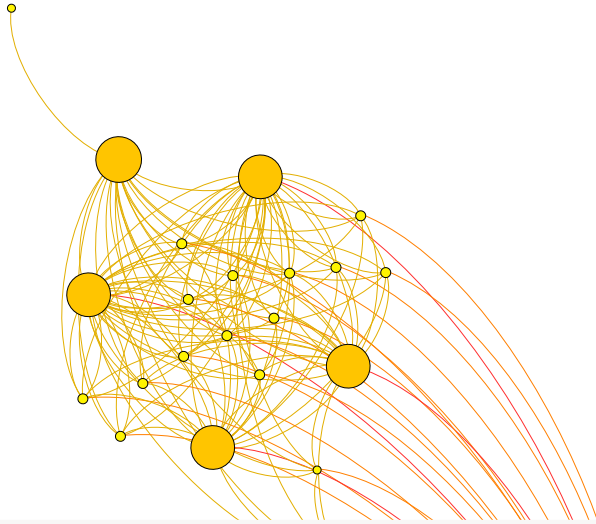
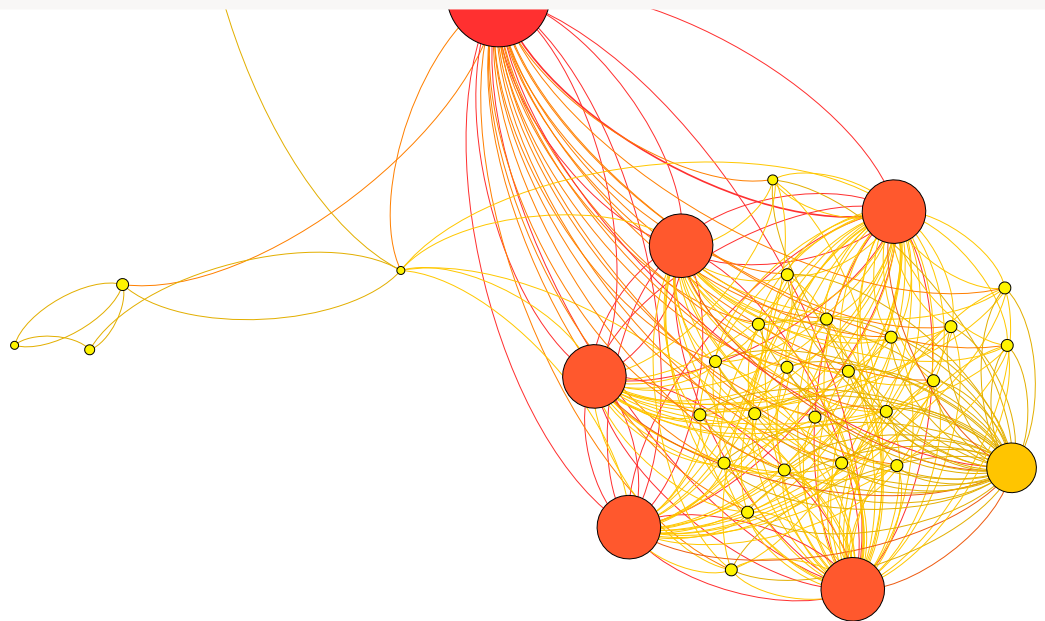


Technical SEO for web developers



A reference guide of technical SEO
for software developers

Rubén Martínez



{ paradigm

SEO for Web Developers

Table of Contents

Introduction	4
Disclaimer.....	4
About the author	4
Special thanks	5
What is SEO?.....	6
Which are the differences between technical and off-page SEO?.....	6
Why is SEO important?	7
Is SEO <i>free</i> ?	8
Google's official stance on SEO.....	8
Jargon buster	9
SEO deals with the bottlenecks in the search flow.....	12
Bottleneck 1: Limitations of keywords	13
Bottleneck 2: The World Wide Web & Search Engines	15
Bottleneck 3: Web servers, websites and code	15
Bottleneck 4: The content itself.....	15
SEO for new websites	17
Why is technical SEO relevant?.....	17
Hosting	17
Host your site on reliable servers with excellent connectivity.....	17
Check your neighbours in shared hosting environments.....	18
Hosting services with dynamic IP addresses	19
Information Architecture	19
Google PageRank	19
Design a lean site architecture.....	19
Link your internal pages sensibly	21
Configuration of mobile rendering	27
Uniform Resource Identifier URL.....	28
URI Syntax	28
Compose a simple URL path	28
URL encoding	29
Friendly URLs.....	29

SEO for Web Developers

Automate the generation of URLs with intuitive rules.....	31
Mark-up your content.....	32
Title Tag.....	32
Meta elements.....	32
Headings.....	34
Main and aside tags.....	34
Rich media (images, videos).....	34
Canonicalization.....	35
Anchor text.....	35
Structured Data.....	36
Authorship.....	37
Robots.txt protocol.....	38
Monitor your site for hacked content.....	39
HTML, JavaScript, AJAX and CSS.....	39
Code for speed.....	39
Debug for crawlers.....	40
Avoid cloaking.....	40
Make AJAX content crawlable.....	40
<noscript> tag for content on JavaScript.....	42
Avoid frames and Flash.....	43
Avoid using CSS to hide text.....	43
Generate sitemaps.....	44
HTML sitemaps.....	44
XML sitemaps.....	46
If-Modified-Since HTTP header.....	47
Set the crawling rate of Googlebot.....	47
SEO for established websites.....	48
Off-page SEO.....	49
Backlinks.....	49
Quantity of backlinks.....	49
Quality of backlinks.....	51
Growth rate of backlinks.....	52
Content inventory.....	53
Internal duplication.....	53
Plagiarism.....	53

SEO for Web Developers

Count of indexed pages	53
HTTP Status Codes	54
Server-side redirects	55
Migration from older versions or consolidated properties	56
Manage the rotation of content	56
Site Architecture	57
First step – Crawl a website	57
Second step - Filter the pages with internal links only	57
Third and last step - Visualize the network and analyze it	58
Watch the health of your site	61
Crawling by Google	61
Server logs.....	62
Health check of indexed URLs.....	63
Log file parsing	64
Block bots other than search engines.....	64
Tools and references.....	65
What now?	66
Epilogue.....	66
Appendix – Google Updates	67
SERP volatility.....	67
Appendix – Target keywords	69
Appendix – Domain names	71
Internationalization of domains	71
Subdomains and subfolders.....	72
Appendix - Google Analytics	73
Engagement	74
Split or A/B tests	75
Licensing.....	76
References	77

SEO for Web Developers

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.

SEO for Web Developers

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.

SEO for Web Developers

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

SEO for Web Developers

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.

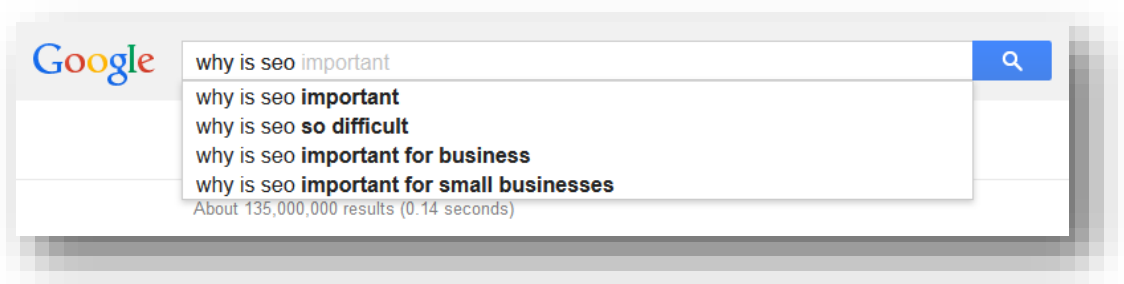


Figure. Screenshot of Google search box for the query “why is SEO”. The search engine suggests auto-completions of related queries

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.

SEO for Web Developers

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

SEO for Web Developers

provide tools, posts on forums by Google employees, videosⁱⁱ, 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.

SEO for Web Developers

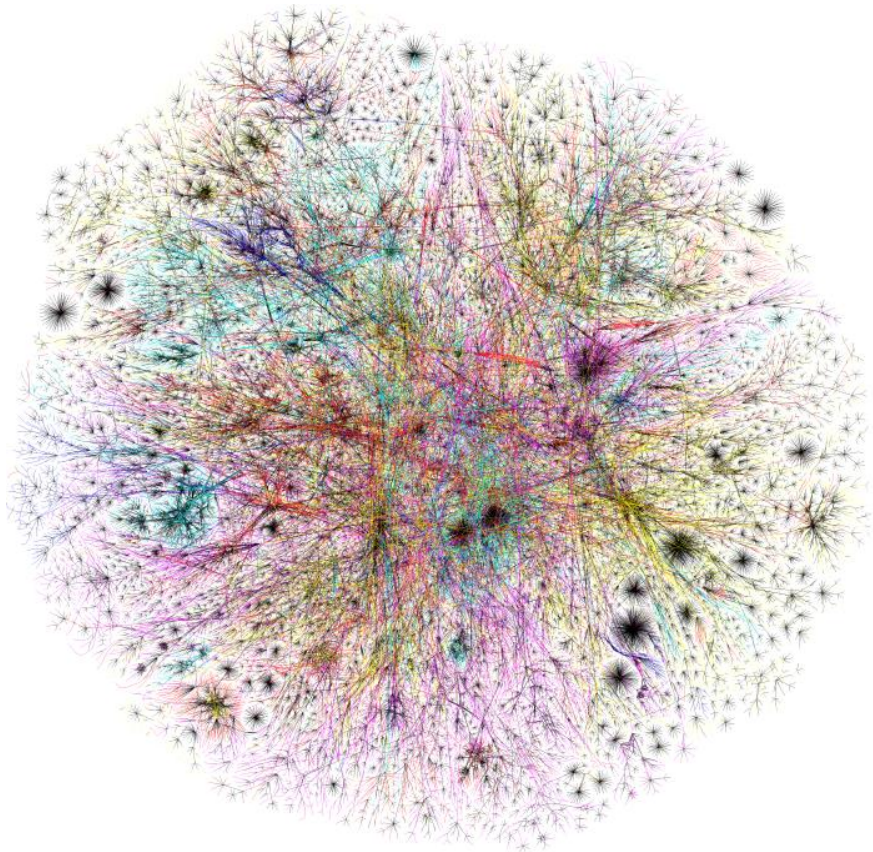


Figure 1 of a graph of 5 million edges and an estimated 50 million hop count by the Opte project modelling the Internet in 2003. Colours modified from the original representing different regions of IP addresses like Asia Pacific, Europe, North and South America, etc.

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.

SEO for Web Developers

- **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 for Web Developers

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.

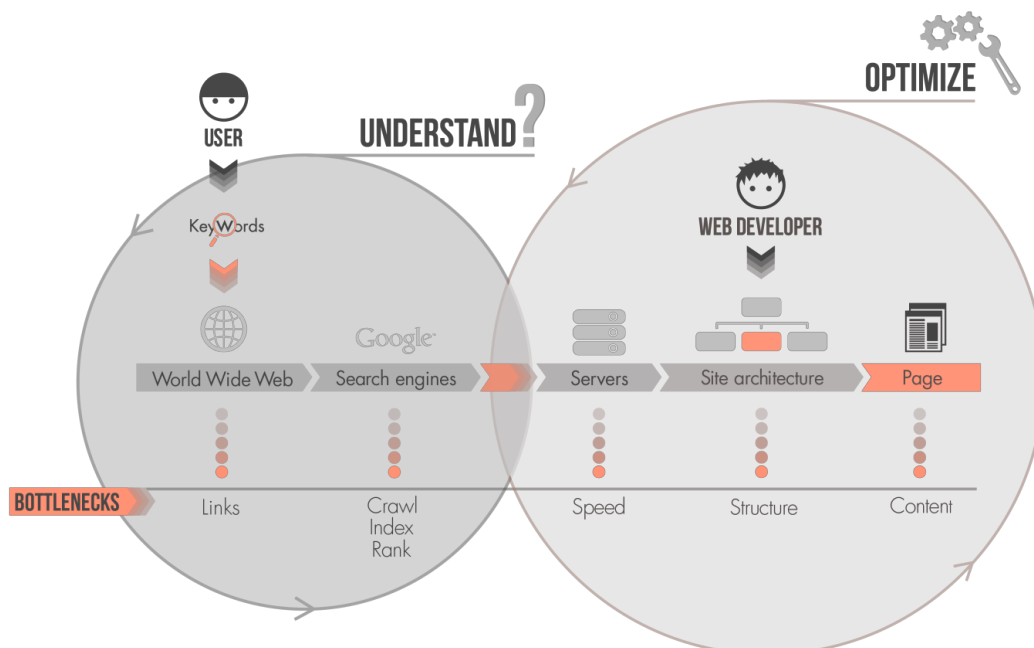


Figure 2 the flow of search is represented on the diagram above from the left (users) to the right (content). There are a few bottlenecks that affect the efficiency of the search.

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.

SEO for Web Developers

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.

SEO for Web Developers

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.

SEO for Web Developers

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:

SEO for Web Developers

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 Web Developers

SEO for new websites

Web developers that pay attention to on-page SEO in the early stages of new projects usually save lots of 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 to 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 *tracert* (*tracert* on Windows) displays the path and transit delays of packets across a given route at a point in time.