through site architecture and prioritization of conversion funnels scale up beautifully.

We list below those SEO techniques that you need to consider in loose chronological order:

Why is technical SEO relevant?

Nowadays on-page SEO remains an important technique in the online marketing toolbox because:

1. Web servers and search engines only have in common the fact that they are extremely fast but essentially dumb machines. The technical SEO helps close the gap between both systems.
2. Search engines fall short of the expectations of users: it is very hard to determine search intent of a query, never mind matching it with the purpose of the content.

SEO helps close the gap between software and users and minimize their limitations.

Hosting

Host your site on reliable servers with excellent connectivity

You need a server uptime of 99.9% or higher over any period of time and as much bandwidth, memory and processing power as it takes. The good news is that all of the infrastructure costs keep dropping in price over time.

Measure the number of hops from your LAN the host of your website. If you are using a well interconnected local ISP, chances are that the number of hops that you are measuring is not too different from the number of hops that sets apart the Googlebot from your server.

The command *traceroute* (*tracert* on Windows) displays the path and transit delays of packets across a given route at a point in time.

E.g.

```
$ traceroute github.com

Tracing route to github.com [207.97.XXX.XXX] over a maximum of 30
hops:

  1     1 ms     1 ms     1 ms  XXX.XXX.XX.X
  2     2 ms     4 ms     2 ms  XX.XX.X.XX [XX.XX.X.XX]
  3     1 ms     1 ms     1 ms  XX.XX.X.XX
  4     1 ms     1 ms     1 ms  XX.XX.X.XX
  5     3 ms     3 ms     3 ms  mad-b1-link.telia.net [213.248.93.21]
  6    32 ms    30 ms    30 ms  prs-bb1-link.telia.net [80.91.245.58]
  7   123 ms   123 ms   112 ms  ash-bb3-link.telia.net [80.91.251.98]
  8   117 ms   123 ms   111 ms  ash-bb1-link.telia.net
[80.91.248.161]
  9   217 ms   118 ms   123 ms  rackspace-ic-138625-ash-
bb1.c.telia.net [213.248.98.218]
 10   260 ms   213 ms   274 ms  vlan905.core5.iad2.rackspace.net
[72.4.122.10]
 11   118 ms   118 ms   124 ms  aggr301a-1-core5.iad2.rackspace.net
[72.4.122.121]
 12   118 ms   118 ms   118 ms  github.com [207.97.227.239]

Trace complete.
```

There were 12 hops between my computer and github.com (the IP addresses are obfuscated in the example above) at the time of executing the command.

If you are physically in your target local market (country) and you are connected to an incumbent telecom operator, it is safe to assume that most of your visitors will be at roughly the same network distance as you are from the host.

If you are offered two or more hosting locations by the same host and your users are evenly distributed worldwide, you may want to choose the location with the lower average count of hops as measured a number of times.

Check your neighbours in shared hosting environments

Use reverse IP services like http://reverseip.domaintools.com to find which other websites share your hosting server. You do not want to be

on IP addresses blacklisted for email spamming or carrying out illegal activities.

Hosting services with dynamic IP addresss

You are usually safe in terms of SEO using cloud hosting, Content Delivery Networks (CDN) systems, Amazon's AWS or load balancing services that change their IP address regularly. When IP addresses change often Google "*may crawl a bit more conservatively than if we've figured out that your server is sturdy enough to be crawled at higher rates*"[iii].

Unless you are running a very large website with a high frequency of publication of new content, the crawling rate should not be a problem in this instance.

Information Architecture

Google PageRank

While dismissed by many SEOs as irrelevant nowadays, it may still be a useful metric if only at the level of order of magnitude for comparison and contextual purposes (an order of magnitude greater on a base 10 is to be 10 times as large).

E.g. the ESA (European Space Agency) and the NASA both enjoy a PageRank (PR) of 8, while the Indian Space Research Organisation features a PR of 7 and the China National Space Administration a PR of 5. Please note that the PR uses a logarithmic scale so a PR of 8 denotes several times more weighting than a PR of 7.

PageRank is now just one signal in more than 200 metrics that Google include in their ranking algorithms. The values of PageRank take nowadays many months to be publicly updated by Google.

Design a lean site architecture

Restrict the number of categories to low numbers that make sense to your users, keep a flat hierarchy and make sure that your conversion pages (products, SKUs, etc) are as few links away from the home as possible.

Avoid a tree structure with complex, deep or duplicated branches. Usually, but not always, the home is the page with the highest PageRank of a website.

The PageRank is diluted with every level down the structure of categories and sub-categories. When there is duplication, the PageRank is wasted.

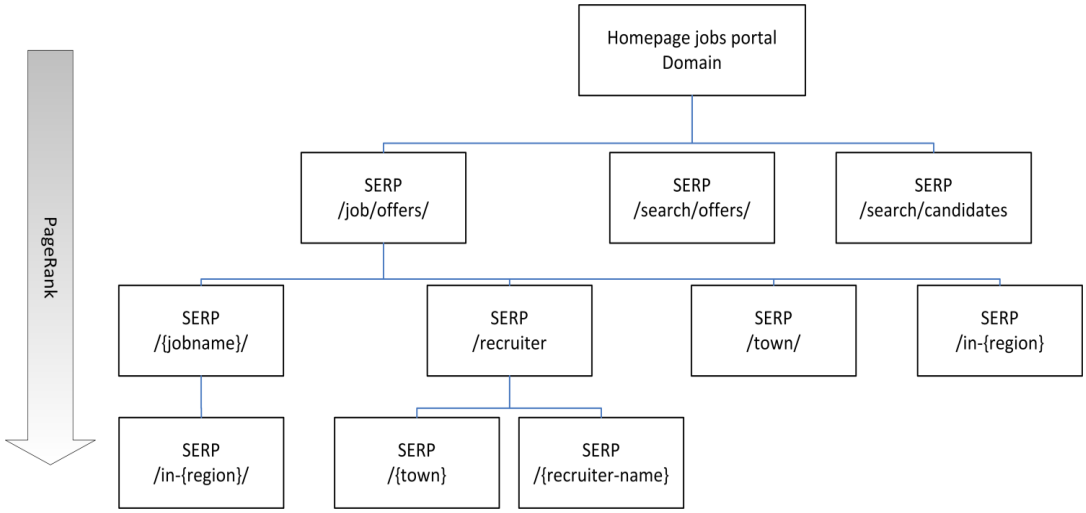


Figure 3 represents part of a (fictional) architecture with 3 levels below the homepage. The symbols curly brackets indicate variables. The type of page, SERP or detail, is indicated at the top of the box. The URL is composed by appending sections to the page. For instance the URL of the bottom left pages is domain/job/offers/{jobname}/in-{region}.

A simpler structure would consist in having only two levels below the homepage. It would be more effective to index the same inventory of content on Google. It also helps direct the user to the best converting webpages, namely detail pages of job vacancies in the example.

It is important to discuss this IA with all the stakeholders of a project in its very early stages. Business, usability, marketing and technical should align to the best architecture to fit all their different interests.

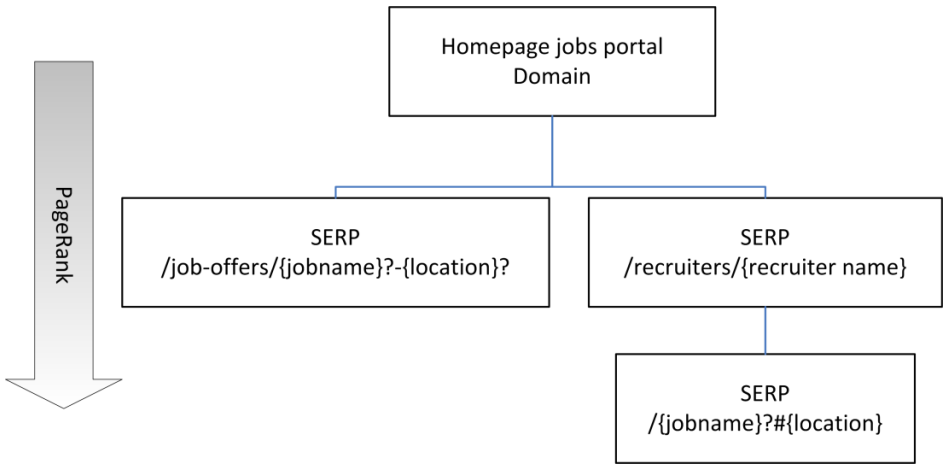


Figure 4 represents part of a lean architecture of 2 levels below the homepage. The symbol "?" indicates sections of the path that are facultative if there is a value for that variable. The symbol # is a character hash on the URL path.

Many product owners and usability consultants tend to "know" how they want to present their content to the end-users. Explain them the concepts above and try to set up an architecture that makes it easy to expand horizontally and vertically.

A successful architecture is often a trade-off between how a website should look, work and how search engines deal with the organization of information.

Link your internal pages sensibly

Densely connected graphs make crawling bots' job easier in principle because they can jump from node to node via multiple edges spreading through the whole network. Following this rationale, the more links you can insert on one page to the rest of the pages of the website, the better.

Avoid however linking to dozens of internal pages unless they are strictly relevant and genuinely useful to visitors. Search engines learned to discern the value of internal links by context, their topology and co-occurrence with other links on the same page. Just like users, search engines disregard any excess of links on-page (see Topology of on-page links below).

New websites have a PageRank value of zero at launch. If you already have a certain amount of PageRank, Google is only willing to crawl so much from your site. Administer your *PageRank budget*[iv] wisely: promote with internal links and backlinks only the pages that really matter to the business. Do not push this to the point of obfuscating links with JavaScript redirects or frames or of annotating them with nofollow attributes (those techniques are known as *PageRank sculpting*[v]).

Backlinks are, for Google, a stronger signal of authority and popularity than internal links. Just a reminder: this document is not covering the techniques of off-page SEO and link building.

Large websites with a long tail of pages do not usually get backlinks for each of them. Internal links are the next best thing you can do to position all those pages. Ideally your internal links should be organised consistently with the IA of the site. Decide the architecture first and then the internal links.

You want to link internally in a relevant way for both the users and Googlebot. It is OK to theme your website by topic organized in directories but beware that some techniques that concentrate links to and in between the pages of each directory are best left to technical SEO professionals who have experience with silos and their risks.

Topology of on-page links

In the early days of the WWW it was assumed that documents are essentially lists of hyperlinks connecting information. One of the proposals to determine the relative value of each document was published as the PageRank (for Larry “Page”, not for “webpage”).

The PageRank in its primordial version assigned each value to each link regardless of its position on the document.

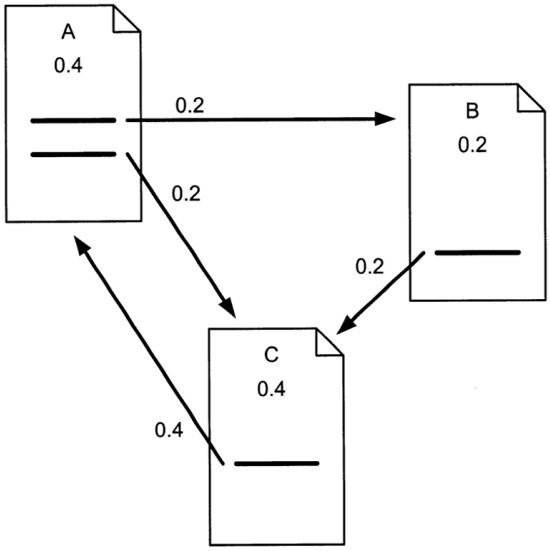


Figure 5 representing the flow of PageRank through documents via outgoing links in a "**random surfer**" model. See how the document A passes equal value of PageRank, 0.2, from its two links regardless of their position on the document. Source: Method for node ranking in a linked database[vi]

Google moved from a paradigm of *random surfer* to *reasonable surfer*.

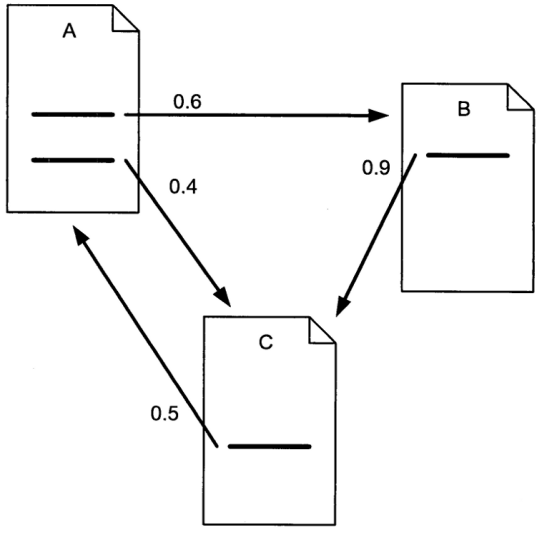


Figure 6 representing a flow of PageRank through documents via outgoing links in a "**reasonable surfer**" model. See how the document A passes more PageRank, 0.6, from the hyperlink in the middle of the document than the PageRank passed by the hyperlinks regardless of their position on the document. Source: Ranking documents based on user behaviour and/or feature data[vii]

Place the most relevant content and links at the top of your page. Use meaningful anchor texts.

If your layout includes many navigational links, like breadcrumbs, menus or facets with nested menus try not to place them all on top of the page.

Unfortunately for us users many websites still list dozens of deep links at the bottom of the page. Those stacks of hyperlinks are illegible at best and a poor SEO practise in all cases.

Responsive websites

Browsers build the DOM of a web document, parse its tree to render its HTML and then interpret the CSS.

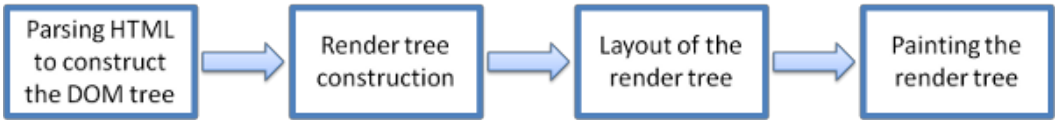


Figure: Rendering engine basic flow. Source: HTML5rocks.com

The front-end development of responsive websites almost always requires to displace content with CSS to adapt it to screens of many geometries and interactions.

Try to follow the reasonable surfer model in placing code and rendered content in the same order. Do not hide links, content or menus on different versions of the responsive design: just position them sensibly.

You are on the safe side in terms of SEO as long as the bot parses the content as you placed it on the HTML source code of the page - even if it is not exactly in the order we humans read it on all devices.

Number of internal links

Help users find their way around with navigation facets, breadcrumbs, menus and relevant pages. Restrict the number of internal links on each page to a reasonable number.

There is not any clear-cut maximum and minimum limits of links; they depend on the nature of the business or on the value proposition of the website.

It is commonly accepted that 100 links or more on a page (either internal links or outgoing ones) are, as a rule-of-thumb, an excessive number of links.

E.g. enter this on your bash terminal after you have w3m installed

```
$ w3m "http://www.king.com"
```

By pressing “L” on w3m you will get a list of all links, anchors and images of the current page. At the time of writing this guide we counted 373 links on a page. That looks like a few hundred too many! Are they all relevant? The homepage of King.com is designed as a catalogue so such a high number of links are to be expected.

The homepages of some e-commerce sites may need very few internal links pointing to what really matters: trial, support and checkout.

E.g. enter this on your bash terminal and press “L”

```
$ w3m "http://www.perspectivemockups.com/"
```

We counted just 3 links on that page. Those are probably are all the links that the website needed for business.

Alternatively, you may try to get a full list of links of a small website – do not try the following one-liner on a large website:

```
$ MYSITE='http://www.website.com';wget -nv -r --spider
 $MYSITE 2>&1 | egrep ' URL:' | awk '{print $3}' | sed
                "s@URL:${MYSITE}@@g"
```

The shell will go quiet for a while until it lists all the links it found on-page.

Prevent broken links before going to production

Identify broken links, internal and outgoing with regular crawls and fix them. It is easy to do it with just by fetching documents from the net directly from a shell.

E.g. the Library for WWW in Perl has a bash command, lwp-request. If you are using Perl on cygwin for Windows, you may have to install HTML-Tree first so enter this on the shell:

```
$ perl -MCPAN -e shell
```

```
cpan[1]> install HTML::TreeBuilder
```

Then enter this one-liner back at the shell prompt[viii]:

```
lwp-request -o links http://www.bigdataspain.org | perl -
pe 'chomp; $_ =~ s{ ^ \w* \s* ( [^#]+ ) .* $ }{$1}x; undef
$_ unless m/^http/; $_ = qq{\"$_\"\n} if $_' | sort | uniq
| perl -ne 'chomp; print $_ . qq{\t} . qx{lwp-request -ds
$_}'
```

The (long) one-liner results in this output:

```
"http://bigdataspain.us5.list-
manage.com/subscribe/post?u=d86f829047abbeaff0d30973f&id=8
832bb34a7"      200 OK
"http://fonts.googleapis.com/css?family=Days+One"
200 OK
"http://microformats.org/profile/hcalendar"     200 OK
"http://www.bigdataspain.org/"  200 OK
"http://www.bigdataspain.org/2012/"     200 OK
"http://www.bigdataspain.org/2012/Terms-and-Conditions-
Big-Data-Spain-conference.pdf"   200 OK
[…]
"http://www.bigdataspain.org/en/infographic-conference-
2012"    200 OK
"http://www.bigdataspain.org/jobs/en/"  200 OK
"http://www.flickr.com/photos/87479166@N04/"    200 OK
```

There are no broken links to report in this example. There are plenty of online checkers that are fast and reliable also.

Link attribute rel *nofollow*

The *nofollow* value of the rel attribute is great to control the crawling of specific pages[ix].

The *nofollow* value helps Googlebot understand that it should not crawl the linked document. In fact, the bot does crawl as much as it can for the given PageRank of a website, including nofollow-tagged pages.

Google claims that *nofollow* links do not pass PageRank and that their anchor text is skipped altogether.

This is useful for IA purposes when you try to avoid indexing duplicate content on Google or to indicate Google that a link with rel="*nofollow*" is a paid link or an external document that you do not necessarily endorse.

This attribute was popular for PageRank sculpting [X] and waterproofing silos. Those practices are not recommended.

Configuration of mobile rendering

There are basically three approaches to SEO for mobile web (as in webpages for interactive screens of smartphones and tablets):

1. Sites that serve all devices on the same set of URLs, with each URL serving the same HTML to all devices. CSS just changes how the page is rendered on the device. If all Googlebot user agents are allowed to crawl the page assets (CSS, javascript, and images), Google's algorithms automatically detect responsive designs.

2. Sites that dynamically serve all devices on the same set of URLs, with each URL serving different HTML (and CSS) depending on whether the user agent is a desktop or a mobile device.

3. Sites that have separate mobile and desktop URLs ej. http://m.website.com and http://www.website.com respectively

The first approach – the responsive web design – saves resources and it is the one preferred by Google [x]

Uniform Resource Identifier URL

URI Syntax

URI stands for Uniform Resource Identifier. URLs or Uniform Resource Locators are just a subset of URIs describing the primary access mechanism.

Eg. URLs http://www.example.org:80 or mailto:me@example.org

```
URI         = scheme ":" hier-part [ "?" query ] [ "#" fragment ]

Example:

          foo://example.com:8042/over/there?name=ferret#nose
          \_/   \______________/\_________/ \_________/ \__/
           |           |            |            |        |
        scheme     authority       path        query   fragment
           |   _____________________|__
          / \ /                        \
          urn:example:animal:ferret:nose
```

Compose a simple URL path

Web browsers request webpages from servers by using Uniform Resource Locators (URLs). Write URLs paths as similar to natural language as possible. Try to use keywords as a user would read them.

E.g. Majestic is a wine retailer in the UK. In 2013 the URLs of their SERPs were for instance
http://www.majestic.co.uk/find/category-is-Wine/category-is-Spain
From the point of view of a user the URL reads unnatural. In terms of SEO, the strings of text “category-is” in the URL path are probably redundant.

In 2014 their equivalent URL is http://www.majestic.co.uk/spanish-wine which is more legible and relevant.

Less is more about URL paths: try to limit the number of folders with categories and subcategories to a maximum of three separated by forward slash. Separate keywords by hyphens instead of underscores or any other symbol.

URLs can theoretically be hundreds or even thousands of characters long (depending on the browser) and still be operational but optimal ones should of course be much shorter.

URL encoding

A URI is a sequence of characters from a very limited set: the letters of the basic Latin alphabet, digits, and a few special characters.

URLs can only be sent over the Internet using the ASCII set of printable characters. Eg. in German /München/ is best written as /muenchen/ or even /munchen/. Avoid punctuation symbols and blanks to compose URLs. They encode in an ugly, unreadable strings of text if encoded with percent-encoding in an URL.

E.g. The sentence in Spanish "*La eñe es una letra del alfabeto español, bretón y guaraní entre otros*" looks terrible on a URL with encoding:

```
http://www.domain.com/La%20e%C3%B1e%20es%20una%20letra%20del%20alfabeto%20espa%C3%B1ol%2C%20bret%C3%B3n%20y%20guaran%C3%AD%20entre%20otros
```

You want something like this, free from triplets including "%":

```
http://www.domain.com/la-ene-es-una-letra-del-alfabeto-espanol-breton-y-guarani-entre-otros
```

All programming languages have URL encode and decode functions to normalize characters in the URI.

Friendly URLs

Pay particular attention to reserved characters[xi] . Reserved characters provide a set of delimiters to distinguish data within a URI.

```
reserved    = gen-delims / sub-delims

gen-delims  = ":" / "/" / "?" / "#" / "[" / "]" / "@"

sub-delims  = "!" / "$" / "&" / "'" / "(" / ")"
       / "*" / "+" / "," / ";" / "="
```

Restrict the delimitation with sub-delims and gen-delims as much as you can. You want to keep the number of parameters in queries to no more than 2. This is a ballpark numbers suggested by Google long time ago. It is a rule particularly applicable to sessions with logged in users or at ecommerce checkouts with multiple fields in their forms. Those are typically identified with symbols in URLs composed dynamically such as question marks "?" and ampersands "&" combined with equal to "=".

E.g. Zara is a fashion retailer whose website used to composed URLs that contained too many folders and parameters in the path:

```
https://www.zara.com/webapp/wcs/stores/servlet/ShopCartPag
                    e?calculationUsageId=-
    1&updatePrices=1&catalogId=24052&orderId=.&langId=-
              1&storeId=10701&URL=ShopCartPage
```

E.g. pingg.com is an online e-cards retailer with shorter, more intuitive URLs on its website. Their URLs are OK although they can be improved

```
http://www.celebrations.com/public_event/dw72fy6hz2sxyby8d
```

Client-side modifications of the content of a webpage that don't reach the server use fragment identifier preceded by a hash mark # (see section "Make AJAX crawlable").

```
    http://www.domain.com/section1/subsection2#fragment
```

Google standardizes URLs by removing any fragments from the URL, i.e. does not reach the optional last part of a URL for a document from the hash onwards. See the section "Making AJAX crawlable".

Prepare to scale up the URL paths. Internet allows for economies of scale so it is only natural that new projects should prepare to scale across languages, countries, categories and verticals.

There are no defined rules to deal with scaling. Check the Appendix Domain names for some considerations about scaling vertically and horizontally.

Automate the generation of URLs with intuitive rules

Some frameworks escalate the generation of URLs easily, e.g. the URL mappings plug-in on Grails[xii]

URL re-writing is a technique to reverse engineer the links from the URL mappings. Given a mapping:

```
static mappings = {
   "/$blog/$year?/$month?/$day?/$id?"(controller:"blog",
action:"show")
}
```

If you tag as follows:

```
<g:link controller="blog" action="show"
        params="[blog:'fred', year:2007]">
    My Blog
</g:link>
<g:link controller="blog" action="show"
        params="[blog:'fred', year:2007, month:10]">
    My Blog - October 2007 Posts
</g:link>
```

Grails will automatically re-write the URL in the correct format on HTML:

```
<a href="/fred/2007">My Blog</a>
<a href="/fred/2007/10">My Blog - October 2007 Posts</a>
```

Mark-up your content

Insert tags into the appropriate HTML elements to define textual content within a page.

Each webpage needs to be relevant to its content. After backlinks and relevant anchor texts, marking up your documents is the most effective technique to help Google correlate your documents with their content purpose.

The importance of copywriting relevant and informative titles is often underestimated. Good titles are important for the CTR of the links to your results on Google's SERPs. In some cases, titles help to associate your brand to a category or niche if you include your brand name.

Title Tag

Write descriptive titles less than 70 characters long in the head container of your HTML document.

```
<head>
<title>Not too many keywords here</title>
</head>
```

Unless you have a reason not to, include your brand name in the title.

```
<head>
<title>Keyword/s Category | Your brand</title>
</head>
```

Meta elements

It is usually trivial to generate, from a coding perspective, meta tags for webpages. In real life however you will code webpages with missing, duplicate, too long, too short or non-informative HTML meta elements so reserve some time to check everything before beta launches.

Content is the attribute of the meta elements that help SEO the most. The <meta> tags always go inside the <head> element.

Description attribute

Think of it as a tweet-long summary of the most relevant content on your page. It should help you improve the Click Through Rate of your results on Google's SERPs.

```
<meta name="description" content="This attribute is a
great opportunity to improve your CTR if Google picks it
up and inserts it as a snippet on its SERP" >
```

Do not exceed 170 characters when composing the text on the tag.

Robots attribute

The values {*index*,*noindex*} and {*follow*,*nofollow*} are directives that help you control the content that you want to index at page-level. They can be search engine-specific. These meta tags are placed in the HTML header.

They can be effective also in combination with the robots.txt protocol so you retain control on your content on Google.

```
<meta name="googlebot" content="noindex">
```

Language attribute

This helps search engine match the language of search queries and locales with your content.

```
<meta http-equiv="content-language" content="pt-br" />
```

Keywords attribute

Common wisdom is that stuffing keywords on this meta tag is "SEO". This practise was so abused that Google gave up using the keywords attribute altogether a few years ago.

You may still want to use it for your own purposes, like semantic tagging, etc but definitely not for SEO.

Headings

Organize the distribution of your content on-page with hierarchical criteria. Place the most important content on the top and the least relevant on the bottom of the webpage as the crawler (or a text-only browser) reads it.

<*h1*> to <*h6*> tags are used to define HTML headings.

```
<h1>Most Relevant Keywords</h1>
<h2>Second Most Relevant Keywords</h2>
<h3>Specific sub-sections Most Important</h3>
```

Tagging with headings does not need to be strictly nested if it takes too much of your time or the editors of the content cannot be briefed.

You can use more than one *h1* tag on the same page in HTML5. HTML5 is the latest standard for HTML.

Main and aside tags

HTML5 enables tagging content-specific elements like <section>, <main> and <aside>. Its impact on SEO is probably null still in 2014.

Rich media (images, videos)

Mark up your videos following schema.org's indications and provide *alt* descriptive and accurate attributes on images.

Not good

```
<img src="conference.jpg" alt=""/>
```

Acceptable

```
<img src="conference.jpg" alt="red dot on white
background"/>
```

Best

```
<img src="conferences.jpg" alt="the flat of Japan">
```

Canonicalization

Duplication of content often occurs when sites render the same content on different URLs. The software of popular Content Management Systems (CMS) like WordPress and Drupal tends to create duplicated content by way of categorisation (archive, pages, categories, tags) or formatting (web, printable, feed).

E-commerce sites may incur in duplication by indexing URLs with multiple parameters (session IDs, shopping cart and checkout parameters, etc).

Offer a unique version of your content: avoid content showing up in more than one page, and if it has to, use the metatag canonical to help Google find with is the important page for your business.

Declare the link attributes rel="*next*" and rel="*prev*" and rel="*canonical*" to help Google understand what are the pages that really matter to you and the user [xiii].

```
<link rel="canonical"
href="http://www.example.com/article?story=abc&page=2"/>

<link rel="prev"
href="http://www.example.com/article?story=abc&page=1&sess
ionid=123" />

<link rel="next"
href="http://www.example.com/article?story=abc&page=3&sess
ionid=123" />
```

You can also help Google handle duplicate content across domains [xiv] with the canonical tag. This tag is just an indication to Google, and it should be used if there is not any other, more efficient, ways to tackle duplication.

Anchor text

Use short, descriptive and relevant anchor texts on hyperlinks. Try not to stuff the anchor text with keywords. Anchor text used to be very important for ranking purposes but context and other techniques like Latent Dirichlet allocation (LDA) [xv] are gaining in weight.

If you link more than once to the same destination URL on the same webpage, only the anchor text of the first link is the one picked up by Google for correlation statistics.

Structured Data

Consider marking up individual pieces of text that can be itemised such as details of people, events, places, resumes for job boards, listings of classifieds, etc. Tagging microdata helps Google identify what is your content is about at very deep minutiae. There are 3 standards of structured data markup that you can follow with similar effects:

- Microdata Schema.org - a well-organized and documented HTML5 microdata
- Microformats http://microformats.org/ - simpler and earlier standard
- RDFa http://rdfa.info/

The tangible SEO benefits of tagging structured data are:

- you can get larger snippets on Google SERPs if you use smart microformats. Eg. hListing for classifieds sites, hResume for job boards, hReview
- you can rank higher on Google if you are organising an event with hCalendar

E.g. case of use of the Open Graph protocol to enable webpages to become a rich object in a social graph

```
<meta name="og:title" content="Big Data Spain 2012
conference - Madrid Nov 16th"/>
<meta name="og.summary" content="Timetable of the schedule
of the sessions of the Big Data Spain 2012 conference"/>
<meta name="og:type" content="website"/>
<meta name="og:url"
content="http://www.bigdataspain.org/en-
2012/program.php"/>
<meta name="og:site_name" content="Big Data Spain"/>
<meta name="og:latitude" content="40.452537"/>
<meta name="og:longitude" content="-3.726412"/>
```

```
<meta name="og:street-address" content="ETSI
Telecomunicaciones, Avenida Complutense n° 30, Ciudad
Universitaria"/>
<meta name="og:locality" content="Madrid"/>
<meta name="og:region" content="Madrid"/>
<meta name="og:postal-code" content="28040"/>
<meta name="og:country-name" content="Spain"/>
<meta property="og:locale" content="en_GB"/>
<meta property="og:locale:alternate" content="es_ES"/>
```

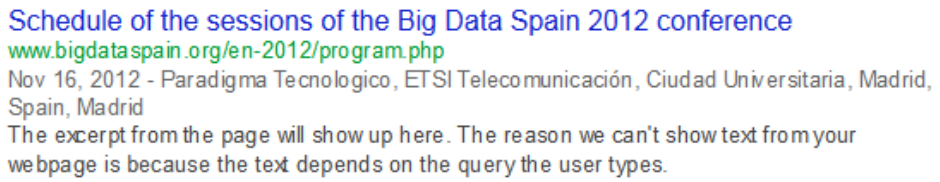


Figure 7 Snapshot of the key information that users would see right in Google's SERPs thanks to tagging content with the standards of hCalendar

Authorship

Google requests the help from webmasters to match content with its author[xvi]. This can be done by marking up links with rel="me" and rel="author" bi-directionally between the authors' profile page on Google+ and the webpage hosting the authors' content.

Google allows rel="author" markup through a <link> element in the head:

```
<link rel="author"
href="http://plus.google.com/117486480899235875959/"/>
```

Identified authors get an avatar picture displayed on Google SERPs, which is great for higher CTR and possibly rankings for some queries:

Figure 8 screenshot of the profile of the page of Google+ of an identified author

Robots.txt protocol

The robots.txt standard is quite popular among webmasters. It is only an advisory protocol to allow or disallow pages or sections to Googlebot.

Google complies with it for but it may still fetch the disallowed content. Depending on a number of circumstances it may display the URLs and titles of disallowed content on its SERPs. The difference between the results of disallowed pages on Google SERPs and allowed ones is that snippets are missing on the former.

If you site architecture is lean and flat, you might not need to make an intensive use of robots.txt. There are of course exceptions of large, complex or legacy websites. E.g. enter this on your bash terminal

```
$ w3m -dump "http://www.bbc.co.uk/robots.txt" | wc -l
```

```
User-agent: Googlebot
Disallow: /iplayer/episode/*?from=r*
Disallow: /iplayer/cy/episode/*?from=r*
Disallow: /iplayer/gd/episode/*?from=r*
Sitemap: http://www.bbc.co.uk/news_sitemap.xml
Sitemap: http://www.bbc.co.uk/video_sitemap.xml
Sitemap: http://www.bbc.co.uk/sitemap.xml
Disallow: /_programmes
Disallow: /606/
Disallow: /apps/cbbc
Disallow: /apps/flash
[…]
```

At the time of writing this, the output of the entire robots.txt was unusually long even for a website as large of the BBC: a count of 362

lines! Those many lines are probably the result of legacy directories. You have a chance at the start of each new project to have a simple robots.txt with a proper IA.

You can use regular expressions but test the syntax thoroughly on Google Webmasters Tools first if you do not want to inadvertently disallow or allow entire sections or website.

Monitor your site for hacked content [xvii]

Keep an eye on injected, added or hidden content and malware scripts and remove tjhe as soon as it appears. Google Webmasters Tools includes a section “Malware” to help deal with situations.

HTML, JavaScript, AJAX and CSS

Code for speed

As a general rule, you want to minimize the traffic between your server and any client browser. Analyse your webpages and diagnose their performance as you code.

Use third party tools like www.webpagetest.org or YSlow. YSlow's web page analysis is based on 23 rules that affect web page performance and are testable:

- Minimize HTTP Requests
- Use a Content Delivery Network
- Avoid empty src or href
- Add an expire or a cache-control Header
- Add gzip components
- Put stylesheets at the top
- Put scripts at the Bottom
- Avoid CSS expressions
- Make JavaScript and CSS external
- Reduce DNS lookups
- Minify JavaScript and CSS
- Avoid redirects

- Make AJAX cacheable
- Use GET for AJAX requests
- Reduce the number of DOM Elements
- Reduce cookie size
- Use cookie-free domains
- Do not scale images in HTML

Some programming languages allow you to minify HTML too. E.g. Twig is a template engine for PHP that features spaceless capabilities.

Debug for crawlers

Avoid crawling pitfalls in forms and redirect chains and loops, e.g. test thoroughly regular expressions in JavaScript.

Dynamic URLs are OK but avoid session IDs and tracking arguments on URL paths for webpages that do not require authentication because they interfere with how crawlers access websites and de-duplicate content.

Avoid cloaking

Cloaking[xviii] consists in serving different content to human users and to search engines. Technically it starts by identifying user agents by the IP address and then serving up different content on the server side. Legitimate coding tactics might inadvertently result in cloaking.

Examples of cloaking:

- Serving text to search engines while showing images or Flash to users for the same page or section
- Inserting text or keywords into a page only when the User-agent requesting the page is a search engine, not a human visitor

Make AJAX content crawlable

The asynchronous capabilities of AJAX allow to update part of the content of a webpage without having to reload all of it client-side. Since Google does not interpret JavaScript to get content on the fly the way a browser can, i.e. Google misses all the dynamic content loaded via AJAX.

Google are now asking webmasters to make their AJAX-based websites crawl-able [xix]. As we will see below, their requirements are a tall order for most webmasters – you need to invest some significant amount of resources to deal with them.

Google have a method to allow for dynamically created content to be visible to crawlers[xx]. The method is convoluted and intricate. Its number one requirement is to present AJAX URLs, that is, URLs containing a #! hash fragment. This requirement is contentious issue because exclamation marks happen to be illegal in HTML, XHTML, and XML URL identifiers and they make pages inaccessible to non-JavaScript-enabled browsers.

If the URLs of your AJAX-based webpages are

```
www.domain.com/ajax-based.html#key=value
```

you want to turn them to

```
www.domain.com/ajax-based.html#!key=value
```

so the crawler can modify the URL to temporarily be

```
www.domain.com/ajax-
based.html?_escaped_fragment_=key=value
```

The rest of the requirements are based on what Google calls the HTML snapshot mechanism.

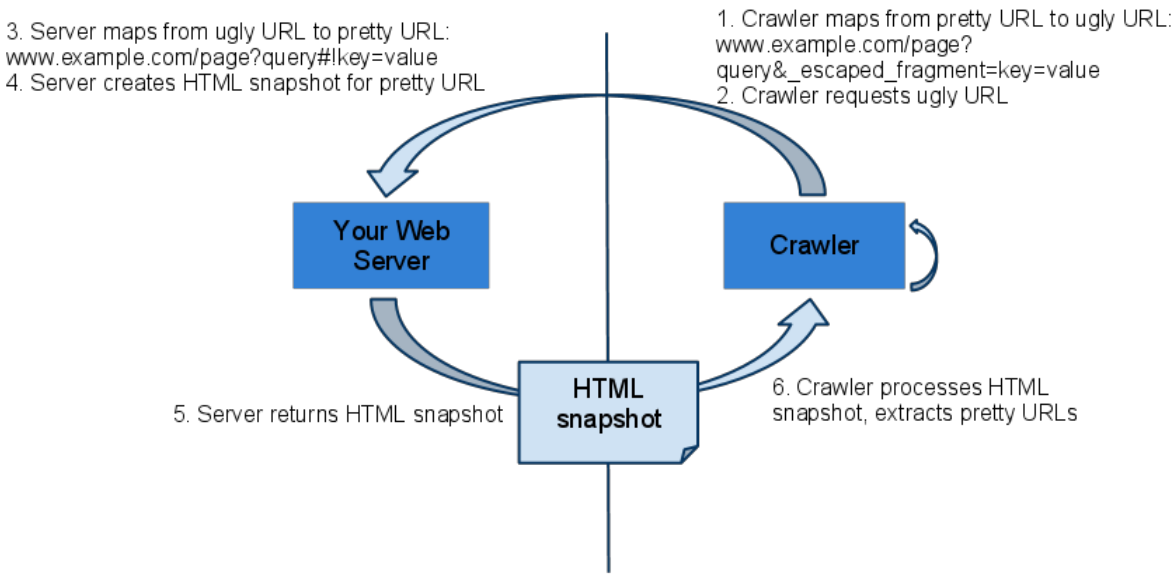


Figure 9 Google's HTML snapshot method to duplicate AJAX dynamic webpages to bot-readable content

Since many of your URLs might be just rendered as static HTML whose paths don't contain hash-identified fragments, you will have to include a new special meta tag in the head:

```
<meta name="fragment" content="!">
```

<noscript> tag for content on JavaScript

Use a <noscript> tag to ensure that the content contained in JavaScript can be read by Google. The text has to be exactly the same as what's contained in the JavaScript.

Using <noscript> this way is best practice although Google will not pass PageRank on pages linked from within the <noscript> tag or treat <noscript> content with the same value as the rest of normal text on-page.

```
<script type="text/javascript">
document.write("Out of sight, out of mind")
</script>
<noscript>Out of sight, out of mind</noscript>
```

Disable JavaScript on your browser to check the result.

Avoid frames and Flash

Search engines cannot read Flash and they often skip frames. Images, videos are challenging at best for search engines. Google has trouble dealing with dynamically generated DOMs whose URLs contain session IDs in the path.

Try to use readable text rendered on HTML as your preferred format. You can also get pdf, doc and txt files indexed on Google if you link to them.

Use a text browser like w3m or lynx to get a closer picture to what search engines "get" your content in the sequential order they crawl your webpages.

E.g. install w3m, a text-based browser that cannot handle JavaScript. You can find at Sourceforge. Enter this on your bash terminal

```
$ w3m -dump "http://www.ft.com/" | less
```

Inspect the output and compare it with the output of the parsed HTML on your browser. Is the text sorted as you can see it on a browser?

Avoid using CSS to hide text

CSS Image Replacement places some text off-screen when calling a logo image or similar from a CSS stylesheet. E.g. a CSS class with the text moved way off any screen size using the -9999px hack.

```
h1.mylogo {
width: 200px; height: 60px;
background: url(/images/logo.jpg);
text-indent: -dpx;
}
```

Invoking the class "hides" any piece of text, in the example the one wrapped by h1:

```
<h1 class="mylogo">The name of my company</h1>
```

There are legitimate reasons to use the technique, accessibility being one of them e.g. text to speech and browsers with images or JavaScript disabled for instance. The use of the alt label may be more appropriate in those instances.

There are other shortcuts to hide text, such as matching the colour of the font and the background or setting the font size to zero. Google learned to deal with them. Specifically CSS image replacement is a contentious issue when it comes to SEO. Check with an expert before you implement it.

Generate sitemaps

Help Google find your content when you launch a new website or when you implement significant changes to it by supplying comprehensive and updated HTML and XML sitemaps. Set up and follow a simple policy to let Google know when you renew your content.

HTML sitemaps are best for the SEO of small websites. They are impractical for large websites: users and bots are not expected to go through long lists of backlinks on-page. XML sitemaps are effective for large inventories of content that rotates regularly.

HTML sitemaps

HTML sitemaps are more important for users than for SEO. An HTML sitemap should link only to the most relevant pages of a website. It is tempting to link to all your inventory of documents from a HTML sitemap. This usually results in pyramidal architectures.

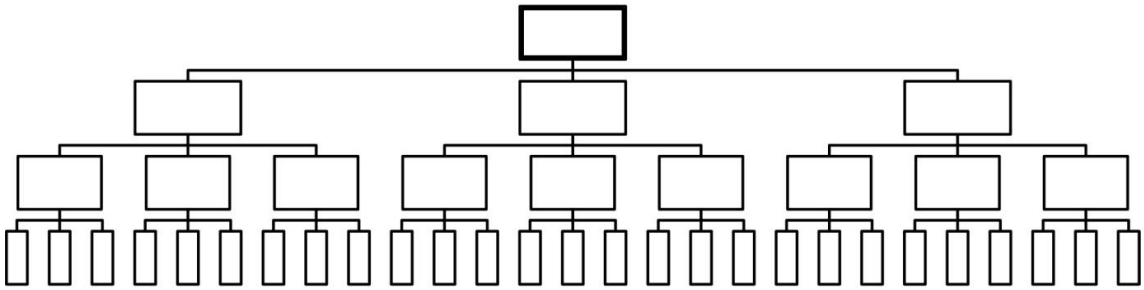


Figure 10 diagram of pyramidal structure of a HTML sitemap listing links to deer pages. Each webpage contains 3 links. The results in 39 pages linked from the HTML home page.

This practice is detrimental to the user experience and ineffective with search engines.

E.g. Indeed.com links to all their individual profile pages from a wide and deep tree of linked pages

Home of the HTML sitemap

```
http://www.indeed.com/resumes/directory/
```

Top category page

```
http://www.indeed.com/resumes/directory/1
```

Sub-category page

```
http://www.indeed.com/resumes/directory/1-1
```

Detail page

```
http://www.indeed.com/r/2e24d4a2aba4f0f1
```

The category and sub-category page are meaningless to end-users. In order to find out how many of these pages are indexed on Google you may use the advanced operator “site:”

```
http://www.google.com/search?q=site%3Aindeed.com%2Fdirecto
ry
```

Indeed’s scheme is linking to around 85,000 pages. Most of those pages are just listings of hyperlinks with terrible, if any, usability. In addition to that, it does not work in terms of SEO because Google counts about

1,980 indexed pages or 2.3% of all the ones linked to from the HTML sitemap.

XML sitemaps

XML sitemaps are a good indicator of the inventory of the content that the webmaster intends to promote on Google.

Upload specific XML sitemaps by categories or types of pages of large websites [xxi] and for images[xxii]. Images may drive in sizeable organic traffic from the Google Images index if properly tagged.

```
<?xml version="1.0" encoding="UTF-8"?>
 <urlset
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:image="http://www.google.com/schemas/sitemap-
image/1.1">
 <url>
   <loc>http://example.com/sample.html</loc>
   <image:image>
     <image:loc>http://example.com/image.jpg</image:loc>
   </image:image>
   <image:image>
     <image:loc>http://example.com/photo.jpg</image:loc>
   </image:image>
 </url>
</urlset>
```

If you have a mobile version of your site that you want to be specifically indexed on Google, supply a mobile sitemap

```
<?xml version="1.0" encoding="UTF-8" ?>
 <urlset
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:mobile="http://www.google.com/schemas/sitemap-
mobile/1.0">
    <url>

<loc>http://mobile.example.com/article100.html</loc>
        <mobile:mobile/>
    </url>
</urlset>
```

Update your sitemaps as often as you add new pages of large websites to help Google find them.

If-Modified-Since HTTP header

Make sure your web server supports the If-Modified-Since HTTP header. This feature allows your web server to tell Google whether your content has changed since the last crawl.

Set the crawling rate of Googlebot

If your server/s is/are too loaded or if you are releasing lots of new documents and pages at launch, you want to have a smooth take up on Google index.

You can set the speed of the crawler of Google from a minimum of 0.002 to a maximum of 2 requests per second and from a pace of 500 to 0.5 seconds between requests – that is a factor of 1,000! In other words, you have three orders of magnitude of gearing of the crawling speed and pace of the Googlebot at your server!

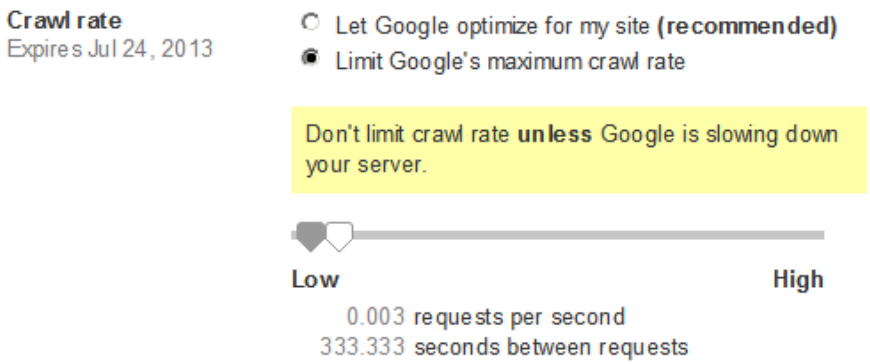


Figure 11 Screenshot: menu Settings of Crawl rate of the section of Configuration on the interface of Google Webmasters Tools

SEO for established websites

The same way that we would only trust a licenced doctor to diagnose a health issue, we strongly advise to get professional help when auditing the SEO potential and status of a given website. Interpreting and contextualizing SEO information takes a lot of experience and specialization.

SEO tries to turn the available resources of a website (content and PageRank) into output (traffic, conversions and engagement). The output should be in direct proportion to the input over time.

IN	⇨	OUT
- Content: quantity, quality and freshness - PageRank [xxiii]	⇨	- Traffic levels - Conversions and Engagement

Google measures more than 200 signals and updates their algorithms several times per week on its ranking software. Even if we knew which ones they are, it would not possible to measure all of them, reverse-engineer Google's algorithms and last but not least, re-build their training data in order to run models of their rankings.

Auditing the degree and potential of the SEO of websites can fortunately be economical for the trained eye. You want to **carry out highly accurate reports with the minimum precision**.

You can get a faithful snapshot of the SEO of a website with few measurements: backlinks, target keywords, content inventory and site architecture.

Variable	Measurement
Backlinks	Quantity, quality, growth rate of backlinks, Google PageRank [xxiv]
Targeted keywords	Monthly searches and competition
Content inventory	Quality and quantity of content, frequency, duplication, keyword stuffing, available pages, pages indexed by Google
Site architecture	Organization of the website, URL structure, internal links

Off-page SEO

Backlinks

Backlinks are part of off-page SEO. We cover them here for their considerable weighting in the ranking software of search engines. There are quick ways to assess the importance of backlinks for any given website. It is highly recommended that you work first hand with a professional SEO if or when considering the improvement or increase in number of your backlinks.

Quantity of backlinks

Popular websites tend to get many backlinks on an ongoing basis.

E.g. Wired.com is a technology news website with a reputation for publishing original and high quality content. Ahrefs.com is a tool for checking backlinks and the social signals of individual websites.

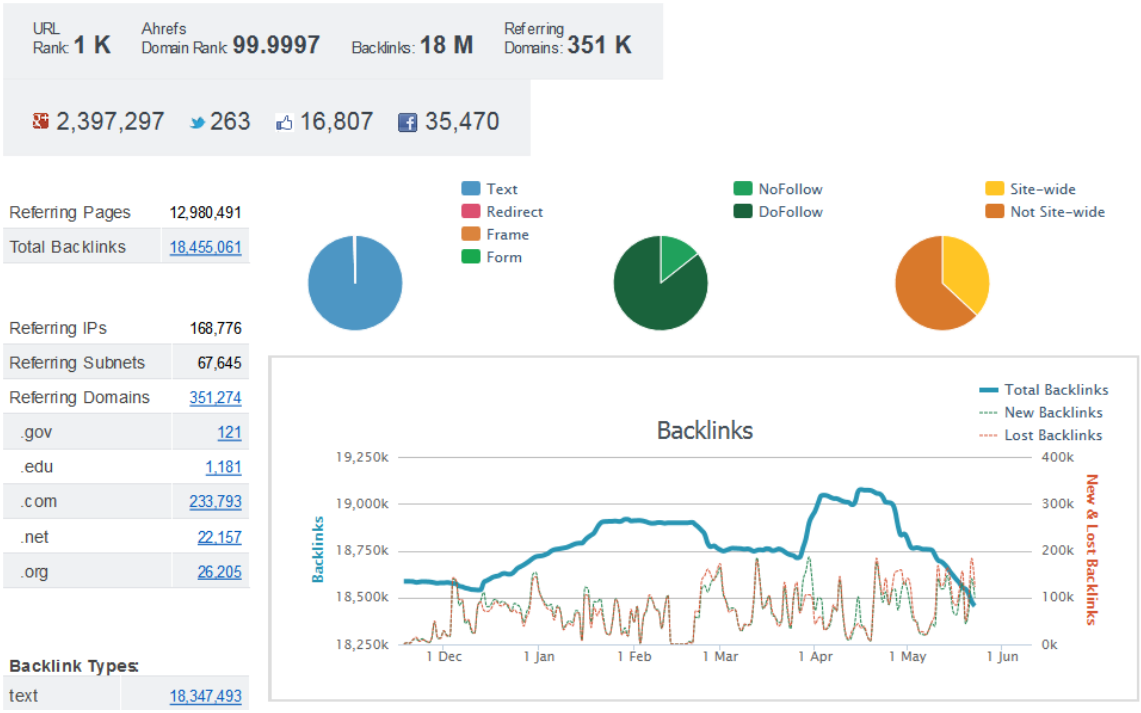


Figure 12 screenshot of a report by ahrefs.com for the domain wired.com with a graph plotting the evolution of the count of backlinks over time.

In this example, the count of backlinks of 18 Million is a remarkably high number from a high number of referring domains: 351,000

You can also compare side by side competing websites with OpenSiteExplorer.org

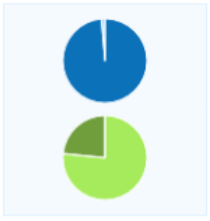
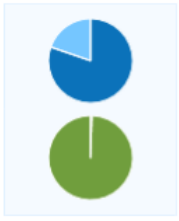
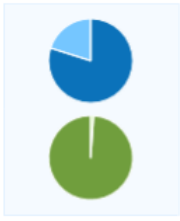
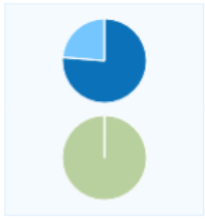
	www.sublimetext.com	notepad-plus-plus.org	www.jedit.org	bluefish.openoffice.nl
Page Authority:	84	91	80	82
Page MozRank:	6.62	6.8	5.98	6.27
Page MozTrust:	6.75	6.36	6.31	6.26
Internal Followed Links:	73,036	631	177	21
External Followed Links:	19,734	98,164	11,849	26,847
Total Internal Links:	73,036	631	178	21
Total External Links:	22,323	130,002	15,848	35,268
Total Links:	95,359	130,633	16,026	35,289
Followed Linking Root Domains:	3,052	12,871	2,793	3,115
Total Linking Root Domains:	3,425	15,142	3,208	3,696
Linking C Blocks:	2,078	7,794	2,159	2,376
Followed Links vs Nofollowed Links:				
Internal Links vs External Links:				

Figure 13 screenshot of Opensiteexplorer.org comparing the linking profiles of the websites of popular text editors: sublimetext.org, notepad-plus-plus.org, jedit.org and bluefish.openoffice.nl. Notepad++ is the best websites for most metrics of quantity and quality of backlinks

The competition for backlinks depends on categories, locales and languages. English-driven websites may potentially enjoy more backlinks from many more websites than other languages on the WWW although competition is theoretically more intense for the same reason.

These are the estimates of the percentages of Web sites using various content languages as of 2014[xxv]:

Rank	Language	Percentage
1	English	56.2%
2	Russian	5.9%
3	German	5.9%
4	Japanese	4.9%
5	Spanish	4.5%
6	French	3.9%
7	Chinese	3.2%

Some online backlinks checker tools do a great job at helping to understand just how many links any website gets. Services like MajesticSEO, Ahrefs and SEOMoz crawl the WWW at a rate of billions of URLs a day. The size of their databases is of hundreds of billions of Unique URLs, which makes them effective for short term information and alerts.

Quality of backlinks

Relevant links on sites with authority and good reputation are beneficial for SEO. The anchor text and the context of links are also signals used by Google to determine the relevancy of backlinks.

Paid links do not comply with the guidelines of Google web search. Finding paid links to third party sites takes often some research with a high degree of manual inspection and knowledge of a vertical and a local market. Knowing that a webmaster is buying sponsored links and that those links do not have the nofollow attribute is a sign of poor off-page SEO practice.

If you are in charge of the SEO of an established website, watch the nature of the backlinks you are obtaining. Stay away from low quality, spam sites linking to you. Paid links or advertorials are legitimate as long as the publishers flag them as such so that Google understands that they are sponsored links.

E.g. Interflora operates an e-commerce website selling the delivery of flowers. Its UK operations incurred in the purchasing of mass

advertorials. Google penalised Interflora.co.uk manually in early 2013. The results on Google of Interflora dropped from 1st and 2nd positions to the 49th place overnight for many converting keywords like “St. Valentine’s Day flowers” in the UK.
Worse come to worse, if you suspect that your site's ranking is being harmed by low-quality links, you can ask Google not to take them into account when assessing your site.

- Decommission all the unnatural links pointing to your site
- Report them via the disavow links tool on Google Webmasters Tools

Growth rate of backlinks

Natural links build up over time. Google is sensitive to sudden bursts of new links. It learned to interpret the sudden peaks as spam behaviour. Inversely, you want to keep a natural, steady rate of new links without dramatic troughs or peaks.

Content inventory

Internal duplication

Avoid duplication of your content among your own pages. A first step to diagnose duplication consists in estimating the percentage of matched words among pages of the same website. Siteliner is an online tool that counts the co-occurrence of the same words on pages of the same website.

E.g. the tool highlights the words on a page that are matched with other pages, e.g. a section of The Guardian
http://www.siteliner.com/www.guardian.co.uk/technology?siteliner=duplicate
39 matches found covering 974 of 1,658 words – 58% of the page

Plagiarism

Use an online plagiarism detection service like Copyscape to find out copies of any given webpage.

E.g. the text of the patent of Google of the PageRank, when checked on the online tool, lists the other pages on the Internet with duplicate snippets of text.
http://copyscape.com/?q=google.com/patents/US6285999

If your content is systematically copied and published somewhere else without your authorisation or citing you as the source, and if this scraped content outranks you on Google, you may consider reporting it on their Scraper Report[xxvi].

Count of indexed pages

The search operator “site:” can be used to get a rough estimate of the number of pages indexed by Google of a given website

E.g. comparison of websites competing for the same audience, for instance, B2C telecom operators with data provided by the count of indexed pages by Google for queries with the “site” operator:

http://www.google.es/search?q=site%3Awww.movistar.es

Subdomain	# pages indexed by Google
www.vodafone.es	107,000
www.movistar.es	24,500
www.orange.es	6,680
www.yoigo.com	2,790
www.simyo.es	369

The differences of several orders of magnitude in the count indexed pages on Google by different operators may point to vast differences in their SEO or in their content marketing strategy.

HTTP Status Codes

SEOs hate webpages with 404 HTTP standard response code from the server. The number 404 denotes a “Not Found error message”, that is, the server has nothing to return for its URL.

Reduce the generation of 404 errors to an absolute minimum. When 404 are inevitable, provide a useful 404 page that let users and bots find the next best content that you can offer.

Even in an ideal world you should have a “404 page” ready to help the user find his/her way around your site. It is important that your 404 page does return a 404 HTTP response code, instead of a 200 one. Web developers tend to oversee this since the server does have something to respond with.

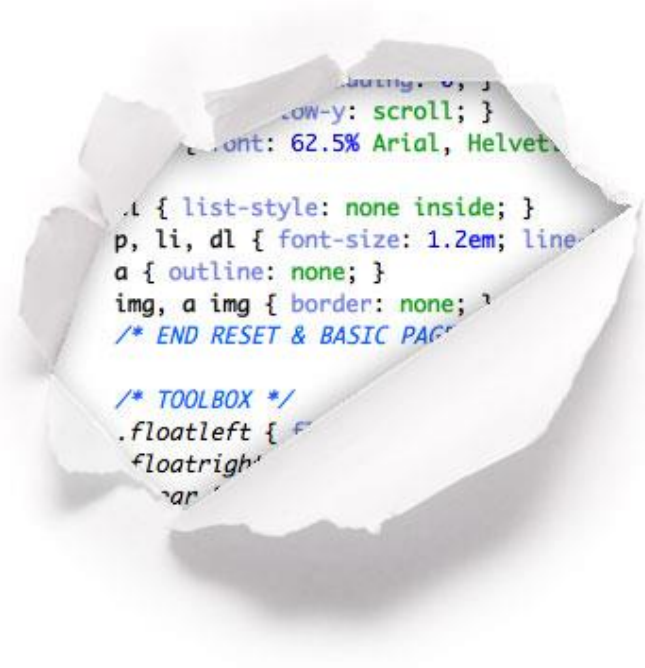


Figure 14 Source: http://css-tricks.com/

Ideally you should the 410 HTTP status code in some cases for obsolete pages. The code indicates that search engines should remove the document from their indices because is no longer available and will not be available again.

Unfortunately the URL whose return code is 410 will still show up on the reports of Site Errors on Google Webmasters Tools.

Server-side redirects

The HTTP response status code 301 status code means that a page has permanently moved to a new location. The 302 ones means temporary redirection. SEOs usually prefer to use the 301 code to the 302 one because the first passes some of the link value, but the second does not.

SEOs usually enter rules which webpages redirect to others on the server's .htaccess file for websites hosted on Apache servers. Mind the regular expressions that your write on the file and test extensively before going to production.

Migration from older versions or consolidated properties

When a medium-sized website launches a new version, almost inevitable some or many URLs will be obsolete and return error codes. This is also the case when domains or subdomains move in under the main www. Subdomain.

Tackle the issue by fixing broken links on your own site, updating your sitemaps and typically, adding a 301 redirect to get users to the "real" page.

An interesting feature of Google Webmasters tools is that the errors are ranked by priority. Google determine the priority with factors such as the presence of the URL in a sitemap, number of backlinks and internal links and, interestingly, whether the URL has gotten any traffic recently from search[xxvii].

Manage the rotation of content

Decide what to do with pages or sections or pages when the content becomes obsolete, redundant or when it is no longer available. Write a content management protocol to deal with obsolete content. If your server supports .htaccess (Apache does), customise the error responses for all the foreseeable situations of missing webpages, obsolete or deprecated ones.

Even when you try to cater for all contingencies, you still may need to remove URLs on Google. While cumbersome and impractical for very large sites, in some cases, submitting individual URLs from Google's index may save your day.

Learn the differences between 301 and 302 redirects and use preferably 301 codes to redirect webpages on the same domain or different domains /obsolete-page.html to /new-page.html

Google Webmasters Tools lets you remove an entire site off Google or redirect it to a new domain.

Site Architecture

This section will be illustrated step by step with an example. We analyse the website of an event, www.bigdataspain.org

First step – Crawl a website

Download Xenu's Link Sleuth (for Windows). It is highly recommended that you download the latest 1.3.9 or higher versions.
Run a crawl (menu File > "Check URL") for the URL www.bigdataspain.org
Once the crawl is finish, export it (Menu File > Export Page Map to a TAB separated file)
At the window "Export Pagemap", save the file with the name crawl.txt

Second step - Filter the pages with internal links only

Inspect the fields of the exported file with bash commands or Perl on a bash terminal

```
$ head crawl.txt
```

Remove the redundant text and most of the internal links to accessory files and external links

```
$ cut -f1,2 crawl.txt | sed -e
's/http\:\/\/www\.{domain}\.{tld}//g' -e 's/\t/,/g' |grep
-v "\.jpg\|http\:|\.css\|\.js" > filtered.csv
```

```
$ head -5 filtered.csv
```

This results in this output with the comma-separated fields where a forward slash only represents the homepage. E.g. the page program.php links up to the home.

```
OriginPage,LinkToPage
/2012/conference/business-big-data/jon-bruner,/
/2012/program.php,/
/2012/speakers.php,/
```

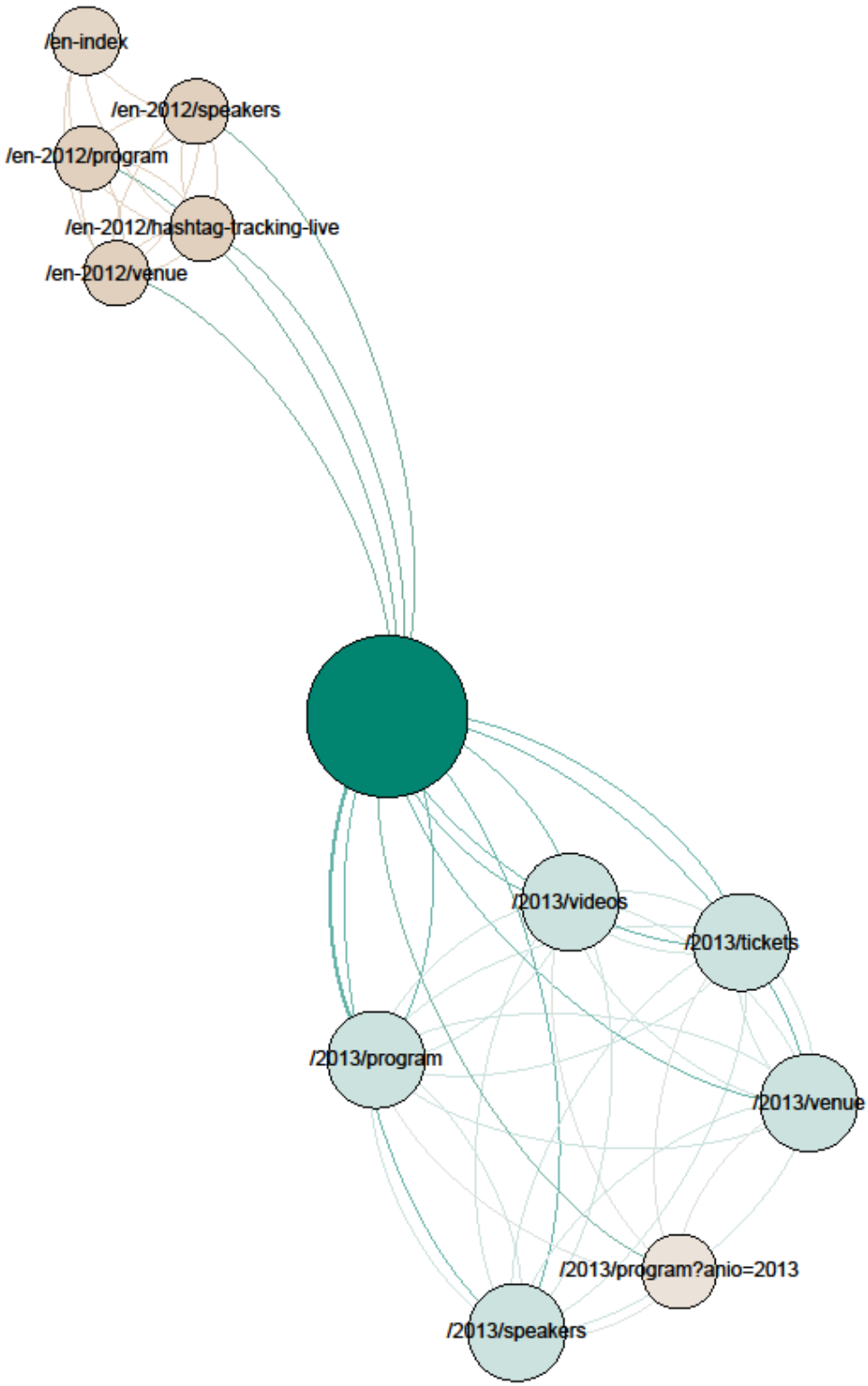
```
/en-2012/speakers.php,/
```

Third and last step - Visualize the network and analyze it

Gephi[xxviii] is a network analysis and visualization software application. Download it from Gephi.org and install it.

Menu “New Project” > File “Open” filtered.csv > Window “Import report”. Options “Graph Type” select the “Directed” value from the menu.
# of Nodes 55
# of Edges 416
Choose a layout ForceAtlas, for instance, from the Window > Layout menu and click on “Run”
Iteratively adjust the settings of the algorithm to plot the heavily linked nodes in a visually effective way.

Figure 15 representing the graph of nodes and edges of www.bigdataspain.org. The size and colour of some nodes have been adjusting for illustration purposes only.



The node *program.php* is more important than the *venue.php* one. This relative importance on the graph is well aligned with the goals of the website. The graph allows detecting at first sight many issues that might otherwise escape the attention of the webmaster.

Note statistics such as average path length (2.559) in the example if you are evolving the website over time so you can compare them.

Examine the graph and find topological information such as:

- Colour of nodes and edges represent different languages on the same website. Red and orange for the English version and grey for the Spanish version
- The number of inbound links that they get from the rest of the pages of the site should be proportional to the value of the node for the site

Some interesting metrics are derived from the mathematics of graphs.

Menu Window Statistics
Graph Density Directed, (simulated) 0.143
PageRank Directed distribution[xxix]: how often a user following links will non-randomly reach the node “page”.

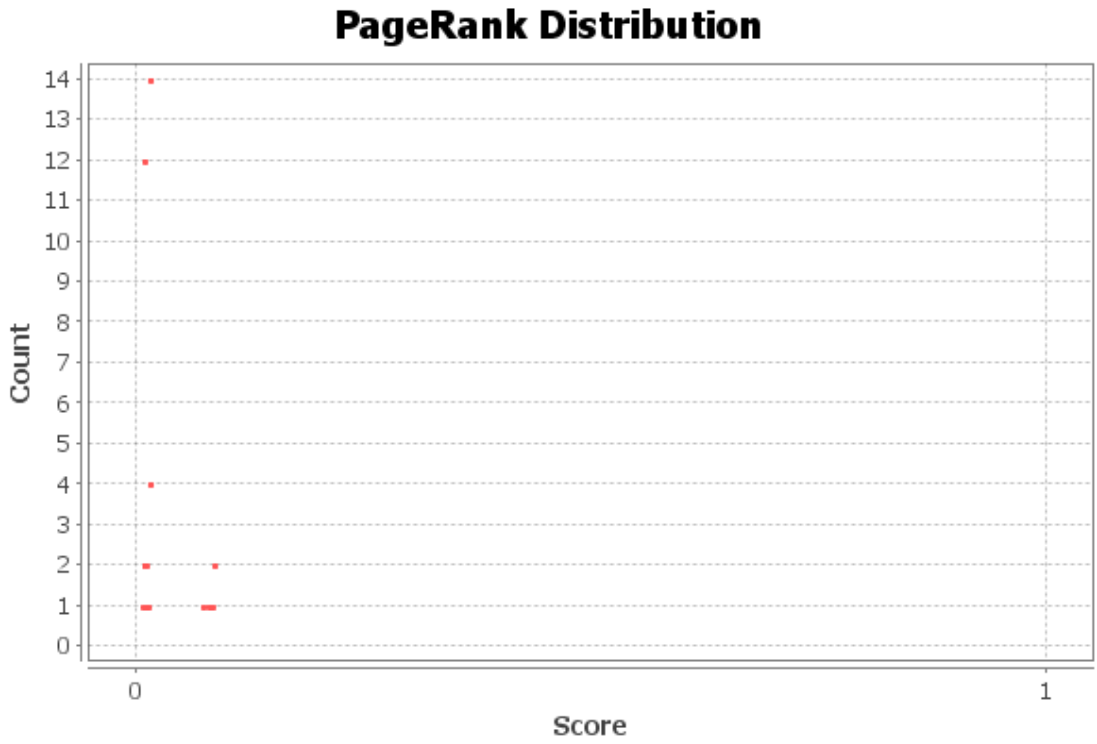


Figure 16 PageRank distribution for a small graph with the parameters Epsilon = 0.001 and Probability = 0.85 Algorithm Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale

Hypertextual Web Search Engine, in Proceedings of the seventh International Conference on the World Wide Web (WWW1998):107-117

The internal PageRank of a given graph is not to be confused with Google Toolbar's PageRank. The PageRank of a local network is a metric of internal authority of each node in a network. In other words, you may assess with webpages enjoy the highest PageRank and thus deduce how important is that webpage for the webmaster or the SEO of that website.

Repeat this exercise of analysis of site architectures with other websites and compare these values over time.

Watch the health of your site

If you are proficient in Perl, a quick script and some bash magic can do wonders to analyse your access logs every now and then. For deeper analysis you may consider Apache Log Viewer or tools to look for issues with the Googlebot. Fluentd is an open source alternative to Splunk for websites with heavy traffic loads consists in combining ElasticSearch, and Kibana.

Crawling by Google

Use Google Webmasters Tools before, during your coding project and after launch. Google Webmasters Tools helps to find out about Googlebot's crawling stats and identify crawl errors. The section "Index Status" is a great indication of just how deep and wide the Googlebots crawls your site over time.

E.g. If Google is indexing everything that it crawls since you launched your site you did a great job.

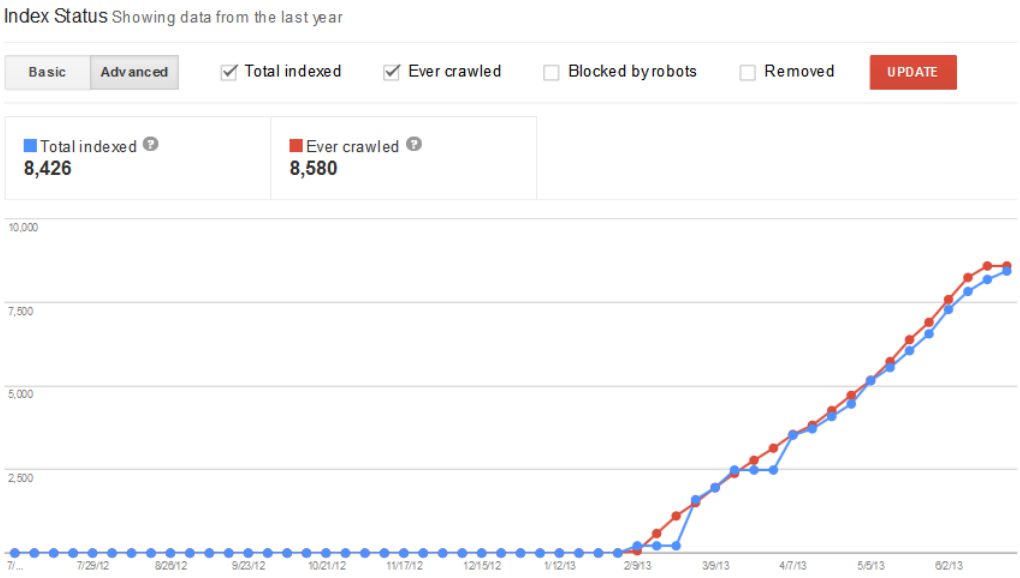


Figure 17 of Index Status of the section Health of Google Webmasters Tools. The take up by Google of virtually all the content that it crawls on a new website is good news yet quite unusual

Server logs

Google used to include almost all the original search queries of its users in the referrer. It started to omit the query for those sessions of logged-in users, i.e. those using Google under the secured protocol https instead of http. Depending on a number of factors, the percentage of visits on https is 30-40% of the total. Nevertheless extracting search queries from the logs remains a basic tool in the SEO trade.

- Setup your servers to archive logs automatically.
- Check regularly the IP range of Googlebot instances and the most important search engines
- Report the frequency of crawling of the top pages, identify all the redirects and all the error return codes
- Extract queries from the referrer. Google does not pass on the queries of sessions of logged-in users using the securitised protocol https. Varying over time and depending on the websites, the percentage of missing queries on referrers from Google is around 30% of the total

- You can count the frequency and rankings of the captured keywords on Google on a per session basis. Use the statistics package of the R project to understand the number of visits by query [xxx] and by ranking of your webpages on Google SERPs.

Health check of indexed URLs

Obsolete or deprecated pages result in 404 error response codes. Google interprets these as a signal of poor quality of a given website. It is important to keep these errors under control.

Put all the URLs in a txt file on your bash shell, e.g. *urls.txt*

```
http://www.website.com/j2ee-java-developer
http://www.website.com/java-developer-j2e3-spring-eclipse
http://www.website.com/software-consultant-java
http://www.website.com/technical-lead-java
http://www.website.com/engineer-full-stack-javascript-
python-java
```

Then check them with a simple shell script as long as they are only a few hundred so you do not overload the server with your loop.

```
$ while read i; do echo $i && curl -s -w "%{http_code}\\n"
              $i -o /dev/null; done < urls.txt
```

This is the output you should expect:

```
http://www.website.com/j2ee-java-developer
404
http://www.website.com/head-developer-j2e3-spring-eclipse
410
http://www.website.com/software-consultant-java
200
http://www.website.com/technical-lead-java
301
http://www.website.com/engineer-full-stack-javascript-
python-java
404
```

Obviously you are interested in the URLs that present server errors, i.e. those with 404 for instance.

Log file parsing

LogFormat "%h %l %u %t \"%r\" %>s %b" common

%r "GET /apache_pb.gif HTTP/1.0" (\"%r\"). The request line from the client is given in double quotes

%>s status code returned by the host

Extract individual queries and the positions of your pages on Google SERP with a few lines of code on R [xxxi]

```
log <-
getURL("sftp://user:password@host:/path/to/apache/accesslo
g.log")
```

Block bots other than search engines

Block –or try to- bad bots and rippers that may weight down on your server with .htaccess or via an algorithmic approach.

Tools and references

The tools are as good as the questions you ask their data. We list a few of them here for you to start getting a feeling of what to ask an SEO about.

Disclaimer: the author is not affiliated with any of the vendors whose tools or names are provided below.

Measurement	Tools
Backlinks	ahrefs, OpenSiteExplorer, MajesticSEO
PageRank	Google Toolbar (only available for Internet Explorer)
Frequency & competition of keywords	Adwords' Keyword Planner
Rankings	Google Webmasters Tools, Advanced Web Rankings
Content inventory	Xenu (desktop app Windows), Screaming Frog (SaaS)
Duplicate content	Copyscape, Siteliner
Pages indexed by Google	Operator "site:" in web search[xxxii]
Site architecture	Gephi
Server logs	Apache Log Viewer, Splunk, R-project
Crawling	Google Webmasters Tools, w3m, LWP, Live HTTP Headers add-on for Firefox
Engagement	R-project web metrics SaaS like Google Analytics

What now?

Experiment with SEO as much as your business and your resources can afford. Ask expert SEOs and read regularly blogs about search and IA.

Epilogue

Was this eBook useful to you? If you like it, pass it on to your colleagues and friends!

We wrote this eBook to help people but not everyone reads English, so you can contribute with translating this handbook to other languages as long as you let us know, provide attribution and generally abide by its licensing terms.

For questions, reports of errors or suggestions or just to say hello by dropping a message on Paradigma's website, mine or via:

- Google+ Rubén Martínez
- Twitter @RubenMartinezS
- LinkedIn rubenm
- Quora Ruben Martinez

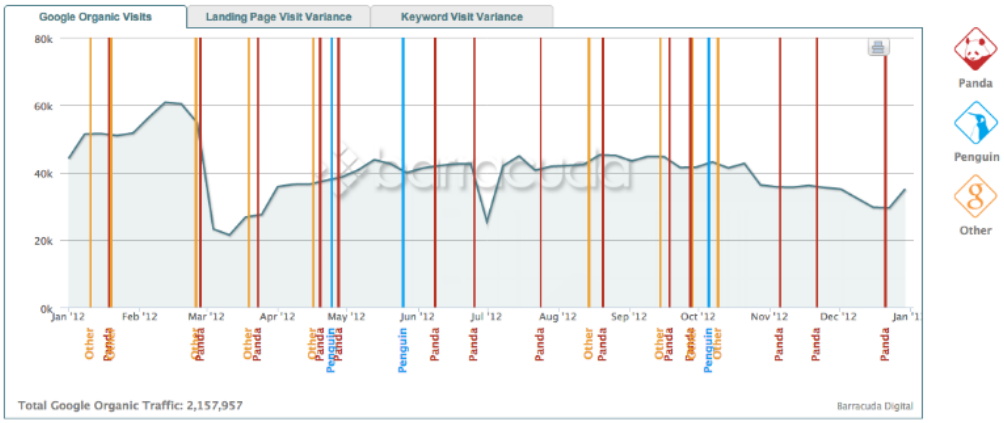
Appendix – Google Updates

Google updates are to marketing what Black Fridays are to the stock market. Some SEO veterans still remember the Google "dances" which basically made or break many businesses online.

More recently Google Caffeine changed the way Google deals with its index. Google could, by the first time, update its index incrementally when it was released in June 2010. "*When a new page is crawled, Google can update its index with the necessarily changes rather than rebuilding the whole thing"* - Eisar Lipkovitz, a senior director of engineering at Google in 2010[xxxiii].

Following the big leap in technological Google Panda, Venice, Penguin and Hummingbird are other important updates that changed

You may find a historical review of Google updates at Moz's updates calendar. You can also check if your Google Analytics' reported visits match with known Google updates. You can give Barracuda, an online marketing agency, access to your Google Analytics account so they can plot the occurrence of updates.



SERP volatility

Search engines like Google update their datasets and adjust their algorithms more than 400 times every year. SEO professionals jump on

their seats when those updates change significantly the rankings of the best converting queries.

Some tools helps understand the changes in positions on SERPs. One of them is Serpmetrics.com. Their charts plot a volatility index. It is composed by comparing the results on Google regularly. A scoring system applies weighting to the top results and calculates the turbulence over time.

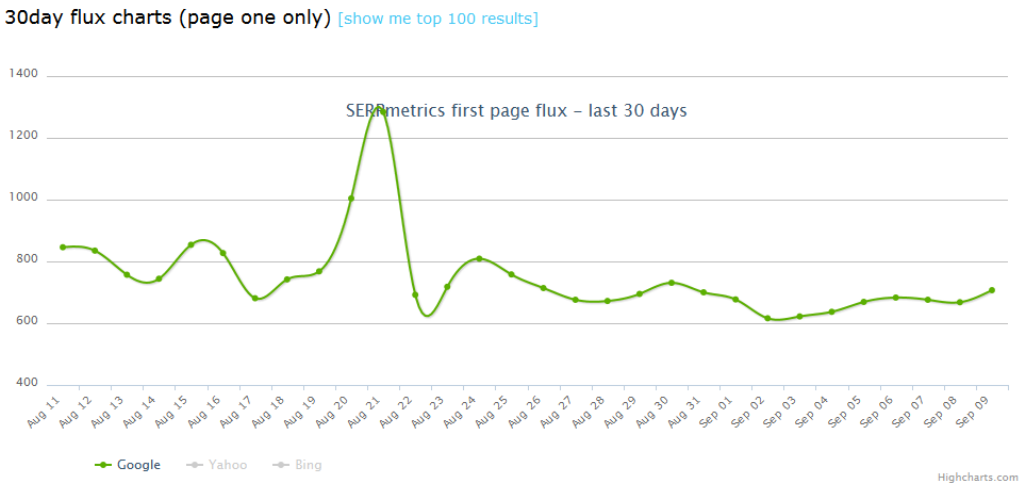


Figure 18 Graph of the flux of changes on Google's SERPs over time. Some turbulence is clearly identifiable before and after August 20th

Appendix – Target keywords

Target keywords are those that are used in titles, URLs, anchor texts of links and other elements of the HTML. Think of the mark up of targeted keywords as important assets in your SEO. Many webmasters ignore which targeted keywords they are really promoting on their websites.

The stage of keyword research is not strictly speaking technical SEO. We mention it on this document because it is crucial for the success of your mark-up.

The keywords that define your business and your target audience are often overlooked at the beginning of web development projects.

E.g. look at the top 5 sites ranking for the query “Terms and Conditions”, searched for on average 135,000 times in the United Kingdom monthly.

```
http://www.google.co.uk/#hl=en&output=search&q=terms+and+conditions
```

At the time of writing this document, Tesco.com, a grocery retailer and O2.co.uk, a mobile network operator, rank 3rd and 4th respectively. Do they really want to rank for those search terms? The query “*Terms and Conditions*” has not transactional or commercial search intent among the customers of Tesco or O2. The rankings of their websites are not helping to sell more groceries or phone contracts.

Find if your audience is increasingly interested in your targeted keywords or not. A way of finding our quickly is by checking the normalized frequency of monthly searches by Google Trends.

E.g. Arduino and Raspberry Pi are open-source hardware devices. Arduino became popular before Raspberry Pi was launched.

```
http://www.google.com/trends/explore#q=arduino%2C%20raspberry%20pi&cmpt=q
```

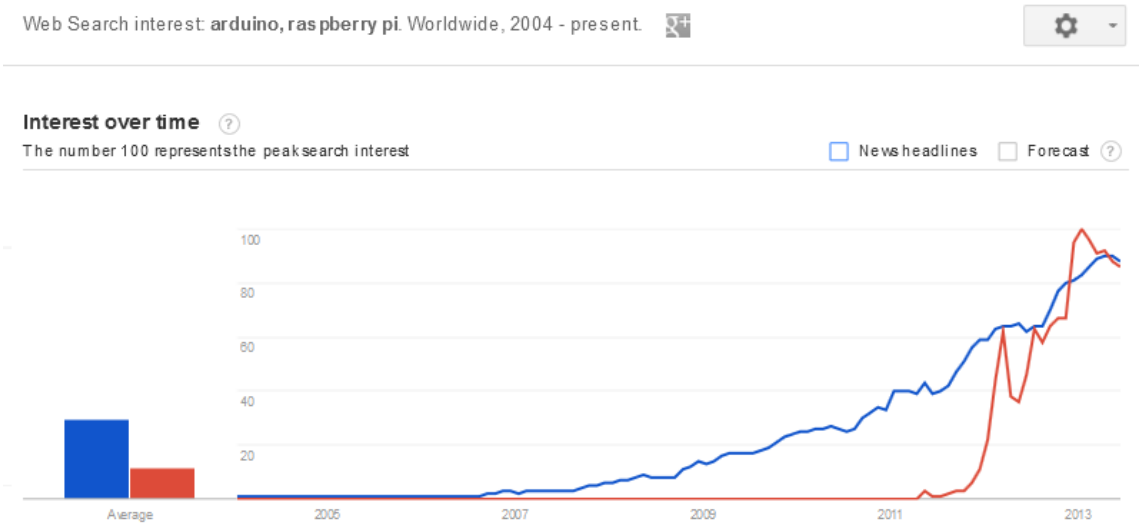


Figure 19 The interest in the Raspberry Pi project (red line on the graph) has recently overtaken the popularity of the Arduino project (red line) only to drop both in the last few weeks.

Find out the competition for your targeted keywords and whether they are head or long tail keywords.

E.g. in English and locale United Kingdom

flights Majorca
flights Menorca
flights Ibiza
flights Tenerife
flights Las Palmas

Google's Keyword Planner helps you find the popularity of your keywords in number of monthly searches and their relative "difficulty" to rank for them.

Keyword	Competition	Global Monthly Searches	Local Monthly Searches
flights Majorca	High	49,500	1,000
flights Menorca	High	22,200	320
flights Ibiza	Medium	74,000	2,400
flights Tenerife	Medium	135,000	2,400
flights Las Palmas	Medium	5,400	390

Appendix – Domain names

The projects of new websites often require the registration of a new domain name. While the topic is related to off-page SEO, we mention it here because we usually are asked about it by developers.

Top level domains (TLDs) typically have the *.com* name for commercial and *.org* for organizations. You can register a domain with your brand name or including keywords.

Google has a reputation of being biased towards big and established brands. Investing in new, niche or second tier branded domains is a strategic decision that makes all the sense in terms of marketing. In terms of SEO however it may take time to pay off even if Google picks up the correlation between your brand and a category or niche of keywords.

Domain names with keywords on the other hand tend to rank high on Google's SERPs and definitely on Bing's. Exact Match Domain (EMD) names contain the highest converting keywords of the business.

E.g. Forextrading.com ranks top for the query "*forex trading*". The owner, Saxo Bank, opted to use an EMD for their business than to promote a brand like Oanda.com do for the same segment of the market.

Google claims to reduce the low-quality "exact-match" domains in search results but in many countries and languages their SERPs are still plagued with them.

Internationalization of domains

Country-code top-level domains (ccTLDs) are indicated to target local markets like domain.ca in Canada and domain.es in Spain if you can afford the complexity in their administration.

In some cases, if you do not want or you cannot localize your domains with ccTLDs, you can setup folders by language, e.g. both Zara and H&M organize their local websites with folders
www.zara.com/jp/ and www.hm.com/jp/ both for Japan.

Subdomains and subfolders

Sub-domains let you organize your domains sensibly by markets, target audience and products and services. As a rule of thumb, create sub-domains for unrelated content.

E.g. the revenue of Suzuki Motor Corporation in Europe comes mainly from 2 consumer categories: cars and motorcycles. It markets each category with a different sub-domain: http://auto.suzuki.es/, http://moto.suzuki.es/ respectively.

Sub-domains are usually ideal for horizontal scaling of categories of products or services but they do not benefit from the PageRank of their root domain. On the other hand, Google Webmasters Tools will not take sub-domains as addresses when you want to move your site to a new domain and notify Google.

Appendix - Google Analytics

Many developers have questions about Google Analytics - mostly about how to implement it and where. The tool is a powerful SaaS suite that is popular partly for being free for sites under 10M hits processed per month, i.e. for the vast majority of websites.

Implement the tracker on your website before or at launch of a new website- You need to get Google to verify your ownership or authority over a website with a few steps:

1. On the Webmaster Tools Home page, click the **Manage Site** button next to the site you want, and then click **Verify this site**.
2. In the **Recommended method tab**, follow the steps on your screen
3. Click **Verify**. Removing the code from your page can cause your site to become unverified, and you will need to go through the verification process again.
4. Copy and paste your tracking code on every page you want to track

Alternatively, create a PHP file named "analyticstracking.php" with the code above and include it on each PHP template page. Then, add the following line to each template page immediately after the opening <body> tag or use a common include or template to paste the code above instead of manually adding it to every page:

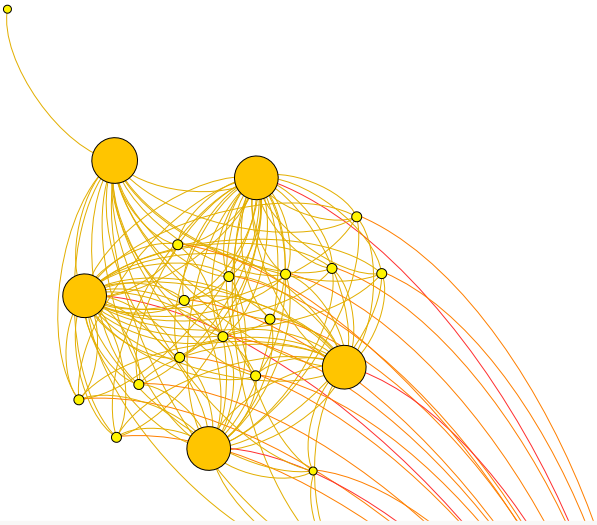
```
<?php include_once("analyticstracking.php") ?>
```

Once you set up Google Analytics tracking, you want to track Traffic Sources – Advanced Segments – Non-paid Search Traffic and Conversion metrics among many other metrics and reports.

Take Google Analytics' data with a pinch of salt for a number of reasons:

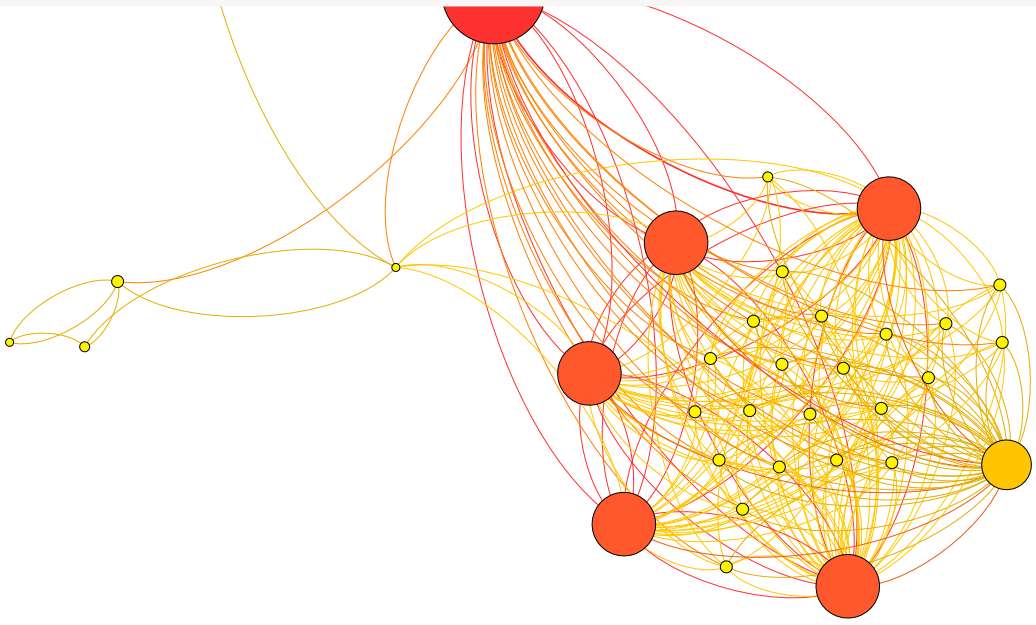
- asynchronous technologies have limitations of their own
- individual visits are not reported – all the data are aggregated
- visits under the secured protocol https are aggregated under the "not provided" category for search queries, etc.

Technical SEO for web developers



A reference guide of technical SEO for software developers

Rubén Martínez



{ paradigma

The best way forward is to combine metrics from your server logs and from Google Analytics or other metrics tools.

Engagement

Your content needs to engage with your audience. Usage signals – like clicks on tagged elements, conversions (goals), time per page, pages/session, social signals, etc. are now part of SEO.

Detect losses in the navigation flow of your website by drilling Google Analytics and correct them appropriately.

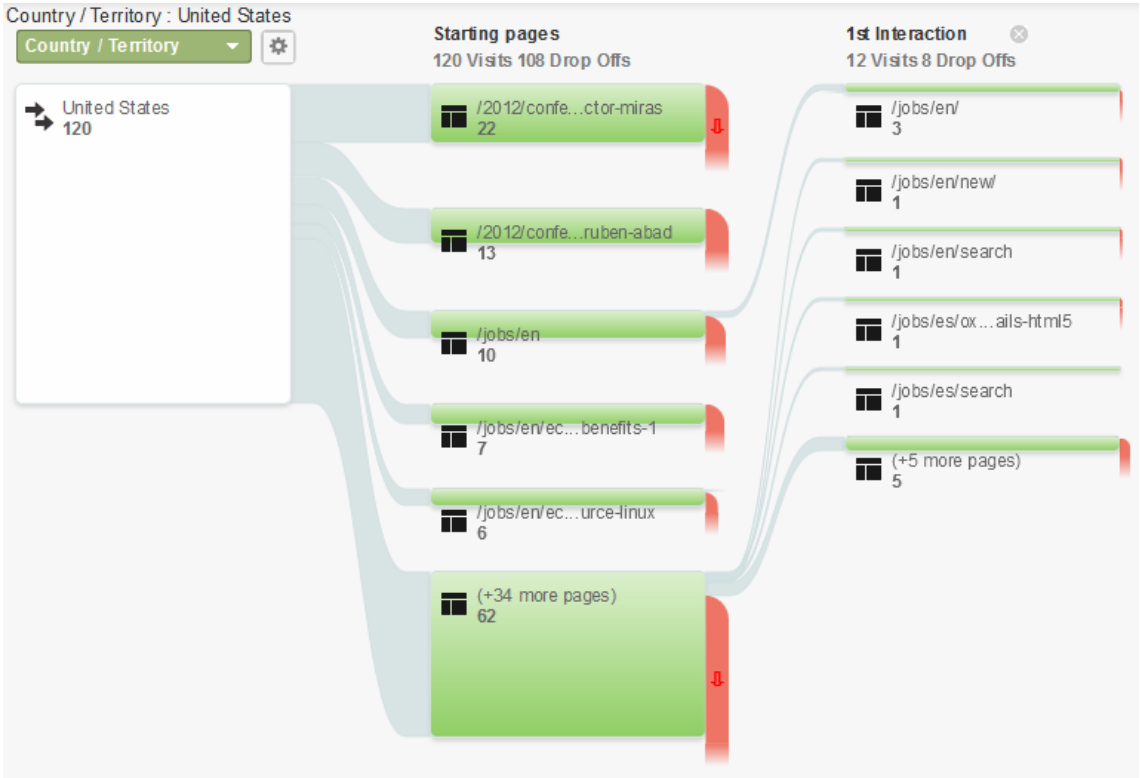


Figure 20 with a snapshot of the downstream interactions of users by geographic origin on the jobs section of www.bigdataspain.org. Losses are represented by red flows pouring from the green boxes with counts of visits.

Split or A/B tests

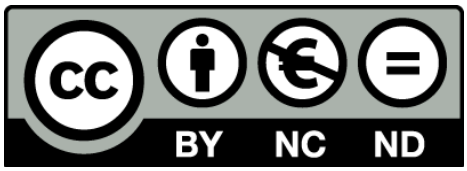
Split tests are formidable tools to optimize conversion ratios. You can use them for SEO also. You should test every component of your technical SEO, from the mark-up of your webpages to the architecture of your website. You may want to verify for instance the effect of the reasonable surfer paradigm by testing different distributions of links on-page.

Just about every aspect of a website can be split-tested. Be prepared to accept insignificant differences and counter-intuitive results. In order to get significant differences you may have to try to test radically different things.

Licensing

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License [xxxiv].

If you want to translate this work to other languages, the resulting work is licensed under the same license with attributions to this copy and version, the author and Paradigma.



References

[i] Search Engine Optimization (SEO)
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35291

[ii] Google for Webmasters
http://developers.google.com/webmasters/googleforwebmasters/?hl=en

[iii] Dynamic IP addresses on cloud or load balancers
http://www.seroundtable.com/seo-amazon-elastic-load-balancing-14123.html

[v] PageRank sculpting
http://www.mattcutts.com/blog/pagerank-sculpting/

[vi] Method for node ranking in a linked database, invented by Page, Lawrence, US Patent 6285999, Filing date: Jan 9, 1998

[vii] Ranking documents based on user behavior and/or feature data, Invented by Jeffrey A. Dean, Corin Anderson and Alexis Battle, Assigned to Google Inc.
US Patent 7,716,225, Filing date: June 17, 2004

[viii] Shell one-liner for check for broken links
http://snipplr.com/view/18344/

[ix] Nofollow attribute
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=96569

[x] Building Smartphone-Optimized Websites
https://developers.google.com/webmasters/smartphone-sites/details

[xi] URI Generic Syntax, Berners-Lee et al, Jan 2005
http://www.ietf.org/rfc/rfc3986.txt

[xii] URL mappings plug-in on Grails
http://grails.org/doc/2.2.x/ref/Plug-ins/URL%20mappings.html

[xiii] Pagination
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=1663744

[xiv] Handling legitimate cross-domain content duplication
http://googlewebmastercentral.blogspot.com.es/2009/12/handling-legitimate-cross-domain.html

[xv] Latent Dirichlet allocation LDA
http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[xvi] Authorship on Google Webmasters Tools
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=1229920

[xvii] Hacked content
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=2721435

[xviii] Cloaking
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=66355

[xx] Making AJAX Applications Crawlable
https://developers.google.com/webmasters/ajax-crawling/

[xxi] Multiple XML Sitemaps: Increased Indexation and Traffic
http://www.seomoz.org/blog/multiple-xml-sitemaps-increased-indexation-and-traffic

[xxii] Image Sitemaps
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=178636

[xxiii] Webpages with a higher PageRank are more likely to appear at the top of Google search results
http://support.google.com/toolbar/bin/answer.py?hl=en&answer=79837

[xxiv] Crawl budget
http://www.stonetemple.com/articles/interview-matt-cutts-012510.shtml

[xxv] Usage of content languages for websites
http://w3techs.com/technologies/overview/content_language/all

[xxvi] Scraper report form
https://docs.google.com/a/google.com/forms/d/1Pw1KVOVRyr4a7ezj_6SHghnX1Y6bp1SOVmy60QjkF0Y/viewform

[xxvii] Crawl errors reported by Google Webmasters Tools
http://googlewebmastercentral.blogspot.com.es/2012/03/crawl-errors-next-generation.html

[xxviii] Gephi
https://gephi.org/

[xxix] PageRank on Gephi http://wiki.gephi.org/index.php/PageRank

[xxx] Una aplicación SEO con R
http://www.datanalytics.com/blog/2013/01/10/una-aplicacion-seo-con-r/

[xxxii] Advanced Operators for Web Search
http://sites.google.com/site/gwebsearcheducation/advanced-operators

[xxxiii] Google Caffeine explained
http://www.theregister.co.uk/2010/09/09/google_caffeine_explained/

[xxxiv] http://creativecommons.org/licenses/by-nc-nd/3.0/