



ADVANCED TECHNICAL SEO

A COMPLETE GUIDE

**404 VS. SOFT
404 ERRORS:**

What's the Difference
& How to Fix Both

HTTP OR HTTPS?

Why You Need a Secure Site

How to Use
XML SITEMAPS
to Boost SEO

How to Perform an

**IN-DEPTH
TECHNICAL SEO
AUDIT**

SEJ SearchEngine®
Journal



CONTENT

INTRODUCTION

Technical SEO Is a Necessity, Not an Option

By Andy Betts

CHAPTER 1

SEO-Friendly Hosting: 5 Things to Look for in a Hosting Company

By Amelia Willson

CHAPTER 2

The Ultimate Guide for an SEO-Friendly URL Structure

By Clark Boyd

CHAPTER 3

How to Use XML Sitemaps to Boost SEO

By Jes Scholz

CHAPTER 4

Best Practices for Setting Up Meta Robots Tags & Robots.txt

By Sergey Grybniak

CHAPTER 5

Your Indexed Pages Are Going Down – 5 Possible Reasons Why

By Benj Arriola

CHAPTER 6

An SEO Guide to HTTP Status Codes

By Brian Harnish

CHAPTER 7

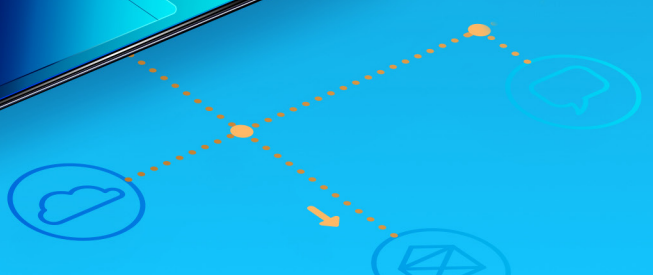
404 vs. Soft 404 Errors: What's the Difference & How to Fix Both

By Benj Arriola

CHAPTER 8

8 Tips to Optimize Crawl Budget for SEO

By Aleh Barysevich





Monitor your website's performance at scale

- ✓ Javascript rendered crawling
- ✓ Customized dashboards
- ✓ Automated workflows
- ✓ Task management

Monitor your sites technical health with DeepCrawl. Maintain sound site architecture, proper internal linking and provide positive user experiences for visitors across both mobile and desktop devices.

[Sign up for a free trial today](#)



@DeepCrawl

+44 (0) 207 947 9613

<https://www.deepcrawl.com>





CONTENT

CHAPTER 9

How to Improve Your Website Navigation: 7 Essential Best Practices

By Benj Arriola

CHAPTER 10

HTTP or HTTPS? Why You Need a Secure Site

By Jenny Halasz

CHAPTER 11

How to Improve Page Speed for More Traffic & Conversions

By Jeremy Knauff

CHAPTER 12

7 Ways a Mobile-First Index Impacts SEO

By Roger Montti

CHAPTER 13

The Complete Guide to Mastering Duplicate Content Issues

By Stoney G deGeyer

CHAPTER 14

A Technical SEO Guide to Redirects

By Vahan Petrosyan

CHAPTER 15

SEO-Friendly Pagination: A Complete Best Practices Guide

By Jes Scholz

CHAPTER 16

What Is Schema Markup & Why It's Important for SEO

By Chuck Price





CONTENT

CHAPTER 17

Faceted Navigation: Best Practices for SEO

By Natalie Hoben

CHAPTER 18

Understanding JavaScript Fundamentals: Your Cheat Sheet

By Rachel Costello

CHAPTER 19

An SEO Guide to URL Parameter Handling

By Jes Scholz

CHAPTER 20

How to Perform an In-Depth Technical SEO Audit

By Anna Crowe



Introduction

Technical SEO Is a Necessity, Not an Option

SEJ
EBOOK

Written By
Andy Betts
Executive & CMO Advisor



The practice of SEO has changed more than any other marketing channel over the last decade.

Through a **succession of algorithmic evolutions**, SEO has also remained the foundation of a successful digital strategy – **51 percent** of online traffic arrives at websites via organic search, after all.

SEO has gone mainstream.



Still, we must take stock of the fact that SEO in 2018 requires **new skills** and approaches to succeed in an increasingly competitive world.

With **more than 5,000** devices integrated with Google Assistant and **voice search** on the rise, the focal points of search have become decentralized.

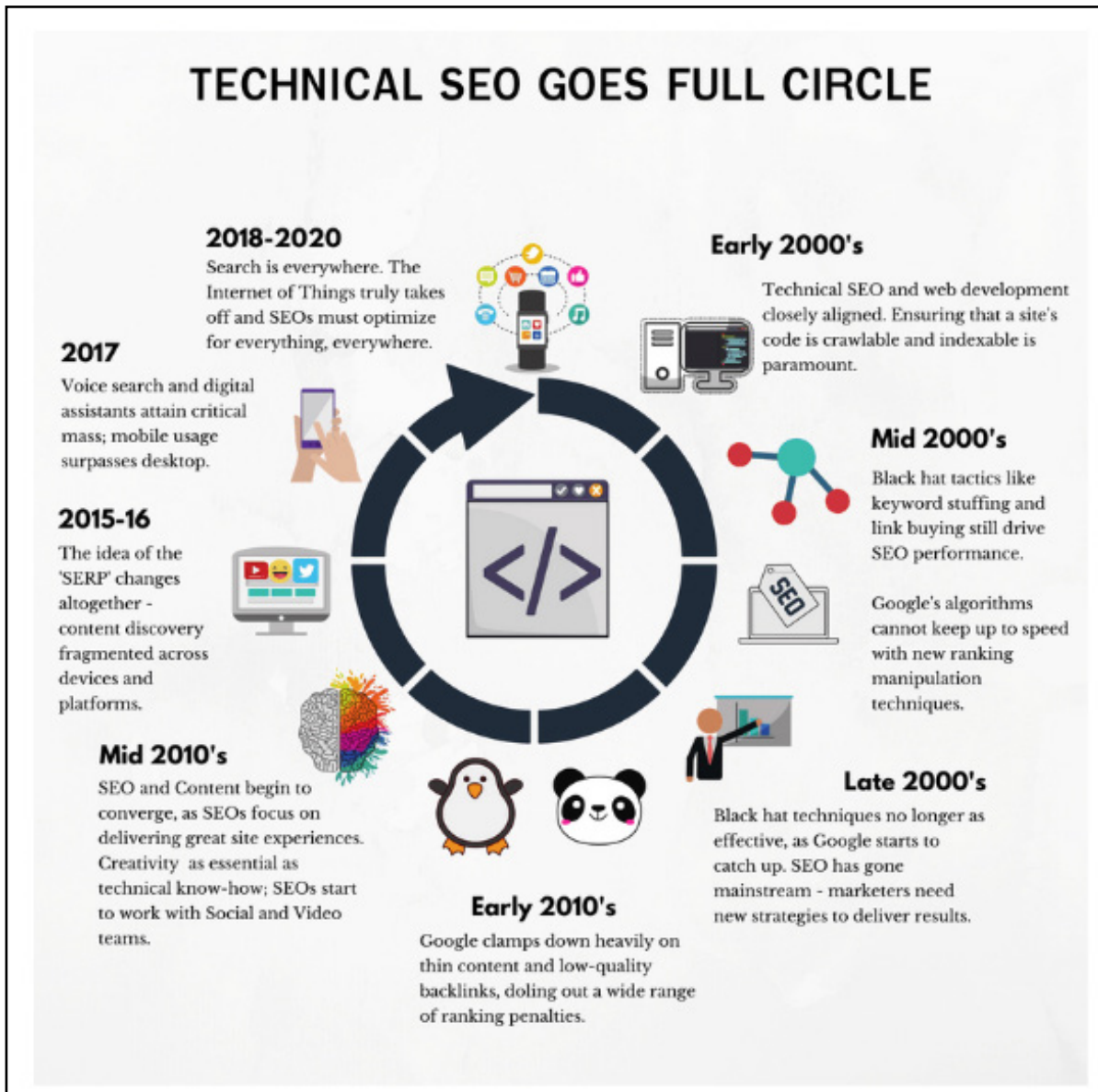
The **SERP** as we knew it is long gone; search is dynamic, visual, and everywhere now.

This has a very significant impact on organizations, as SEO is a collaborative discipline that requires a synthesis of multiple specialisms to achieve optimal results. At the heart of this lies the domain of **technical SEO**, which has remained the foundation upon which any successful strategy is built.



A Brief History of Technical SEO

All roads lead back to technical – it's how you now use your skills that has changed.





SEO has always entailed driving high-quality traffic through organic search.

The means of achieving this goal have altered significantly since the early days of SEO, when technical skills were dominant.

Crawlability was then – as it is now – a foremost consideration when setting up an SEO strategy.

Content was secondary – a vehicle to include keywords and improve rankings. This evolved over time to encompass **link building**, based on Google’s key early innovation of using links to evaluate and rank content.

The goal of marketers remained constant: to attract organic search traffic that converted on their website.

As a result, we endured a cat and mouse game with some marketers doing whatever it took to gain high search rankings.

As soon as Google caught up with keyword cloaking, black hat SEO practitioners moved on to link buying in an attempt to manipulate their rankings.

The **Panda** and **Penguin** algorithm updates put paid to a lot



of those murky tactics and even (briefly) raised the discussion of whether **SEO was dead.**

This question missed one key point.

As long as people are using search as a means to discover information, SEO will continue in rude health. Those discussions are a distant memory as we embrace modern SEO, especially its convergence with content marketing.

The industry has gone from strength to strength and the best strategies are now justly rewarded with increased search presence.

In the process, SEO has moved from an entirely rational discipline to something more rounded, including the typically “right-brained” domain of creative content. This has changed the shape of SEO departments and demanded collaboration with other digital marketing departments.

Technical SEO, for its part, now encompasses all search engine best practices and allows no room for manipulation. This specialism never went away, but it has seen a recent renaissance as senior marketers realize that it drives performance as well as crawler compliance.

There are four key areas to this:

- **Site Content:** Ensuring that content can be crawled and indexed by [all major search engines](#), in particular making use of log file analysis to interpret their access patterns and structured data to enable efficient access to content elements.
- **Structure:** Creating a site hierarchy and [URL structure](#) that allow both search engines and users to navigate to the most relevant content. This should also facilitate the flow of internal link equity through the site.
- **Conversion:** Identifying and resolving any blockages that prevent users from navigating through the site.
- **Performance:** A key development has been the evolution of technical SEO into a performance-related specialism. This has always been the case, but marketers of all stripes have realized that technical SEO is about a lot more than just “housekeeping.” Getting the three areas above in order will lead to better site performance through search and other channels, too.

Within this context, it is worth questioning whether “SEO” is even an adequate categorization for what we do anymore.

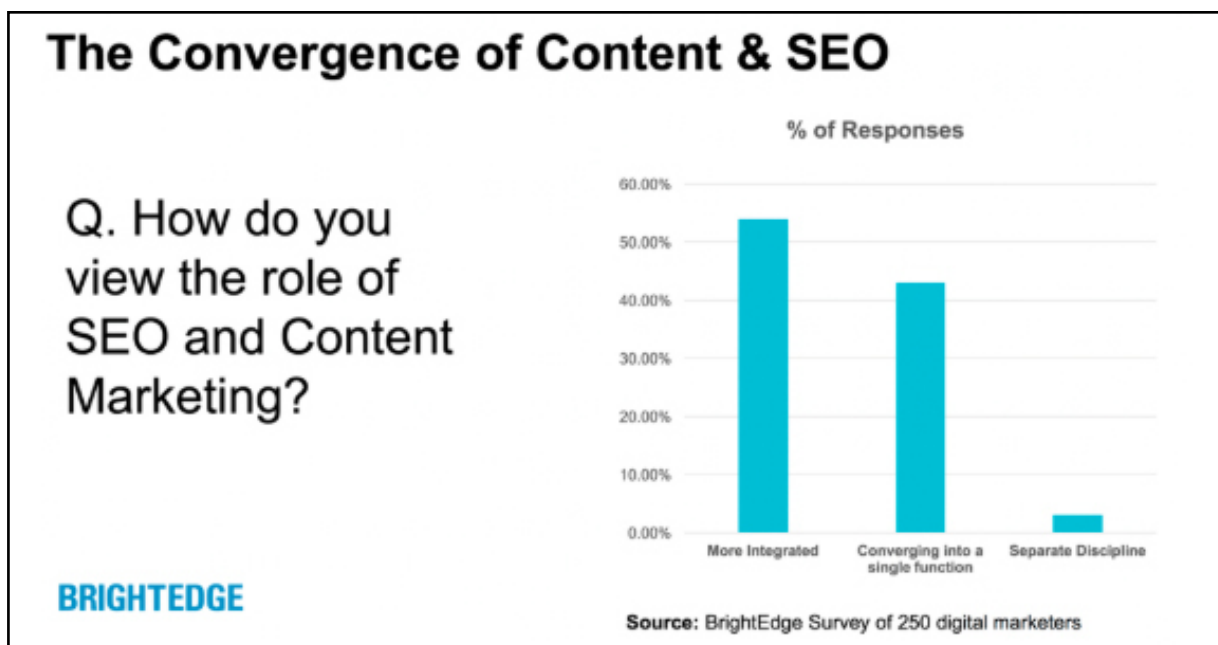
A New Approach: Site, Search & Content Optimization

The term “search engine optimization” is arguably no longer fit for purpose, as we extend our remit to include content marketing, conversion rate optimization, and user experience.

Our work includes:

- Optimizing the site for users.
- Ensuring accessibility of content for all major search engines and social networks.
- Creating content that engages the right audience across multiple marketing channels.

According to research from BrightEdge, only **3 percent** of 250 marketers surveyed believe SEO and content are separate disciplines.



We should therefore be looking at this set of skills as site, search, and content optimization – especially as the role of a search engine continues to evolve beyond the 10 blue links of old.

Our responsibility is to get in front of consumers wherever they are searching, which is an ever-changing set of contexts. This would be a more fitting depiction of a marketing channel that plays an increasingly pivotal role in digital and business strategy.

After all, when major technological trends develop, technical SEO pros are often at the forefront of innovation. This looks set to be further entrenched by recent industry developments.

Now that [**Accelerated Mobile Pages \(AMP\)**](#) and [**Progressive Web Apps \(PWAs\)**](#) are center stage, brands must ensure that their web presence meets the highest standards to keep pace with the modern consumer.



Being **“mobile-first”** has big implications for how we engage our audiences, but it is also a technological consideration. PWAs will soon be coming to Google Chrome on desktop, which is a further manifestation of the “mobile-first” approach to site experiences that we all need to adopt.

It would be hard to argue that these fit uniquely under the remit of ‘Search Engine Optimization’, and yet it is likely SEO pros that will lead to change within their respective organizations.

Brands need to think beyond search engines and imagine the new ways their content could – and should – be discovered by customers.

A different approach to SEO is required if we are to tap into the full potential of emerging consumer trends. That approach should expand to include site experience optimization, as well as traditional SEO techniques.

There are plentiful new opportunities for those who adapt; a process that can be accelerated by creating a collaborative working environment.

6 Thinking Hats & SEO

However we choose to label it, it should be clear that SEO has never existed in a vacuum. From its early symbiosis with web development to its latter-day convergence with content, SEO has always been about collaboration.

It is therefore helpful to consider frameworks that can bring this idea to life and bring together the specialist skills required for a modern organic search campaign.

We typically talk only about black hat and white hat in SEO (with the occasional mention of gray), but Edward de Bono's [Six Thinking Hats](#) approach can add structure to collaboration.

Each hat reflects a way of thinking and separates out the different functions required to achieve successful outcomes. These could be entirely different departments, different individuals, or even different mindsets for one person.

The objective is to improve the collaborative process, but also to erode the fallibility of subjectivity by approaching every challenge from all angles before progressing.





1. White Hat

A well-known term for most SEO pros, White Hat thinking in this context depends purely on facts, statistics, and data points. This is the most objective way of approaching a situation.

Who Should Wear This Hat?

Data analysts and analytics specialists are typically naturals at adopting this approach.

Why Is It Needed for SEO?

Looking purely at the data is a perfect starting point for discussion. It keeps everyone focused on the objective truths of cross-channel performance. Data without context is meaningless, of course, so this approach in isolation lacks the color needed to understand consumers.



2. Yellow Hat

The Yellow Hat approach brings optimism to the table, focusing on the potential benefits a strategy may bring for brands and the consumer.

Who Should Wear This Hat?

Anyone can be an optimist, so this could be a mindset that all parties take on for a period of time. Equally, this could be handed to one person as a responsibility; the key thing is to maintain some structure.

Why Is It Needed for SEO?

We tend to have a lot of ideas, so it is easy to jettison some of them before their full potential has been explored. Taking an alternative view allows for full exploration of an idea, even if only to retain some of its components.



3. Black Hat

The Black Hat is anathema to advanced SEO pros, but the concept does have value in this particular context. We can use this interchangeably with the “devil’s advocate” approach, where someone purposefully points out obstacles and dangers for the project.

Who Should Wear This Hat?

No one really, but be aware of the dangers of people offering SEO solutions and little transparency into the how. Keep an eye out for negative SEO attacks.



4. Red Hat

The Red Hat approach relates to feelings and emotions, often based on the gut reaction to an idea. This can be very beneficial for a digital project, as we can sometimes be overly rational in our data-driven approach.

Who Should Wear This Hat?

It can be helpful to assign this role to someone who works closely with the target audience, or who analyzes and interprets a lot of audience data.

Why Is It Needed for SEO?

When fighting for vital – and dwindling – consumer attention, first impressions matter. Content marketing campaigns can depend on getting this right, so it's worth listening to gut instinct sometimes.



5. Green Hat

The Green Hat brings creativity and spontaneity to the process, tackling challenges from a new perspective when possible. Where others see obstacles, this approach will see new opportunities.

Who Should Wear This Hat?

Anyone can be creative. However, it may be best to assign this role to someone who feels comfortable sharing their ideas with a group and is not easily disheartened if they don't take off!

Why Is It Needed for SEO?

There are best practices, but those only take us so far. They are a leveling force; new ideas are what really make the difference. In an industry, as nascent as ours, there are plenty of trails yet to be explored. The Green Hat brings that element of innovation to a discussion.

6. Blue Hat

The Blue Hat organized the thinking process and takes ultimate responsibility for bringing together the different strands into a cohesive whole.

Who Should Wear This Hat?

The project lead or the person closest to the brand's objectives can help keep things focused. Project managers also have a natural affinity for this role.

Why Is It Needed for SEO?

SEO is an increasingly diverse set of disciplines, which makes this role indispensable. To maximize the organic search opportunity, various departments need to be working in tandem on an ongoing basis. The Blue Hat keeps this collaboration going.

Actual hats are optional, but may help the adoption of this approach.

Regardless, these ways of thinking have a range of benefits across any organization:

- Opportunities to integrate more digital functions into the SEO process.
- Ways to learn new skills, both by doing and by observing.
- Integration of SEO best practices across more digital channels.
- A central role for SEO, without reducing the importance of other specialists.



Technical SEO Is Important Now More Than Ever

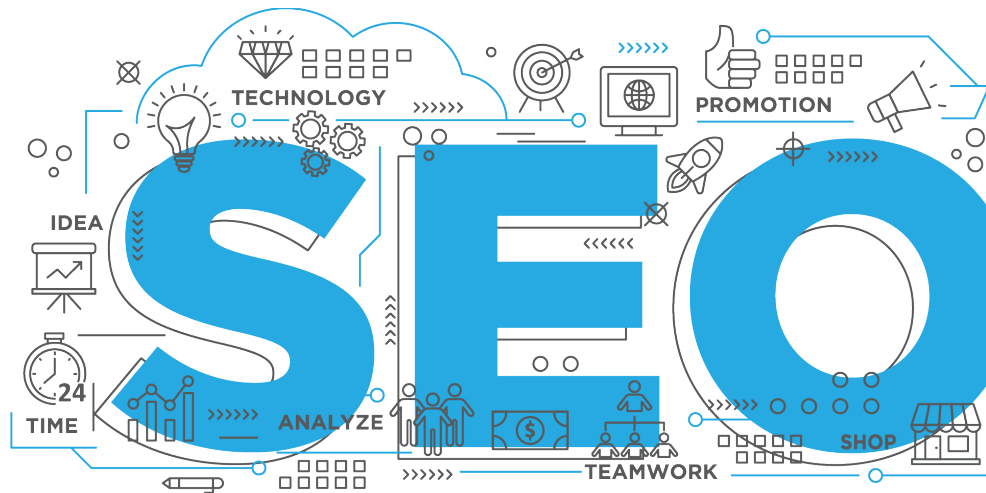
SEO has evolved to be part of something bigger and technical skills must be applied in a different manner.

If anything, it has expanded into a much more sophisticated and nuanced digital channel that has outgrown the “Search Engine Optimization” category.

The core tenets of organic search remain firmly in place, with technical SEO given overdue prominence as a driver of web, mobile and device performance.

SEO professionals are often at the forefront of technological innovations and this looks unlikely to change in a world of voice search, digital assistants, and Progressive Web Apps.

New approaches are required if we are to maximize this opportunity, however. That begins with the definition of what exactly SEO entails and extends to the ways we lead collaboration within our organizations.



The level of technical acumen needed for success has changed back to the levels it once was.

However, where and how you apply that knowledge is key to technical success. Focus your skills on optimizing:

- Your site.
- Mobile and desktop devices.
- Mobile apps.
- Voice search.
- VR.
- Agents.
- Vertical search engines (it's not just Google anymore – think Amazon for example).

The [AI revolution](#) is begging for more help from technical SEO professionals and data scientists to help drive it forward.

Mastering SEO fundamentals is only the bare minimum. If you really want to win against the competition, you must go beyond the basics.

If you act now and take a slightly different viewpoint on your role, organic search can assume a central role in both business strategy and cross-channel digital marketing.

Chapter 1

SEO-Friendly Hosting: 5 Things to Look for in a Hosting Company

SEJ
EBOOK

Written By
Amelia Willson
Owner at AWCopywriting



As SEO professionals, we have no shortage of things to worry about.

There are the old standbys: **backlinks**, **content creation**, **sitemaps** and **robots.txt** files.

And there's new(er) stuff to get excited about as well: voice search, featured snippets, the mobile-first index.

Amidst the noise, one factor often goes overlooked, even though it can impact your site's uptime and your page speed – both of which are essential elements for maintaining positive organic performance.

I'm talking about web hosting, folks.

The web host you choose determines the overall consistency of the site experience you offer organic visitors (and all visitors, for that matter).

If you want to prevent server errors and page timeouts – and **stop users from bouncing back** to Google – you need a solid web host you can rely on.

Ultimately, you want a web host that supports your organic efforts, rather than impeding them. Let's look at five key features that define an SEO-friendly web hosting company.

1. High Uptime Guarantee

Your host's uptime guarantee is arguably the most important factor in whether they're SEO-friendly.

Uptime refers to the percentage of the time your site is online and accessible. The higher your uptime, the less likely visitors will visit your site only to discover it's down, sending them back to the search engines and potentially risking your rankings in the process.

Better, more reliable hosts offer higher uptime guarantees.

For best results, choose a host with at least 99.9 percent uptime guarantee (or higher, if you can get it). That translates to roughly 1.44 minutes of downtime a day and 8.8 hours per year. Not bad.

However, be wary of any host that claims 100 percent uptime. There's always going to be some downtime. The key is to keep it as short as possible. That way, it won't affect your SEO performance.



2. Server Location

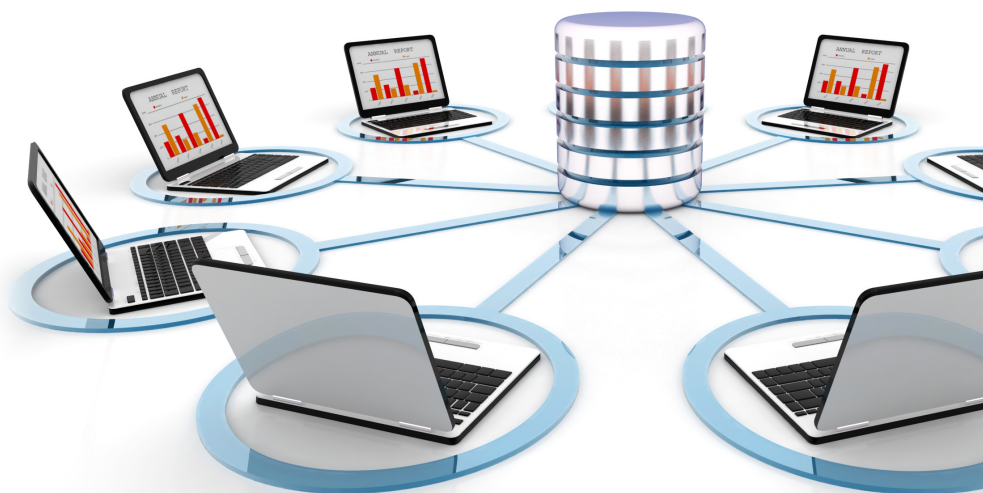
While uptime refers to your site content being accessible to users, your server location may dictate how quickly it's accessible to them.

If you're on a shared, VPS, or dedicated server hosting plan, your site lives on a physical server in a data center somewhere (as opposed to cloud hosting, where your data is housed in the cloud).

Ideally, you want that data center located as close as possible to the majority of your site visitors. The farther away your server is, the longer it can take for your site to load.

Server location can also look fishy to search engines, which may affect your SEO. If you operate in one country but use a host located halfway around the world, there may be something nefarious going on.

It goes without saying that servers themselves should also be fast, and that the host should further boost performance through a **[Content Delivery Network \(CDN\)](#)**.



3. Multiple Options

We all like options. You should enjoy them with your web hosting, too.

Beyond hosting itself, many hosting companies offer optional value-adds that can upgrade your site.

Here are some of the SEO-friendly ones you'll want to see:

- **Automatic backups:** If something ever goes wrong, you want a site backup you can quickly restore from. See if your host offers automatic backups for free or for an added cost.
- **SSL: HTTPS** has been a **ranking factor** for years now. If you haven't already transitioned to a secure site, you need to get your act together. Make sure your host supports SSL. Some even include them for free with your hosting package.
- **Multiple hosting plans:** As your site grows, your hosting needs are likely to change (this is a good thing!). Eventually, your traffic numbers may be big enough to warrant switching to your own dedicated server. This transition will be easier (and cheaper) if you don't have to switch hosting providers at the same time.

4. Good Reviews

Alright, let's say you're actually using this list to compare hosts. By this point, you've read through their hosting features, and it appears they're checking off all the right things.

Now it's time to validate that the marketing claims are true. Before you sign up with a host, take a few minutes to read their online reviews.

A caveat: The hosting space tends to attract more unhappy reviews than most.

If a barista messes up your coffee, you're unlikely to be bothered enough to write a scathing review for the cafe on Yelp.

But if your site goes down, even for a moment, or even if you were at fault (as can happen if you choose an inappropriate hosting plan for your traffic needs), you are going to be extremely angry with your host and tweet, post, and blog about it loudly and vociferously.

Unfortunately, that's just the nature of the business.



Having said that, you can still gather a lot of valuable information from reviews. Look for hosts that appear again and again on Top Web Hosts lists, and read the reviews to verify that the hosting plan you're considering is likely to give you what you need.

You won't have trouble finding these lists. A quick Google search for [best web hosting] delivered a slew of results from PCMag, CNET, and more:

[Best Web Hosting Services 2018 - Best Picks | PCMag.com](https://www.pcmag.com/article2/0,2817,2424725,00.asp)

<https://www.pcmag.com/article2/0,2817,2424725,00.asp> ▼

Apr 6, 2018 - These 10 **top web hosting** services give everyone from business owners to bloggers the support and tools they need to build an attractive, professional, and reliable web site, at any budget.

[HostGator Web Hosting](#) · [DreamHost Web Hosting](#) · [GoDaddy Web Hosting](#)

[The Best Web Hosting Providers for 2018 - CNET](https://www.cnet.com/web-hosting/)

<https://www.cnet.com/web-hosting/> ▼

Sep 7, 2017 - In this directory, we'll look at a few of the best web site hosting providers like **InMotion Hosting**, **Hostgator**, Web Hosting Pad, 1&1 Hosting and more. In this evaluation, we're featuring commercial hosting providers who offer WordPress, Shared Hosting, VPS and many more hosting products.

[The World's Best Web Hosting Brands Reviewed \[2018\]](https://www.whoishostingthis.com/hosting-reviews/)

<https://www.whoishostingthis.com/hosting-reviews/> ▼

Apr 26, 2018 - 1m+ words of web hosting reviews of the world's biggest & **best web hosts**, including BlueHost, HostGator, Siteground & more. Check out what our experts say, as well as reviews from thousands of real webmasters.

[The best web hosting services for 2018 | TechRadar](https://www.techradar.com/news/best-web-hosting-services)

<https://www.techradar.com/news/best-web-hosting-services> ▼

Apr 25, 2018 - Whatever size of website you have, this article will help you find the **best web hosting** service providers, and the best hosting deals for you. The first step is to identify what your needs are - with one eye on the future growth of your website - then choose an appropriate plan at the right price. Web hosting ...

5. Responsive Support Team

While you're reading through the reviews, pay special attention to how people talk about their support.

In the unlikely event that your site does go down, you want to be able to fix it immediately. Most often, that will involve speaking to a support person.

A good host will offer 24/7 support for free. Verify the operating hours of your potential host's support team, and see how exactly you'll be able to get in touch with them. Is there a phone number, live chat, or email?

Check out their social profiles, too. Web hosts who care about helping their customers tend to make customer support widely available on social media, perhaps even via dedicated support Twitter accounts.

Here's an example from Squarespace:



The image shows a screenshot of the Twitter profile for Squarespace Help (@SquarespaceHelp). The profile picture is a circular logo with the Squarespace 'S' symbol. The bio states: "Providing service updates and support. Please follow @squarespace for product and company news. Check status.squarespace.com for system status." The location is listed as New York, NY, and the website is support.squarespace.com. The account was joined in April 2010. The statistics show 103K tweets, 1,556 following, 35.2K followers, 459 likes, and 1 list. The tweets section shows a pinned tweet from February 22, 2017, and another tweet from May 3, 2017, both providing customer support information.

Tweets	Following	Followers	Likes	Lists
103K	1,556	35.2K	459	1

Squarespace Help @SquarespaceHelp

Providing service updates and support. Please follow @squarespace for product and company news. Check status.squarespace.com for system status.

New York, NY
support.squarespace.com
Joined April 2010

Tweets **Tweets & replies** **Media**

↓ Pinned Tweet

Squarespace Help @SquarespaceHelp · 22 Feb 2017
We're here for you and happy to help. Send us a tweet, direct message, or you can contact Customer Care. sqsp.link/1bDcY
178 6 63

Squarespace Help @SquarespaceHelp · May 3
[status] Resolved: This incident has been resolved and we have confirmed that all systems are operational. Thank you for your patience. stspg.io/194a97a23
1 1

Bonus: Easy-to-Use CMS

This one's not exactly related to hosting, but it's important nonetheless. Being able to easily create outstanding content is key for your SEO success. You know that.

So, you want a host that integrates with a CMS you're either already familiar with or you can easily learn. Otherwise, you're just making things hard on yourself!

Fortunately, most hosts today offer their own drag-and-drop content creation tools. Many also integrate with WordPress and other popular content management systems.

What Defines an SEO-Friendly Web Host?

Good, reliable web hosting is one of those things that runs in the background without you ever having to think about it. That, in essence, is an SEO-friendly web host.

Chapter 2

The Ultimate Guide for an SEO-Friendly URL Structure

SEJ
EBOOK

Written By
Clark Boyd
Founder, Candid Digital



First impressions count.

And when it comes to your website, your URLs are often the first thing Google and customers will see.

URLs are also the building blocks of an effective site hierarchy, passing equity through your domain and directing users to their desired destinations.

They can be tricky to correct if you don't plan ahead, as you can end up with endless redirect loops. Neither Google nor your site visitors will appreciate those.

So they are worth getting right. But getting URL structure right involves a complex blend of usability and accessibility factors, along with some good old-fashioned SEO.

Although there is no one-size-fits-all approach, there are some rules we can all follow to get the most out of our URLs and set our sites up for future SEO success.

1. Use Your Keywords

Every time you launch a page on your domain, it should have a purpose. Whether transactional, informational, or administrative, its reason for existence should be clear at the outset.

You'll want this page to be discovered by the right people (and crawlers), so you will incorporate some keyword research and include the relevant terms. The most descriptive of these — the term that gets to the nub of what this page is about — should be included in the URL, close to the root domain.

We'll deal with multiple pages that broadly tackle the same topic later, but for now, let's assume the simple example of a page that clearly handles one topic. Let's go for whiskey.

Generic example:

<https://example.com/topic>

Whiskey-based example:

<https://example.com/whiskey>

Even this isn't quite as simple as it seems, though.

Should we use
“whiskey” or
“whisky” as our
standard spelling?



Both are valid, with the former being an Irish spelling and the latter Scottish. The Irish spelling has been adopted in the U.S., but we'll need more proof before proceeding with that as our chosen variation.

The Moz Keyword Explorer is great for this sort of predicament, as it groups keywords together to give an estimate of the search volume for particular topics. In this era of vague keyword-level search volumes, it provides a nice solution.

The screenshot shows the Moz Keyword Explorer interface for the keyword 'whiskey'. The filters are set to 'include a mix of sources', 'yes with low lexical similarity', and 'any' for monthly volume. The results show 1,000 keywords in 99 groups. Three rows are highlighted with red boxes:

Keyword	Relevancy	Monthly Volume	Search
<input type="checkbox"/> whiskey Show all 207 grouped keywords	●●●●●●●●	70.8k-118k	
<input type="checkbox"/> whiskey brands Show all 3 grouped keywords	●●●●●●●●	9.3k-11.5k	🔍
<input type="checkbox"/> whisky Show all 25 grouped keywords	●●●●●●●●	30.3k-70.8k	🔍

The search volume is with “whiskey” and our site is based in the U.S., so let’s run with that.

2. Build a Sound Structure for the Future

Perhaps the biggest challenge we all face when defining a sitewide URL hierarchy is ensuring that it will still fit our purpose for years to come.

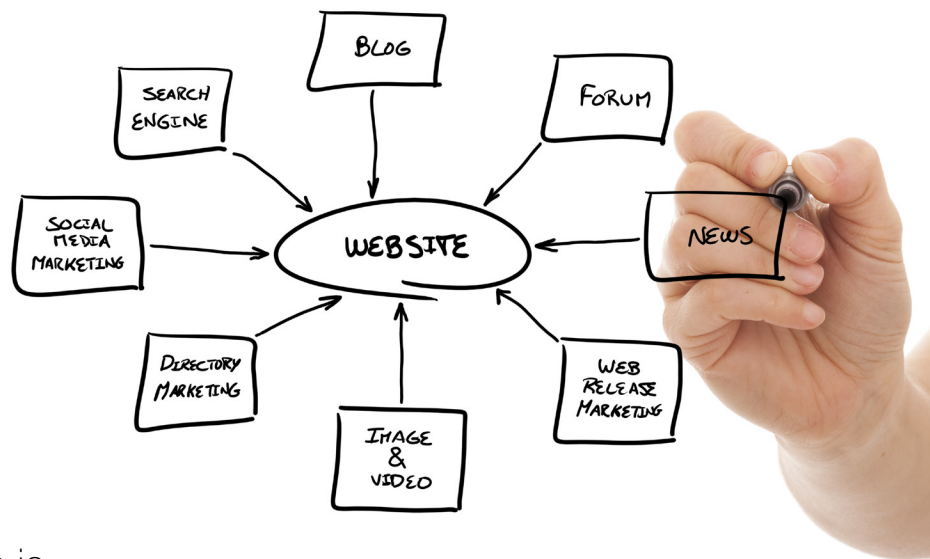
It is for this reason that some websites end up as a patchwork quilt of sub-domains and conflicting paths to arrive at similar products. This is poor from a user's perspective, but it also sends confusing signals to Google about how you categorize your product offering.

An example of this would be:

<https://example.com/whiskey/irish-whiskey/jameson>

<https://example.com/bushmills>

The first URL flows logically from domain to category to sub-category to product. The second URL goes from domain to product. Hierarchically, both products should sit at the same level in the site and the Jameson example is better for SEO and users.





We encounter this a lot, though. Why?

It can be a simple lack of communication, with a product team launching a new item straight onto the site without consulting other parties. It can also be down to a failure of future planning.

Either way, it's essential to lay out your structure in advance. Work together with different teams to understand the future direction of the business, then add your SEO knowledge to shape the site architecture. It will rarely be perfect, but the more you plan, the fewer errors you will have to undo down the line.

3. Avoid Superfluous Words & Characters

As a rule of thumb, make sure a user can understand what your page is about by looking at the URL. That means you don't need to include every single preposition or conjunction.

Words like “and” or “the” are just distractions and can be stripped out of the URL altogether. Just as users can understand what a topic is about without these short words, Google will derive all the meaning it requires too.

You should also avoid keyword repetition within URLs. Adding the same keyword multiple times in the hope of increasing your ranking chances will only lead to a spammy URL structure.

An example of this unnecessary repetition would be:

<https://domain.com/whiskey/irish-whiskey/jameson-irish-whiskey/jameson-irish-whiskey-history>

The first two uses of the main keyword make sense, but the third and fourth are overkill.

A few additional points to bear in mind on this topic:

- **Case Sensitivity:** It is surprisingly common to find multiple versions of the same URL, with one all in lower case and the others using occasional capital letters. Use canonical tags to mark the lower-case URL as the preferred version or, if possible, use permanent redirects.
- **Hashes:** These can be useful to send users to a specific section of a page, but restrict their use in other circumstances if possible. If the content users are sent to after the # symbol is unique, make it available via a simple URL instead.
- **Word Delimiters:** Stick with hyphens to separate words within your URL strings. Underscores will serve to join two words together, so be wary of using these.
- **URL Length:** After 512 pixels, Google will truncate your URL in search results pages. A good rule of thumb is to keep your URLs as short as you can, without losing their general meaning.

4. Minimize Dynamic URL Strings

This one can be harder than it sounds, depending on the content management system you use.

Some e-commerce platforms will automatically spit out character strings that leave you with URLs like: `https://domain.com/cat/?cid=7078`.

These are a bit unsightly and they also go against the rules we've been outlining above. We want static URLs that include a logical folder structure and descriptive keywords.

Although search engines have no problem crawling or indexing either variant, for SEO-based reasons it's better to use static URLs rather than dynamic ones. The thing is, static URLs contain your keywords and are more user-friendly since one can figure out what the page is about just by looking at the static URL's name.

So how do we get around this? You can use rewrite rules if your web server runs Apache, and some tools like [this one](#) from Generate It are helpful. There are different fixes for different platforms (some more complex than others).

Some web developers make use of relative URLs, too. The problem with relative URLs for SEO is that they are dependent on



 `http://www.`

the context in which they occur. Once the context changes, the URL may not work. For SEO, it's better to use absolute URLs instead of relative ones, since the former are what search engines **prefer**.

Now, sometimes different parameters can be added to the URL for analytics tracking or other reasons (such as **sid, utm**, etc.) To make sure that these parameters don't make the number of URLs with duplicate content grow over the top, you can do either of the following:

Ask Google to disregard certain URL parameters in Google Search Console in Configuration > URL Parameters.

See if your content management system allows you to solidify URLs with additional parameters with their shorter counterparts.

5. Consolidate the Different Versions of Your Site

As a rule, there are two major versions of your domain indexed in search engines: the www and the non-www version of it. We can add to this the complexity of having a secure (https) and non-secure (HTTP) version too, with Google giving preference to the former.

Most SEOs use the **301 redirect** to point one version of their site to the other (or vice versa).

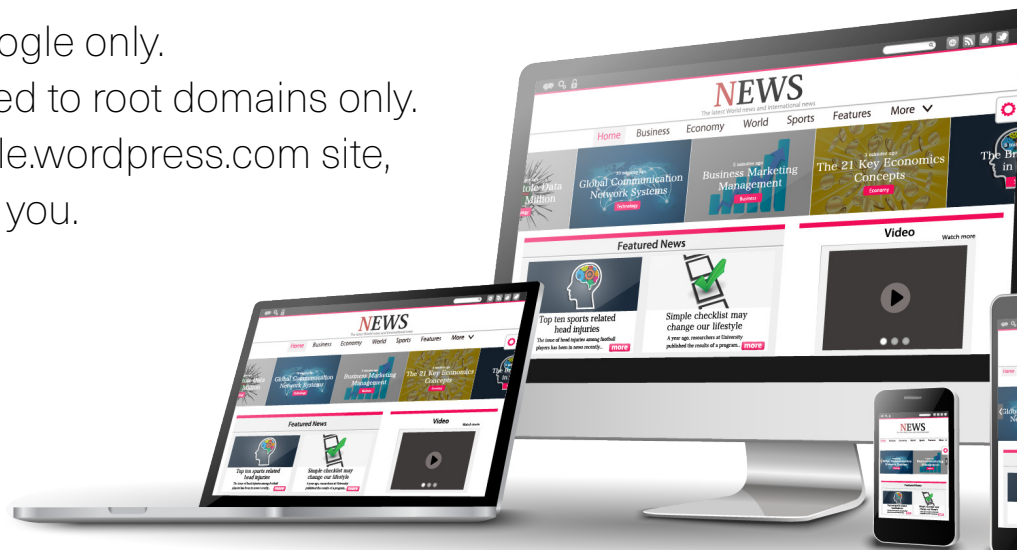
This tells search engines that a particular URL has moved permanently to another destination.

Alternatively (for instance, when you can't do a redirect), you can specify your preferred version in Google Search Console in Configuration > Settings > Preferred Domain. However, this has certain drawbacks:

This takes care of Google only.

This option is restricted to root domains only.

If you have an example.wordpress.com site, this method is not for you.



But why worry about the www vs non-www issue in the first place? The thing is, some of your backlinks may be pointing to your www version, while some could be going to the non-www version.

To ensure all versions' SEO value is consolidated, it's better to explicitly establish this link between them. You can do this via the 301 redirect, in Google Search Console, or by using a canonical tag, the latter of which we will look at in more detail below.

6. Make Correct Use of Canonical Tags

So, canonical tags. These are a very helpful piece of code when you have multiple versions of what is essentially the same page. By adding a canonical tag, you can tell Google which one is your preferred version.

Note: The canonical tag should be applied only with the purpose of helping search engines decide on your canonical URL. For redirection of site pages, use redirects. And, for **paginated content**, it makes sense to employ rel="next" and rel="prev" tags in most cases.

Canonical tags are useful for just about any website, but they are particularly powerful for online retailers.

For example, on Macy's website, I can go to the Quilts & Bedspreads page directly by using the URL (<https://www.macys.com/shop/bed-bath/quilts-bedspreads>), or I can take different routes from the homepage:

- I can go to Homepage >> Bed& Bath >> Quilts & Bedspreads. The following URL with my path recorded is generated:

<https://www.macys.com/shop/bed-bath/quilts-bedspreads?id=22748&edge=hybrid>

- Or I can go to Homepage >> For the Home >> Bed & Bath >> Bedding >> Quilts & Bedspreads. The following URL is generated:

https://www.macys.com/shop/bed-bath/quilts-bedspreads?id=22748&cm_sp=us_hdr_-_bed-%26-bath_-_22748_quilts-%26-bedspreads_COL1

Now, all three URLs lead to the same content. And if you look into the code of each page, you'll see the following tag in the head element:

```
4 <head>
5 <title>Quilts and Bedspreads - Macy's</title>
6 <meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
7 <meta http-equiv="generator" content="JACPKMALPHTCSJDTCR" />
8 <meta http-equiv="X-UA-Compatible" content="IE=edge" />
9 <link href="/favicon.ico" rel="SHORTCUT ICON" >
10 <meta name="format-detection" content="telephone=no" />
11 <link rel="canonical" href="https://www.macys.com/shop/bed-bath/quilts-bedspreads?id=22748" />
```

As you see, for each of these URLs, a canonical URL is specified, which is the cleanest version of all the URLs in the group:<https://www.macys.com/shop/bed-bath/quilts-bedspreads?id=22748>

What this does is, it funnels down the SEO value each of these three URLs might have to one single URL that should be displayed in the search results (the canonical URL). Normally search engines do a pretty good job identifying canonical URLs themselves, but, as Susan Moskwa once [wrote](#) at Google Webmaster Central:

“If we aren’t able to detect all the duplicates of a particular page, we won’t be able to consolidate all of their properties. This may dilute the strength of that content’s ranking signals by splitting them across multiple URLs.”

7. Incorporate Topical Authority

In Google's own [Search Quality Evaluators Guidelines](#) (a must-read document for all SEOs!), there are clear references to both main content and supplementary content.

Main content will be your lead page in each section that really sets out what your category is all about. It will set out your stall as a relevant source for a topic. Supplementary content provides, as the name suggests, additional information that helps users navigate the topic and reach informed decisions.



URL structure is an essential component of getting this right.

So, let's go back to our whiskey example to see how we might tackle this. Our site is e-commerce focused and we want to sell the product, of course. However, going for the jugular and only pushing out product pages is tantamount to SEO tunnel vision.

Our initial research from Moz Keyword Explorer is a great resource as we make these plans. Below, I have exported the keyword list and reduced it to the highest-volume topics. From here, we can start to decide what might qualify as a topic for a main content or supplementary content page.

Keyword	Main/Supplementary	Relevancy	Min Volume	Max Volume
whiskey	Main	5	70801	118000
whisky	Main	5	30301	70800
whiskey sour	Supplementary	5	30301	70800
scotch	Main	5	30301	70800
rye whiskey	Main	3	11501	30300
whiskey brands	Supplementary	5	9301	11500
whiskey neat	Supplementary	3	6501	9300
whisky vs whiskey	Supplementary	5	1701	2900
types of whiskey	Supplementary	3	1701	2900
whiskey types	Supplementary	5	851	1700
whiskeys	Main	5	851	1700
scotch whisky	Main	3	851	1700
whiskey vs whisky	Supplementary	3	851	1700
whiskey club	Supplementary	3	501	850
how is whiskey made	Supplementary	3	501	850
whiskey prices	Supplementary	5	201	500
barley whiskey	Main	3	51	100
whiskey finder	Supplementary	3	51	100

This is a simplified example and just a first step, of course.

However, it is worth noting that this approach goes further than just category > sub-category > product. By thinking in terms of main content and supplementary content, a product is just as likely to qualify as main content as a category is. The question is more about which topics consumers want us to elaborate on to help them make choices.

From here, we can dig into some of these topics and start to flesh out what each hub might look like.

Some clear opportunities already stand out to create content and rank via rich snippets. People want to know how whiskey is made, what different varieties exist, and of course, whether it's spelled 'whiskey' or 'whisky'. This could be the beginning of a business case to create a whiskey tasting guide or a 'history of whiskey' content hub on the site.

Combined with ranking difficulty metrics, business priorities, and content production capabilities, this approach will soon take shape as a site hierarchy and opportunity analysis.

For our whiskey example, it might start to comprise the following structure:

<https://domain.com/whiskey/whiskey-tasting-guide>

<https://domain.com/whiskey/whiskey-tasting-guide/how-to-taste-whiskey>

<https://domain.com/whiskey/whiskey-tasting-guide/how-is-whiskey-made>

<https://domain.com/whiskey/whiskey-tasting-guide/barley-whiskey>

Again, there are decisions to make.

In the last URL, one could argue that the tasting guide page for barley whiskey should sit under the barley whiskey sub-category page in the site hierarchy. Barley whiskey has been earmarked as 'main content' in my spreadsheet, after all. The choice here comes down to where we want to consolidate value; dispersing that value would reduce our chances of ranking for any 'tasting guide' terms.

These are exactly the kinds of decisions that can lead to a confused structure if a consistent logic is not followed.

All of this will contribute to your topical authority and increase site visibility.

This type of content often already exists on site, too. I am not claiming anything revolutionary by saying a website should have lots of useful information, after all. However, the structure of this content and how entities are semantically linked to each other makes the difference between success and failure.

This can be used as a 'quick win' tactic and it tends to be received well by all parties. Updating and moving existing content will always be an easier sell than asking for an all-new content hub.

8. Create an XML Sitemap

Once you've ticked off all of the above, you'll want to make sure search engines know what's going on with your website. That's where sitemaps come in handy — particularly XML sitemaps.

An XML Sitemap is not to be confused with the HTML sitemap. The former is for the search engines, while the latter is mostly designed for human users (although it has other uses too).

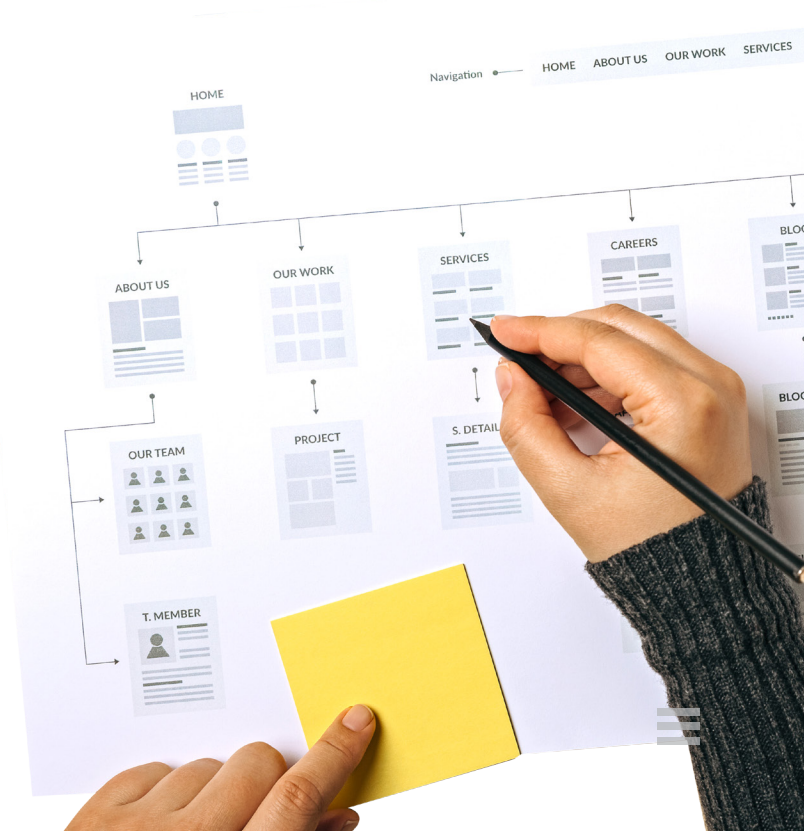
So what is an XML Sitemap? In plain words, it's a list of your site's URLs that you submit to the search engines.

This serves two purposes:

1. This helps search engines find your site's pages more easily.
2. Search engines can use the sitemap as a reference when choosing canonical URLs on your site.

Picking a preferred (canonical) URL becomes necessary when search engines see duplicate pages on your site, as we saw above.

So, as they don't want any duplicates in the search results, search engines use a special algorithm to identify



duplicate pages and pick just one URL to represent the group in the search results. Other web pages just get filtered out.

Now, back to sitemaps. One of the criteria search engines may use to pick a canonical URL for the group of web pages is whether this URL is mentioned in the website's sitemap.

So, what web pages should be included in your sitemap? For purely SEO reasons, it's recommended to include only the web pages you'd like to show up in search. You should include a more comprehensive account of your site's URLs within the HTML sitemap.

Summary

An SEO-friendly URL structure is the following things:

Easy to read: Users and search engines should be able to understand what is on each page just by looking at the URL.

Keyword-rich: Keywords still matter and your target queries should be within URLs. Just be wary of overkill; extending URLs just to include more keywords is a bad idea.

Consistent: There are multiple ways to create an SEO-friendly URL structure on any site. It's essential that, whatever logic you choose to follow, it is applied consistently across the site.

Static: Dynamic parameters are rarely an SEO's best friend, but they are quite common. Where possible, find a solution that allows your site to render static URLs instead.

Future-proof: Think ahead when planning your site structure. You should minimize the number of redirects on your domain, and it's easier to do this if you don't require wholesale changes to URLs.

Comprehensive: Use the concepts of main content and supplementary content to ensure you have adequate coverage for all relevant topics. This will maximize your site's visibility.

Supported by data: It normally requires buy-in from a lot of stakeholders to launch or update a particular site structure. Numbers talk, so make use of search and analytics data to support your case.

Submitted to search engines: Finally, create an XML sitemap containing all of the URLs that you want to rank via SEO and submit it to search engines. That will ensure all your hard work gets the reward it deserves.

Chapter 3

How to Use XML Sitemaps to Boost SEO

SEJ
EBOOK

Written By
Jes Scholz
International Digital Director, Ringier



As the web evolves, so too does Google and SEO.

This means what is considered best practice is often in flux. What may have been good counsel yesterday, is not so today.

This is especially true for sitemaps, which are almost as old as SEO itself.

The problem is, when every man and their dog has posted answers in forums, published recommendations on blogs and amplified opinions with social media, it takes time to sort valuable advice from misinformation.

So while most of us share a general understanding that submitting a sitemap to Google Search Console is important, you may not know the intricacies of how to implement them in a way that drives SEO key performance indicators (KPIs).

Let's clear up the confusion around best practices for sitemaps today.

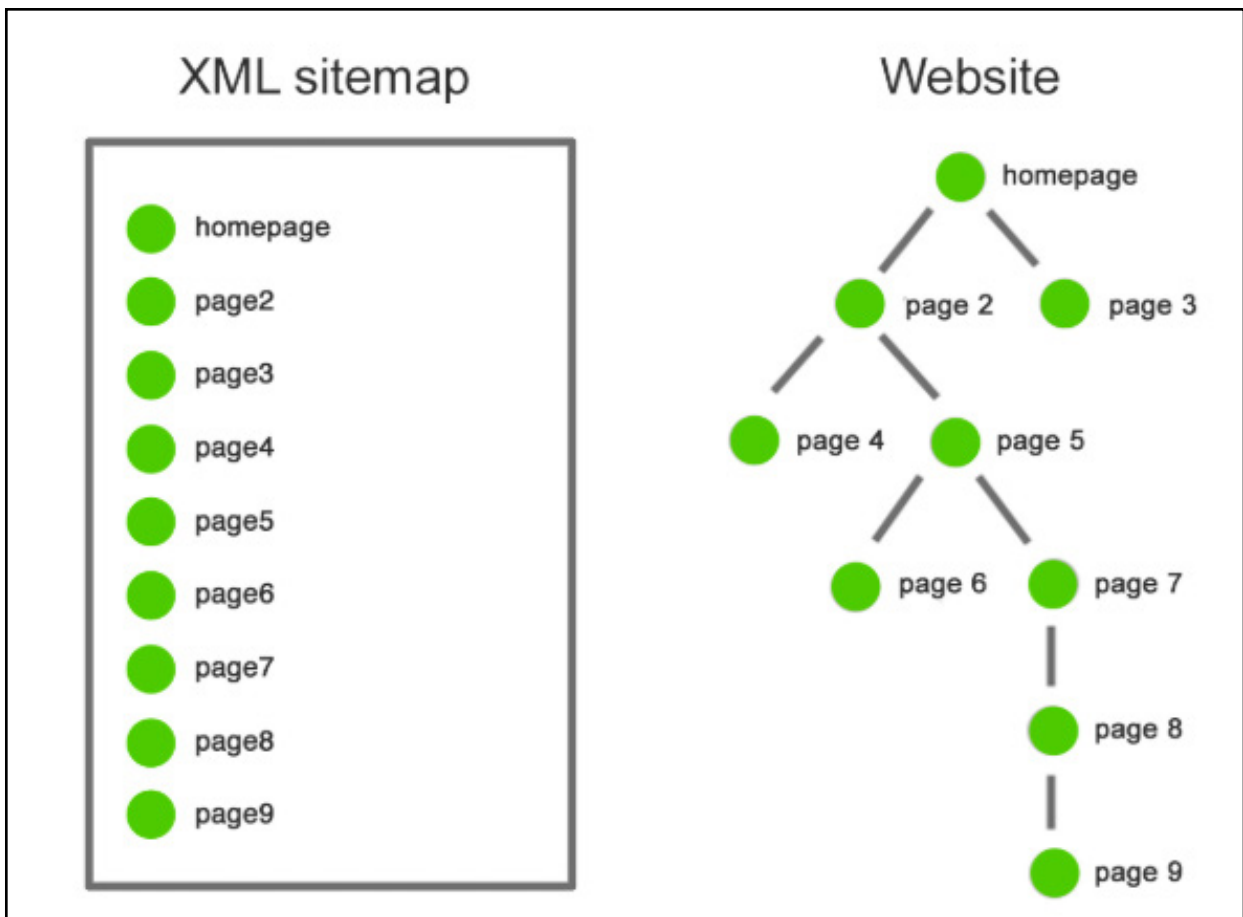
In this article we cover:

- What is an XML sitemap
- XML sitemap format
- Types of sitemaps
- XML sitemap indexation optimization
- XML sitemap best practice checklist

What Is an XML Sitemap

In simple terms, an XML sitemap is a list of your website's URLs.

It acts as a roadmap to tell search engines what content is available and how to reach it.



In the example above, a search engine will find all nine pages in a sitemap with one visit to the XML sitemap file.

On the website, it will have to jump through five internal links to find page 9.

This ability of an XML sitemap to assist crawlers in faster indexation is especially important for websites that:

- Have thousands of pages and/or a deep website architecture.
- Frequently add new pages.
- Frequently change content of existing pages.
- Suffer from weak internal linking and orphan pages.
- Lack a strong external link profile.



Side note: Submitting a sitemap with noindex URLs can also speed up deindexation. This can be more efficient than removing URLs in Google Search Console if you have many to be deindexed. But use this with care and be sure you only add such URLs temporarily to your sitemaps.

Key Takeaway

Even though search engines can technically find your URLs without it, by including pages in an XML sitemap you're indicating that you consider them to be quality landing pages.

While there is no guarantee that an XML sitemap will get your pages crawled, let alone indexed or ranked, submitting one certainly increases your chances.

XML Sitemap Format

A one-page site using all available tags would have this XML sitemap:

```
<?xml version="1.0" encoding="UTF-8"?>

<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

  <url>

    <loc>https://www.example.com/</loc>

    <lastmod>2018-08-24</lastmod>

    <changefreq>weekly</changefreq>

    <priority>0.5</priority>

  </url>

</urlset>
```

But how should an SEO use each of these tags? Is all the metadata valuable?

Loc (a.k.a. Location) Tag

This compulsory tag contain the absolute, canonical version of the URL location.

It should accurately reflect your site protocol (http or https) and if you have chosen to include or exclude www.

For international websites, this is also where you can **implement your hreflang handling.**

By using the xhtml:link attribute to indicate the language and region variants for each URL, you reduce page load time, which the other implementations of link elements in the <head> or HTTP headers can't offer

Yoast has an epic **post** on hreflang for those wanting to learn more.

Lastmod (a.k.a. Last Modified) Tag

An optional but highly recommended tag used to communicate the file's last modified date and time.

John Mueller acknowledged Google does use the lastmod metadata to understand when the page last changed and if it should be crawled. Contradicting advice from **Illyes** in 2015.



The last modified time is especially critical for content sites as it assists Google to understand that you are the original publisher.

It's also powerful to communicate freshness, be sure to update modification date only when you have made meaningful changes.

Trying to [trick search engines that your content is fresh](#), when it's not, may result in a Google penalty.

Changefreq (a.k.a. Change Frequency) Tag

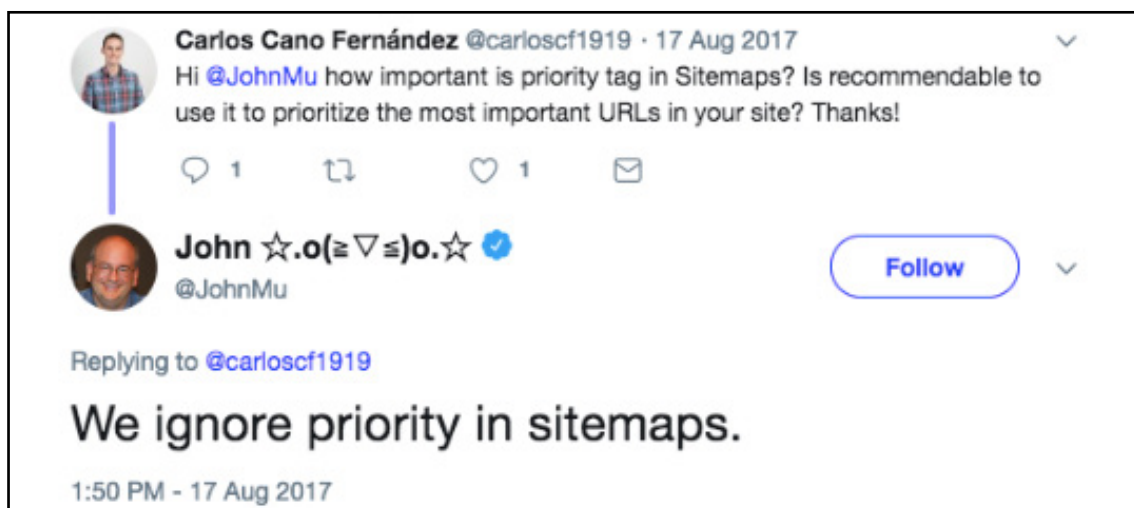
Once upon a time, this optional tag hinted how frequently content on the URL was expected to change to search engines.

But [Mueller has stated](#) that “change frequency doesn't really play that much of a role with sitemaps” and that “it is much better to just specify the time stamp directly”.

Priority Tag

This optional tag that ostensibly tells search engines how important a page is relative to your other URLs on a scale between 0.0 to 1.0.

At best, it was only ever a hint to search engines and both Mueller and [Illyes](#) have clearly stated they ignore it.



Key Takeaway

Your website needs an XML sitemap, but not necessarily the priority and change frequency metadata.

Use the lastmod tags accurately and focus your attention on ensuring you have the right URLs submitted.

Types of Sitemaps

There are many different types of sitemaps. Let's look at the ones you actually need.

XML Sitemap Index

XML sitemaps have a couple of limitations:

- A maximum of 50,000 URLs.
- An uncompressed file size limit of 50MB.

Sitemaps can be compressed using gzip (the file name would become something similar to sitemap.xml.gz) to save bandwidth for your server. But once unzipped, the sitemap still can't exceed either limit.

Whenever you exceed either limit, you will need to split your URLs across multiple XML sitemaps.

Those sitemaps can then be combined into a single XML sitemap index file, often named `sitemap-index.xml`. Essentially, a sitemap for sitemaps.

For exceptionally large websites who want to take a more granular approach, you can also create multiple sitemap index files.

For example:

- `sitemap-index-articles.xml`
- `sitemap-index-products.xml`
- `sitemap-index-categories.xml`

But be aware that you cannot nest sitemap index files.

For search engines to easily find every one of your sitemap files at once, you will want to:

- Submit your sitemap index(s) to Google Search Console and Bing Webmaster Tools.
- Specify your sitemap index URL(s) in your `robots.txt` file. Pointing search engines directly to your sitemap as you welcome them to crawl.

```
User-agent: *  
Disallow:  
  
Sitemap: https://www.example.com/sitemap-index.xml.gz
```

You can also submit sitemaps by pinging them to Google.

But beware:

Google no longer pays attention to hreflang entries in “unverified sitemaps”, which [Tom Anthony](#) believes to mean those submitted via the ping URL.

XML Image Sitemap

Image sitemaps were designed to improve the indexation of image content.

In modern-day SEO, however, images are embedded within page content, so will be crawled along with the page URL.

Moreover, it's best practice to utilize JSON-LD schema.org/ [ImageObject](#) markup to call out image properties to search engines as it provides more attributes than an image XML sitemap.

Because of this, an XML image sitemap is unnecessary for most websites. Including an image sitemap would only waste crawl budget.

The exception to this is if images help drive your business, such as a stock photo website or ecommerce site gaining product page sessions from Google Image search.

Know that images don't have to be on the same domain as your website to be submitted in a sitemap. You can use a CDN as long as it's verified in Search Console.

XML Video Sitemap

Similar to images, if videos are critical to your business, submit an XML video sitemap. If not, a video sitemap is unnecessary.

Save your crawl budget for the page the video is embedded into, ensuring you markup all videos with JSON-LD as a schema.org/VideoObject.

Google News Sitemap

Only sites registered with Google News should use this sitemap.

If you are, include articles published in the last two days, up to a limit of 1,000 URLs per sitemap, and update with fresh articles as soon as they're published.

Contrary to some online advice, Google News sitemaps don't support image URL.

[Google recommends](#) using schema.org image or og:image to specify your article thumbnail for Google News.

Mobile Sitemap

This is not needed for most websites.

Why? Because [Mueller confirmed](#) mobile sitemaps are for feature phone pages only. Not for smartphone-compatibility.

So unless you have unique URLs specifically designed for featured phones, a mobile sitemap will be of no benefit.

HTML Sitemap

XML sitemaps take care of search engine needs. HTML sitemaps were designed to assist human users to find content.

The question becomes, if you have a good user experience and well crafted internal links, do you need a HTML sitemap?

Check the page views of your HTML sitemap in Google Analytics. Chances are, it's very low. If not, it's a good indication that you need to improve your website navigation.

HTML sitemaps are generally linked in website footers. Taking link equity from every single page of your website.

Ask yourself. Is that the best use of that link equity? Or are you including HTML sitemap as a nod to legacy website best practices?

If few humans use it. And search engines don't need it as you have strong internal linking and an XML sitemap. Does that HTML sitemap have a reason to exist? I would argue no.

Dynamic XML Sitemap

Static sitemaps are simple to create using a tool such as Screaming Frog.

The problem is, as soon as you create or remove a page, your sitemap is outdated. If you modify the content of a page, the sitemap won't automatically update the lastmod tag.

So unless you love manually creating and uploading sitemaps for every single change, it's best to avoid static sitemaps.

Dynamic XML sitemaps, on the other hand, are automatically update by your server to reflect relevant website changes as they occur.

To create a dynamic XML sitemap:

- Ask you developer to code a custom script, being sure to provide clear specifications
- Use a dynamic sitemap generator tool
- Install a plugin for your CMS, for example the Yoast SEO plugin for Wordpress

Key Takeaway

Dynamic XML sitemaps and a sitemap index are modern best practice. Mobile and HTML sitemaps are not.

Use image, video and Google News sitemaps only if improved indexation of these content types drive your KPIs.

XML Sitemap Indexation Optimization

Now for the fun part. How do you use XML sitemaps to drive SEO KPIs.

Only Include SEO Relevant Pages in XML Sitemaps

An XML sitemap is a list of pages you recommend to be crawled, which isn't necessarily every page of your website.

A search spider arrives at your website with an "allowance" for how many pages it will crawl.

The XML sitemap indicates you consider the included URLs to be more important than those that aren't blocked but aren't in the sitemap.

You are using it to tell search engines "I'd really appreciate it if you'd focus on these URLs in particular".

Essentially, it helps you use crawl budget effectively.

By including only SEO relevant pages, you help search engines crawl your site more intelligently in order to reap the benefits of better indexation.

You should exclude:

- Non-canonical pages.
- Duplicate pages.
- Paginated pages.
- Parameter or session ID based URLs.
- Site search result pages.
- Reply to comment URLs.
- Share via email URLs.
- URLs created by filtering that are unnecessary for SEO.
- Archive pages.
- Any redirections (3xx), missing pages (4xx) or server error pages (5xx).
- Pages blocked by robots.txt.
- Pages with noindex.
- Resource pages accessible by a lead gen form (e.g. white paper PDFs).
- Utility pages that are useful to users, but not intended to be landing pages (login page, contact us, privacy policy, account pages, etc.).

I want to share an example from [Michael Cottam](#) about prioritising pages:

Say your website has 1,000 pages. 475 of those 1,000 pages are SEO relevant content. You highlight those 475 pages in an XML sitemap, essentially asking Google to deprioritize indexing the remainder.

Now, let's say Google crawls those 475 pages, and algorithmically decides that 175 are "A" grade, 200 are "B+", and 100 "B" or "B-". That's a strong average grade, and probably indicates a quality website to which to send users.

Contrast that against submitting all 1,000 pages via the XML sitemap. Now, Google looks at the 1,000 pages you say are SEO relevant content, and sees over 50 percent are "D" or "F" pages. Your average grade isn't looking so good anymore and that may harm your organic sessions.

But remember, Google is going to use your XML sitemap only as a clue to what's important on your site.

Just because it's not in your XML sitemap doesn't necessarily mean that Google won't index those pages.

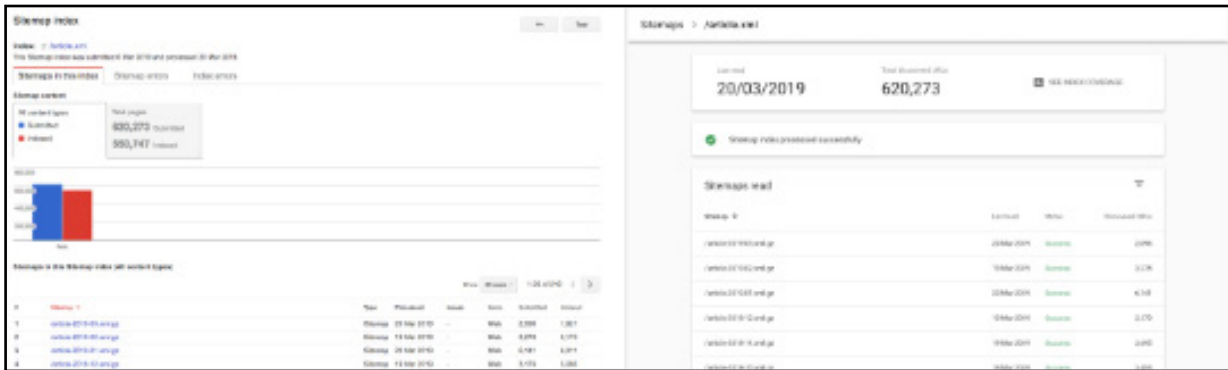
When it comes to SEO, overall site quality is a key factor.

To assess the quality of your site, turn to the sitemap related reporting in Google Search Console (GSC).

Key Takeaway

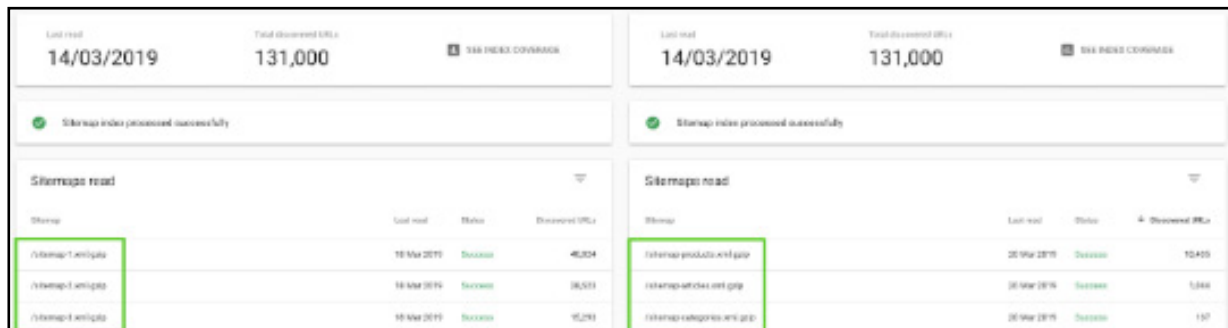
Manage crawl budget by limiting XML sitemap URLs only to SEO relevant pages and invest time to reduce the number of low quality pages on your website.

Fully Leverage Sitemap Reporting



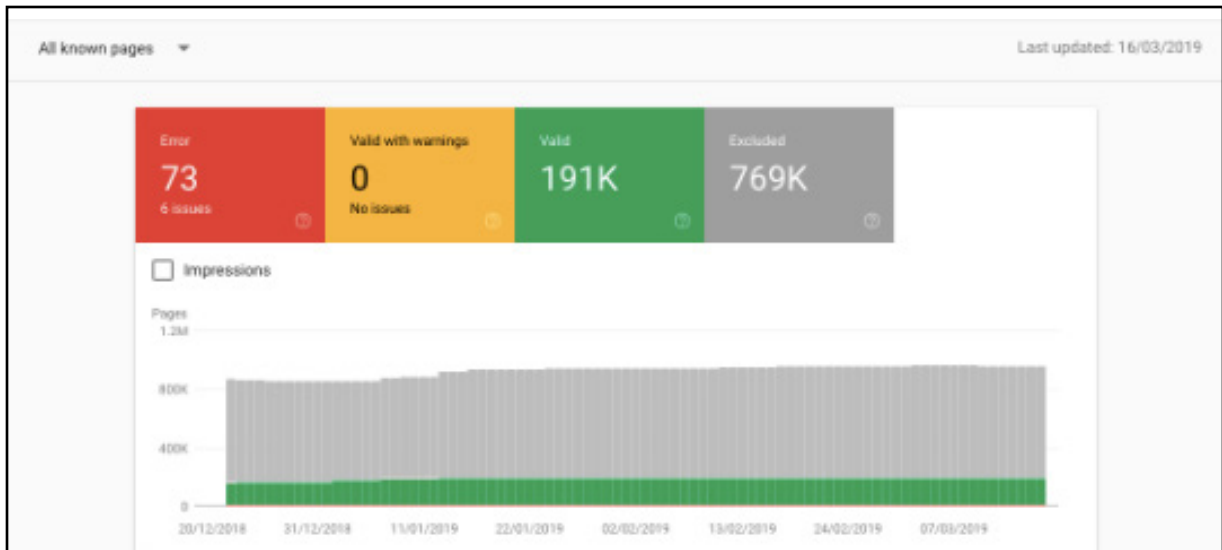
The sitemaps section in the new Google Search Console is not as data rich as what was previously offered.

It's primary use now is to confirm your sitemap index has been successfully submitted.



If you have chosen to use descriptive naming conventions, rather than numeric, you can also get a feel for the number of different types of SEO pages that have been “discovered” - aka all URLs found by Google via sitemaps as well as other methods such as following links.

In the new GSC, the more valuable area for SEOs in regard to sitemaps is the Index Coverage report.



The report will default to “All known pages”. Here you can:

- Address any “Error” or “Valid with warnings” issues. These often stem from [conflicting robots directives](#). One solved, be sure to validate your fix via the Coverage report.
- Look at indexation trends. Most sites are continually adding valuable content, so “Valid” pages (aka those indexed by Google) should steadily increase. Understand the cause of any dramatic changes.
- Select “Valid” and look in details for the type “Indexed, not submitted in sitemap”. These are pages where you and Google disagree on their value. For example, you may not have submitted your privacy policy URL, but Google has indexed the page. In such cases, there’s no actions to be taken. What you need to be looking out for are indexed URLs which stem from [poor pagination handling, poor parameter handling](#), duplicate content or pages being accidentally left out of sitemaps.

Afterwards, limit the report to the SEO relevant URLs you have included in your sitemap by changing the drop down to “All submitted pages”. Then check the details of all “Excluded” pages.

Reasons for exclusion of sitemap URLs can be put into four action groups:

- 1. Quick wins:** For duplicate content, canonicals, robots directives, 40X HTTP status codes, redirects or legalities exclusions put in place the appropriate fix.
- 2. Investigate page:** For both “Submitted URL dropped” and “Crawl anomaly” exclusions investigate further by using the Fetch as Google tool.
- 3. Improve page:** For “Crawled - currently not indexed” pages, review the page (or page type as generally it will be many URLs of a similar breed) content and internal links. Chances are, it’s suffering from thin content, unoriginal content or is orphaned.
- 4. Improve domain:** For “Discovered - currently not indexed” pages, Google notes the typical reason for exclusion as they “tried to crawl the URL but the site was overloaded”. Don’t be fooled. It’s more likely that Google decided “it’s not worth the effort” to crawl due to poor internal linking or low content quality seen from the domain. If you see a larger number of these exclusions, review the SEO value of the page (or page types) you have submitted via sitemaps, focus on optimising crawl budget as well as review your information architecture, including parameters, from both an link and content perspective.

Whatever your plan of action, be sure to note down benchmark KPIs.

The most useful metric to assess the impact of sitemap optimisation efforts is the “All submitted pages” indexation rate - calculated by taking the percentage of valid pages out of total discovered URLs.

Work to get this above 80 percent.

Why not to 100 percent? Because if you have focussed all your energy on ensuring every SEO relevant URL you currently have is indexed, you likely missed opportunities to expand your content coverage.

Note: If you are a larger website who has chosen to break their site down into multiple sitemap indexes, you will be able to filter by those indexes.

This will not only allow you to:

- 1.** See the overview chart on a more granular level.
- 2.** See a larger number of relevant examples when investigating a type of exclusion.
- 3.** Tackle indexation rate optimisation section by section.

Key Takeaway

In addition to identifying warnings and errors, you can use the Index Coverage report as an XML sitemap sleuthing tool to isolate indexation problems.

XML Sitemap Best Practice Checklist

Do invest time to:

- ✓ Compress sitemap files using gzip
- ✓ Use a sitemap index file
- ✓ Use image, video and Google news sitemaps only if indexation drives your KPIs
- ✓ Dynamically generate XML sitemaps
- ✓ Ensure URLs are included only in a single sitemap
- ✓ Reference sitemap index URL(s) in robots.txt
- ✓ Submit sitemap index to both Google Search Console and Bing Webmaster Tools
- ✓ Include only SEO relevant pages in XML sitemaps
- ✓ Fix all errors and warnings
- ✓ Analyse trends and types of valid pages
- ✓ Calculate submitted pages indexation rate
- ✓ Address causes of exclusion for submitted pages

Now, go check your own sitemap and make sure you're doing it right.

Chapter 4

Best Practices for Setting Up Meta Robots Tags & Robots. txt

SEJ
EBOOK

Written By
Sergey Grybniak
Founder, Oppority



First-rate website optimization is fundamental to success in search, but forgetting about the technical part of SEO can be a serious mistake.

Experienced digital marketers and SEO professionals understand the importance of proper search engine indexing. For that reason, they do their best to help Google crawl and index their sites properly, investing time and resources in on-page and off-page optimization.

Content, links, tags, meta descriptions, image optimization, and website structure are essential for SEO, but if you have never heard about robots.txt, meta robots tags, XML sitemaps, microformats, and X-Robot tags, you could be in trouble.

But do not panic.

In this chapter, I will explain how to use and set up robots.txt and meta robots tags. I will provide several practical examples as well.

Let's start!

What Is Robots.txt?

Robots.txt is a text file used to instruct search engine bots (also known as crawlers, robots, or spiders) how to crawl and index website pages.

Ideally, a robots.txt file is placed in the top-level directory of your website so that robots can access its instructions right away.

Why Is Robots.txt Important?

Correct robots.txt operation ensures that search engine bots are routed to required pages, disallowing content duplicates that lead to a fall in position. For that reason, you should make sure your site has a thoughtfully created robot.txt file.

If a robots.txt file is set up incorrectly, it can cause multiple indexing mistakes. So, every time you start a new SEO campaign, check your robots.txt file with [Google's robots texting tool](#).

Do not forget: If everything is correctly set up, a robots.txt file will speed up the indexing process.

Robots.txt on the Web

Yet, do not forget that any robots.txt file is publicly available on the web. To access a robots.txt file, simply type: www.website-example.com/robots.txt.

This availability means that:

- You cannot secure or hide any data within it.
- Bad robots and malicious crawlers can take advantage of a robots.txt file, using it as a detailed map to navigate your most valuable web pages.

Also, keep in mind that robots.txt commands are actually directives. This means that search bots can crawl and index your site, even if you instruct them not to.

The good news is that most search engines (like Google, Bing, Yahoo, and Yandex) honor robots.txt directives.

Robots.txt files definitely have drawbacks. Nonetheless, I strongly recommend you make them an integral part of every SEO campaign.

Google recognizes and honors robots.txt directives and, in most cases, having Google under your belt is more than enough.

Robots.txt Basics

The robots.txt file should:

- Contain the usual text in the UTF-8 encoding, which consists of records (lines), divided by symbols.
- Be situated at the root of the website host to which it applies.

- Be unique.
- Contain not more than 1,024 rules.
- Be under 500KB.

Google bots find all the content available for indexing if:

- There is no robots.txt file.
- A robots.txt file isn't shown in the text format.
- They do not receive the 200 OK response.

Note:

- You can, but are not allowed to, mention the byte order mark (BOM) at the beginning of the robots.txt file, as it will be ignored by bots. The standard recommends the use of a newline before each User-agent directive.
- If your encoding contains symbols beyond the UTF-8, bots may analyze the file incorrectly. They will execute the valid entry only, ignoring the rest of your content without notifying you about the mistake.

Robots.txt Structure

Robots.txt File consists of:

- One or several User-agent directives, meant for robots of various search engines.
- Disallow and Allow directives that allow or restrict indexing. Sitemap directives.

Disallow directives forbid indexing, Allow directives allow indexing.

Each record consists of the directory field (allow, disallow, host or user-agent), two-spot and a value. Empty spaces are not required, but recommended for better readability. You can place comments anywhere in the file and mark them with the # symbol.

“#” is the symbol meant for comment descriptions.

Google bots do not count everything mentioned between the # symbol and the next newline.

- The general format is: <field>:<value><#comment (optional)>.
- Empty spaces at the beginning and the end will be ignored.
- Letter case for <field> element does not matter.
- Letter case might be important for the <value> element, depending on the <field> element.

What to Hide with Robots.txt

Obviously, you do not want to show search engines your private technical page, customers' personal data, and duplicate content.

Robots.txt files can be used to exclude certain directories, categories, and pages from search. To that end, use the “disallow” directive.

Here are some pages you should hide using a robots.txt file:

- Pages with duplicate content
- Pagination pages
- On-site search pages
- Dynamic product and service pages
- Account pages
- Admin pages
- Shopping cart
- Chats
- Thank-you pages

Here is an example of how I instruct Googlebot to avoid crawling and indexing all pages related to user accounts, cart, and multiple dynamic pages that are generated when users look for products in the search bar or sort them by price, and so on.

```
User-Agent: Googlebot
Disallow: /account*
Disallow: /basket*
Disallow: /search*
Disallow: /*&price=
Disallow: /*&sort=
Disallow: /*&page=
Disallow: /*&limit=
Disallow: /*&tag=
Disallow: /*?price=
Disallow: /*?sort=
Disallow: /*?page=
Disallow: /*?limit=
Disallow: /*?tag=
```

How to Use Robots.txt

Robots.txt files are pretty flexible and can be used in many ways.

Their main benefit, however, is that they enable SEO experts to “allow” or “disallow” multiple pages at once without having to access the code of page by page.

For example, you can block all search crawlers from content, like this:

```
User-agent: *  
Disallow: /
```

Or hide your site's directory structure and specific categories, like this:

```
User-agent: *  
Disallow: /no-index/
```

It's also useful for excluding multiple pages from search.

Just parse URLs you want to hide from search crawlers. Then, add the "disallow" command in your robots.txt, list the URLs and, voila! – the pages are no longer visible to Google.

```
Disallow: /help/shipment_payment/order-shipment  
Disallow: /help/shipment_payment/when-shipment  
Disallow: /help/shipment_payment/shipment-in-my-city  
Disallow: /help/shipment_payment/shipment-price  
Disallow: /help/shipment_payment/payment-method  
Disallow: /help/shipment_payment/check-product  
Disallow: /help/faq/different-price  
Disallow: /help/faq/credit-bank  
Disallow: /help/faq/service-collaborate  
Disallow: /help/faq/non-cash-payment  
Disallow: /help/faq/shipping-cities  
Disallow: /help/faq/oversized-cargo  
Disallow: /help/faq/buy-in-office  
Disallow: /help/faq/damaged-product  
Disallow: /help/faq/money-back
```

More important, though, is that a robots.txt file allows you to prioritize certain pages, categories, and even bits of CSS and JS code. Have a look at the example below:

```
User-Agent: *
# Directories
Allow: /wp-content/uploads/
Allow: /wp-content/themes/ /css/
Allow: /wp-content/themes/ /js/
Allow: /wp-content/themes/ /images/
Allow: /wp-content/plugins/contact-form-7/includes/js/
Allow: /wp-content/themes/ /style.css*
Disallow: /wp-login.php
Disallow: /wp-register.php
Disallow: /xmlrpc.php
Disallow: /template.html
Disallow: /wp-admin
Disallow: /wp-includes
Disallow: /wp-content
Disallow: /archive
Disallow: */trackback/
Disallow: */feed/
Disallow: */comments/
Disallow: /?feed=
Disallow: /?s=
Disallow: /openid/
Allow: /blog
```

Here, we have disallowed WordPress pages and specific categories, but wp-content files, JS plugins, CSS styles, and blog are allowed. This approach guarantees that spiders crawl and index useful code and categories, firsthand.

One more important thing: A robots.txt file is one of the possible locations for your sitemap.xml file. It should be placed after User-agent, Disallow, Allow, and Host commands. Like this:

```
User-agent: *
Disallow: /wp-admin/
Disallow: /googlesearch/
Disallow: /wp-trackback
Disallow: /wp-feed
Disallow: /plugins/really-simple-captcha/tmp*
Disallow: /wp-comments
Disallow: /wp-login.php
Disallow: /wp-register.php
Disallow: */trackback
Disallow: */feed
Disallow: /cgi-bin
Disallow: *?s=
#Disallow: /wp-content/plugins
#Disallow: /wp-content/themes
#Disallow: /wp-includes/

User-Agent: Googlebot-Mobile
Allow: /wp-content/
Allow: /wp-content/
Allow: /wp-includes/

Host: ██████████
Sitemap: https://██████████/'sitemap_index.xml
```



Note: You can also add your robots.txt file manually to Google Search Console and, in case you target Bing, Bing Webmaster Tools.

Even though robots.txt structure and settings are pretty straightforward, a properly set up file can either make or break your SEO campaign.

Be careful with settings: You can easily “disallow” your entire site by mistake and then wait for traffic and customers to no avail.

Typical Robots.txt Mistakes

1. The File Name Contains Upper Case

The only possible file name is robots.txt, nor Robots.txt or ROBOTS.TXT.

2. Using Robot.Txt Instead of Robots.txt

Once again, the file must be called robots.txt.

3. Incorrectly Formatted Instructions

For example: Disallow: Googlebot

The only correct option is:

User-agent: Googlebot

Disallow: /

4. Mentioning Several Catalogs in Single 'Disallow' Instructions

Do not place all the catalogs you want to hide in one 'disallow' line, like this:

Disallow: /css/ /cgi-bin/ /images/

The only correct option is:

Disallow: /css/

Disallow: /cgi-bin/

Disallow: /images/

5. Empty Line in 'User-Agent'

Wrong option:

User-agent:

Disallow:

The only correct option is:

User-agent: *

Disallow:

6. Using Upper Case in the File

This is wrong and is treated as a bad style:

USER-AGENT: GOOGLEBOT

DISALLOW:

7. Mirror Websites & URL in the Host Directive

To state which website is the main one and which is the mirror (replica), specialists use 301 redirect for Google and 'host' directive for Yandex.

Although the links to <http://www.site.com>, <http://site.com>, <https://www.site.com>, and <https://site.com> seem identical for humans, search engines treat them as four different websites.

Be careful when mentioning 'host' directives, so that search engines understand you correctly:

✘ Wrong

User-agent: Googlebot

Disallow: /cgi-bin

Host: <http://www.site.com/>

✔ Correct

User-agent: Googlebot

Disallow: /cgi-bin

Host: www.site.com

If your site has https, the correct option is

User-agent: Googlebot

Disallow: /cgi-bin

Host: [https:// www.site.com](https://www.site.com)

8. Listing All the Files Within the Directory

✘ Wrong

```
User-agent: *
Disallow: /AL/Alabama.html
Disallow: /AL/AR.html
Disallow: /Az/AZ.html
Disallow: /Az/bali.html
Disallow: /Az/bed-breakfast.html
```

✔ Correct

```
Just hide the entire directory:
User-agent: *
Disallow: /AL/
Disallow: /Az/
```

9. Absence of Disallow Instructions

The disallow instructions are required so that search engines bots understand your intents.

✘ Wrong

```
User-agent: *
Disallow: /AL/Alabama.html
Disallow: /AL/AR.html
Disallow: /Az/AZ.html
Disallow: /Az/bali.html
Disallow: /Az/bed-breakfast.html
```

✔ Correct

```
Just hide the entire directory:
User-agent: *
Disallow: /AL/
Disallow: /Az/
```

10. Redirect 404

Even if you are not going to create and fill out robots.txt. file for your website, search engines may still try to reach the file. Consider creating at least an empty robots.txt. to avoid disappointing search engines with 404 Not Found pages.

11. Using Additional Directives in the * Section

If you have additional directives, such as 'host' for example, you should create separate sections.

✘ Wrong

```
User-agent: *  
Disallow: /css/  
Host: www.example.com
```

✔ Correct

```
User-agent: *  
Disallow: /css/  
  
User-agent: Googlebot  
Disallow: /css/  
Host: www.example.com
```

12. Incorrect HTTP Header

Some bots can refuse to index the file if there is a mistake in the HTTP header.

✘ Wrong

```
Content-Type: text/html
```

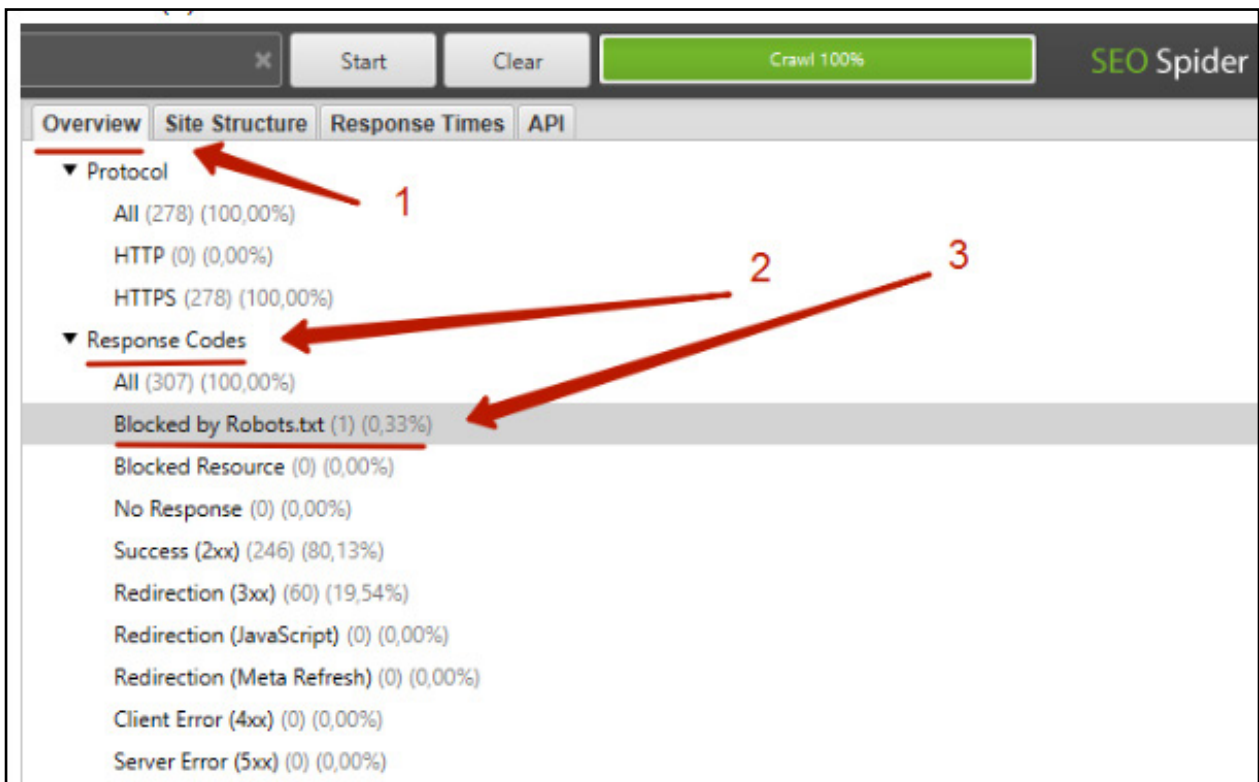
✔ Correct

```
Content Type: text/plain
```

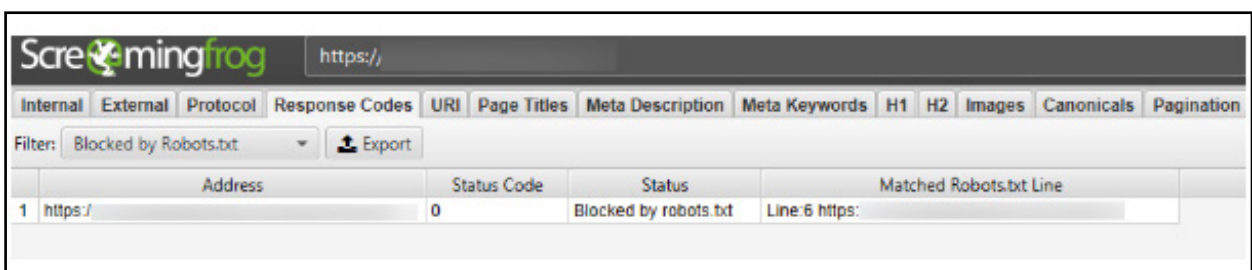

Checking Pages Blocked with Robots.txt

Let's use Screaming Frog to check the web pages that are blocked with our robots.txt file.

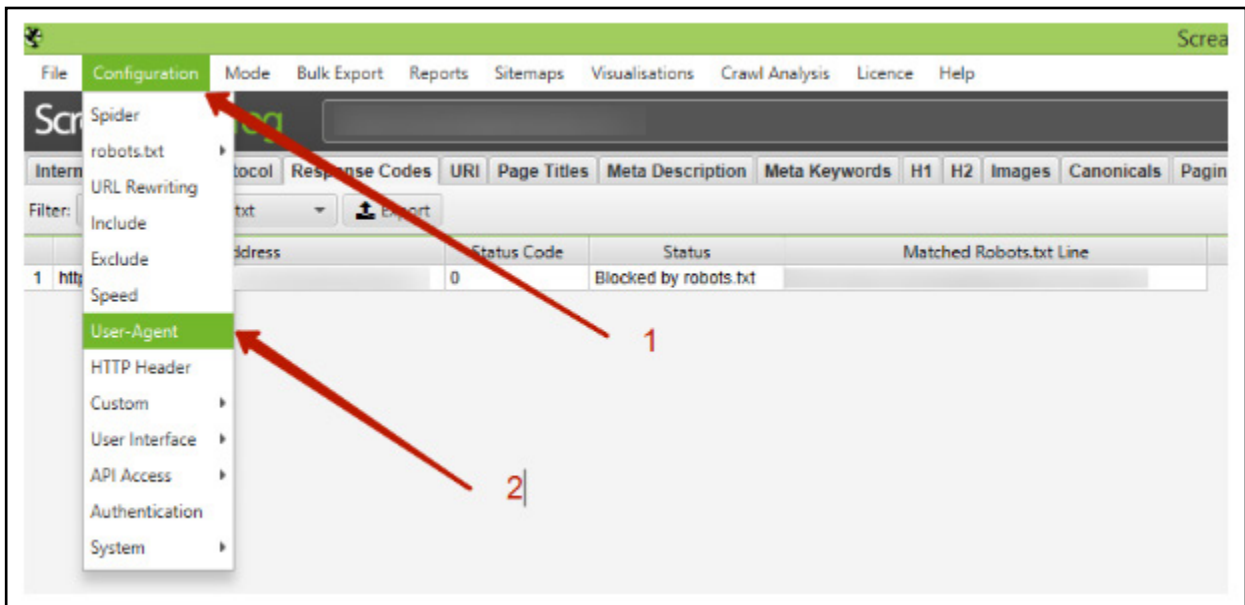
1. Go to the right panel and choose 'Overview' (1), 'Response Codes' (2), 'Blocked by Robots.txt' (3).



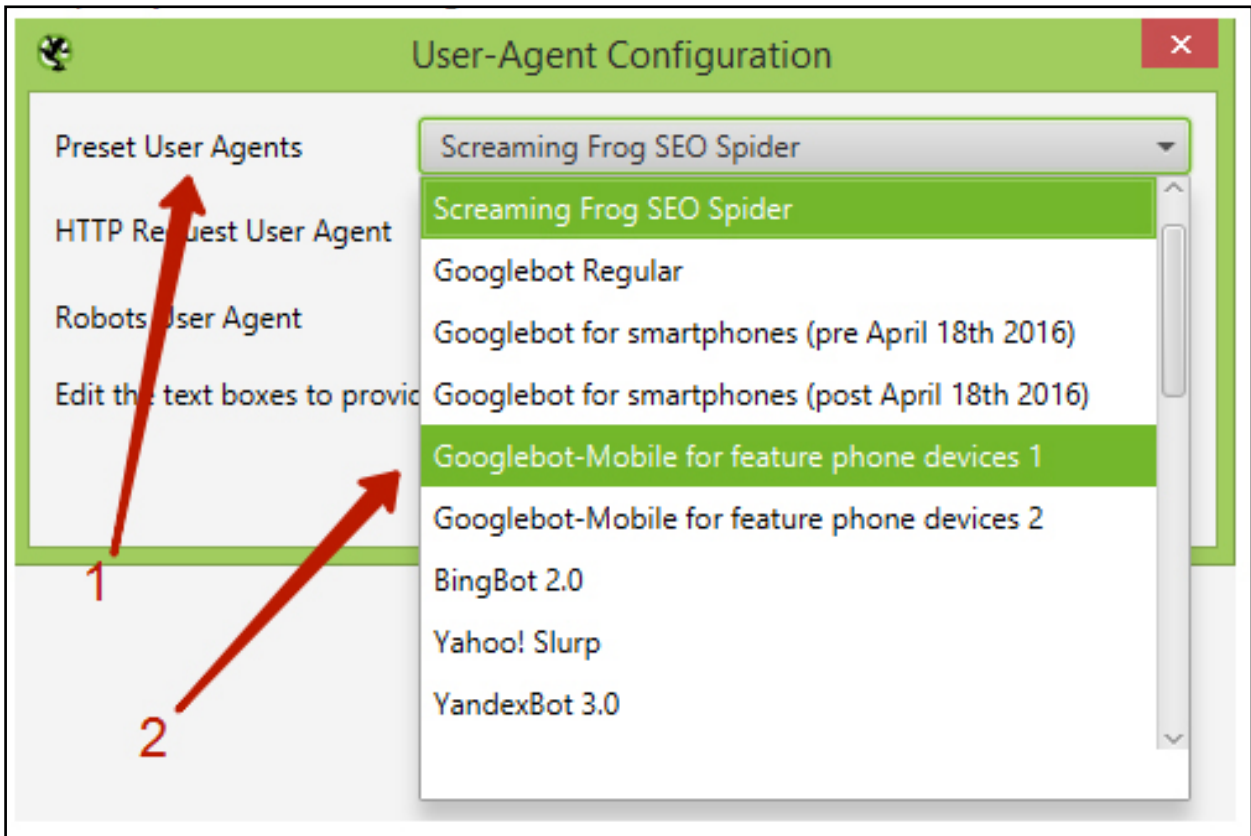
The main panel shows us all the pages that are blocked.



2. Check to ensure that no pages with essential content are occasionally hidden from search engines.
3. Choose 'User Agent' to test robots.txt for various search engines.



4. Specify which search engine bots the tool should imitate.



5. You may test various robots.txt sections by repeating the entire process and pressing 'Start.'

What Are Meta Robots Tags?

Meta robots tags (REP tags) are elements of an indexer directive that tell search engine spiders how to crawl and index specific pages on your website.

They enable SEO professionals to target individual pages and instruct crawlers on what to follow and what not to follow.

Meta Robots Tags Basics

You may hide pages from indexing in several ways, including meta robots tags implementation. Here you can use the following directives:

all – No limitations for indexing and content demonstration. This directive is being used by default and has no impact on the search engines' work, unless otherwise specified.

noindex – Do not show this page and the 'Saved Copy' link in the SERPs.

nofollow – Do not allow following the on-page links.

none – The same as noindex, and nofollow meta tags.

noarchive – Do not show the 'Saved Copy' link in the SERPs.

nosnippet – Do not show the extended description version of this page in the SERPs.

notranslate – Do not offer this page's translation in the SERPs.

noimageindex – Do not index the on-page images.

unavailable_after: [RFC-850 date/time] – Do not show this page in the SERPs after specified date/time. Use RFC 850 format.

How to Use Meta Robots Tags

Meta robots tags are pretty simple to use.

It does not take much time to set up meta robots tags. In four simple steps, you can take your website indexation process up a level:

- 1.** Access the code of a page by pressing CTRL + U.
- 2.** Copy and paste the <head> part of a page's code into a separate document.
- 3.** Provide step-by-step guidelines to developers using this document. Focus on how, where, and which meta robots tags to inject into the code.
- 4.** Check to make sure the developer has implemented the tags correctly. To do so, I recommend using The Screaming Frog SEO Spider.

The screenshot below demonstrates how meta robots tags may look (check out the first line of code):

```
<meta name="robots" content="noindex, follow" />  
<link rel="canonical" href="https://www.example.com/" />  
<link rel="next" href="https://www.example.com/next/" />  
<meta property="og:locale" content="en_US" />  
<meta property="og:type" content="object" />  
<meta property="og:title" content="Acupuncture" />  
<meta property="og:description" content="Read more about Acupuncture in our blog" />  
<meta property="og:url" content="https://www.example.com/tag/acupuncture/" />  
<meta property="og:site_name" content="Example" />
```

Meta robots tags are recognized by major search engines: Google, Bing, Yahoo, and Yandex. You do not have to tweak the code for each individual search engine or browser (unless they honor specific tags).

Main Meta Robots Tags Parameters

As I mentioned above, there are four main REP tag parameters: follow, index, nofollow, and noindex. Here is how you can use them:

- index, follow: allow search bots to index a page and follow its links
- noindex, nofollow: prevent search bots from indexing a page and following its links
- index, nofollow: allow search engines to index a page but hide its links from search spiders
- noindex, follow: exclude a page from search but allow following its links (link juice helps increase SERPs)

REP tag parameters vary. Here are some of the rarely used ones:

- none
- noarchive
- nosnippet
- unavailable_after
- noimageindex
- nocache
- noodp
- notranslate

Meta robots tags are essential if you need to optimize specific pages. Just access the code and instruct developers on what to do.

If your site runs on an advanced CMS (OpenCart, PrestaShop) or uses specific plugins (like WP Yoast), you can also inject meta tags and their parameters directly into page templates. This allows you to cover multiple pages at once without having to ask developers for help.

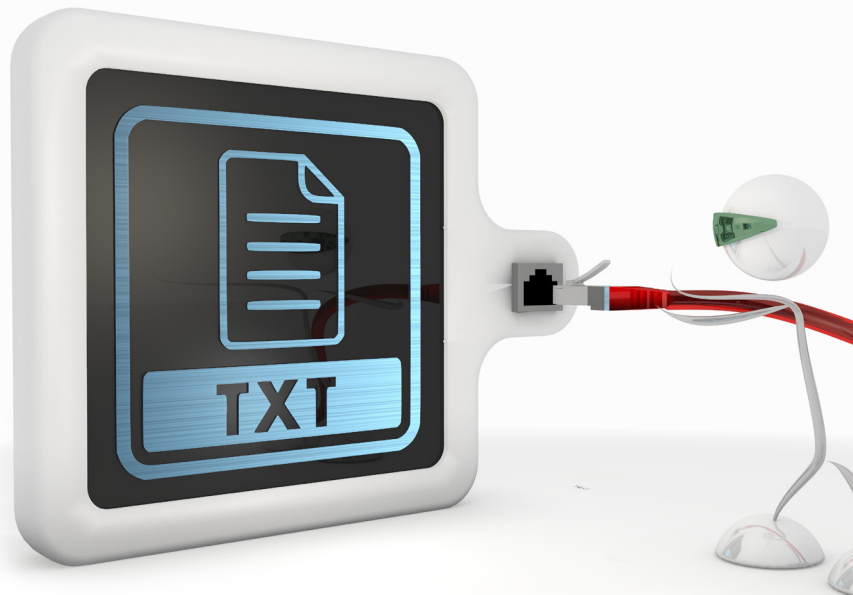
Robots.txt & Meta Robots Tags Non-Compliance

Incoherence between directives in robots.txt and on-page meta tags is a common mistake.

For example, the robots.txt file hides the page from indexing, but the meta robots tags do the opposite.

In such cases, Google will pay attention to what is prohibited by the robots.txt file. Most likely, bots will ignore the directives that encourage indexing of the content.

Pay attention to the fact that robots.txt is a recommendation by Google, but not a demand.



Therefore, you still have a chance to see your page indexed, as long as there are external links that lead to them.

If robots.txt does not hide the page, but the directives do – Google bots will accomplish the most restricting task and will not index the on-page content.

The conclusion is simple: eliminate non-compliance between meta robots tags and robots.txt to clearly show Google which pages should be indexed, and which should not.

Another noteworthy example is incoherence between on-page meta tags.

Yandex search bots opt for positive value when they notice conflicts between the meta tags on a page:

```
<meta name= "robots" content="all"/>  
<meta name="robots" content="noindex, follow"/>  
<!--Bots will choose the 'all' value and index all the links and texts.-->
```

By contrast, Google bots opt for the strongest directive, indexing only links and ignoring the content.

The Sitemap.xml Role

The sitemap.xml, robots.txt and meta robots tags instructions complement one another when set up correctly.

The major rules are:

- Sitemap.xml, robots.txt and meta robots tags should not be conflicting.
- All the pages that are blocked in robots.txt and meta robots tags must be excluded from sitemap.xml as well.
- All the pages that are opened for indexing must be included in the sitemap.xml as well.
- The sitemap.xml, robots.txt and meta robots tags instructions complement one another when set up correctly.

The major rules are:

- Sitemap.xml, robots.txt and meta robots tags should not be conflicting.
- All the pages that are blocked in robots.txt and meta robots tags must be excluded from sitemap.xml as well.
- All the pages that are opened for indexing must be included in the sitemap.xml as well.

However, a few exceptions exist:

Starting the second pagination page, you should add 'noindex, follow' to the meta robots tags, leaving those pages open for indexing in robots.txt.

Consider adding all the pagination pages to the sitemap.xml, so all the links can be re-indexed.

To Sum It Up

Knowing how to set up and use a robots.txt file and meta robots tags is extremely important. A single mistake can spell death to your entire campaign.

I personally know several digital marketers who have spent months doing SEO, only to realize that their websites were closed to indexing in robots.txt. Others abused the "nofollow" tag so much that they lost backlinks in droves.

Dealing with robots.txt files and REP tags is pretty technical, which can potentially lead to many mistakes. Fortunately, there are several basic rules that will help you implement them successfully.

Robots.txt

1. Place your robots.txt file in the top-level directory of your website code to simplify crawling and indexing.

- 2.** Structure your robots.txt properly, like this: User-agent - Disallow - Allow - Host - Sitemap. This way, search engine spiders access categories and web pages in the appropriate order.
- 3.** Make sure that every URL you want to “Allow:” or “Disallow:” is placed on an individual line. If several URLs appear on one single line, crawlers will have a problem accessing them.
- 4.** Use lowercase to name your robots.txt. Having “robots.txt” is always better than “Robots.TXT.” Also, file names are case sensitive.
- 5.** Do not separate query parameters with spacing. For instance, a line query like this “/cars/ /audi/” would cause mistakes in the robots.txt file.
- 6.** Do not use any special characters except * and \$. Other characters are not recognized.
- 7.** Create separate robots.txt files for different subdomains. For example, “hubspot.com” and “blog.hubspot.com” have individual files with directory- and page-specific directives.
- 8.** Use # to leave comments in your robots.txt file. Crawlers do not honor lines with the # character.
- 9.** Do not rely on robots.txt for security purposes. Use passwords and other security mechanisms to protect your site from hacking, scraping, and data fraud.

Meta Robots Tags

Be case sensitive. Google and other search engines may recognize attributes, values, and parameters in both uppercase and lowercase, and you can switch between the two if you want. I strongly recommend that you stick to one option to improve code readability.

Avoid multiple `<meta>` tags. By doing this, you will avoid conflicts in code. Use multiple values in your `<meta>` tag, like this: `<meta name="robots" content="noindex, nofollow">`.

Do not use conflicting meta tags to avoid indexing mistakes. For example, if you have several code lines with meta tags like this `<meta name="robots" content="follow">` and this `<meta name="robots" content="nofollow">`, only “nofollow” will be taken into account. This is because robots put restrictive values first.

Note: You can easily implement both robots.txt and meta robots tags on your site. However, be careful to avoid confusion between the two.

The basic rule here is, restrictive values take precedent. So, if you “allow” indexing of a specific page in a robots.txt file but accidentally “noindex” it in the `<meta>`, spiders will not index the page.

Also, remember: If you want to give instructions specifically to Google, use the `<meta>` “googlebot” instead of “robots”, like this: `<meta name="googlebot" content="nofollow">`. It is similar to “robots” but avoids all the other search crawlers.

Chapter 5

Your Indexed Pages Are Going Down – 5 Possible Reasons Why

SEJ
EBOOK

Written By
Benj Arriola
SEO Director, Myers Media Group



Getting your webpages indexed by Google (and other search engines) is essential. Pages that aren't indexed can't rank.

How do you see how many pages you have indexed? You can:

- Use the **site: operator**.
- Check the **status of your XML Sitemap Submissions** in Google Search Console.
- Check your overall **indexation status**.

Each will give different numbers, but why they are different is another story.

For now, let's just talk about analyzing a decrease in the number of indexed pages reported by Google.

If your pages aren't being indexed, this could be a sign that Google may not like your page or may not be able to easily crawl it.

Therefore, if your indexed page count begins to decrease, this could be because either:

- You've been slapped with a Google penalty.
- Google thinks your pages are irrelevant.
- Google can't crawl your pages.

Here are a few tips on how to diagnose and fix the issue of decreasing numbers of indexed pages.

1. Are the Pages Loading Properly?

Make sure they have the proper 200 HTTP Header Status.

Did the server experience frequent or long downtime? Did the domain recently expire and was renewed late?

Action Item

You can use a free HTTP Header Status checking tool to determine whether the proper status is there. For massive sites, typical crawling tools like Xenu, DeepCrawl, Screaming Frog, or Botify can test these.

The correct header status is 200. Sometimes some 3xx (except the 301), 4xx, or 5xx errors may appear – none of these are good news for the URLs you want to be indexed.

2. Did Your URLs Change Recently?

Sometimes a change in CMS, backend programming, or server setting that results in a change in domain, subdomain, or folder may consequently change the URLs of a site.

Search engines may remember the old URLs but, if they don't redirect properly, a lot of pages can become de-indexed.

Action Item

Hopefully a copy of the old site can still be visited in some way or form to take note of all old URLs so you can map out the 301 redirects to the corresponding URLs.

3. Did You Fix Duplicate Content Issues?

Fixing duplicate content often involves implementing canonical tags, 301 redirects, noindex meta tags, or disallows in robots.txt. All of which can result in a decrease in indexed URLs.

This is one example where the decrease in indexed pages might be a good thing.

Action Item

Since this is good for your site, the only thing you need to do is to double check that this is definitely the cause of the decrease of indexed pages and not anything else.

4. Are Your Pages Timing Out?

Some servers have bandwidth restrictions because of the associated cost that comes with a higher bandwidth; these servers may need to be upgraded. Sometimes, the issue is hardware related and can be resolved by upgrading your hardware processing or memory limitation.

Some sites block IP addresses when visitors access too many pages at a certain rate. This setting is a strict way to avoid any **DDOS** hacking attempts but it can also have a negative impact on your site.

Typically, this is monitored at a page's second setting and if the threshold is too low, normal search engine bot crawling may hit the threshold and the bots cannot crawl the site properly.

Action Item

If this is a server bandwidth limitation, then it might be an appropriate time to upgrade services.

If it is a server processing/memory issue, aside from upgrading the hardware, double check if you have any kind of server caching technology in place, this will give less stress on the server.

If an anti-DDOS software is in place, either relax the settings or whitelist Googlebot to not be blocked anytime. Beware though, there are some fake Googlebots out there; be sure to **detect googlebot** properly. **Detecting Bingbot** has a similar procedure.

5. Do Search Engine Bots See Your Site Differently?

Sometimes what search engine spiders see is different than what we see.

Some developers build sites in a preferred way without knowing the SEO implications.

Occasionally, a preferred out-of-the-box CMS will be used without checking if it is search engine friendly.

Sometimes, it might have been done on purpose by an SEO who attempted to do content cloaking, trying to game the search engines.

Other times, the website has been compromised by hackers, who cause a different page to be shown to Google to promote their hidden links or cloak the 301 redirections to their own site.

The worse situation would be pages that are infected with some type of malware that Google automatically deindexes the page immediately once detected.



Action Item

Using Google Search Console's [fetch and render feature](#) is the best way to see if Googlebot is seeing the same content as you are.

You may also try to translate the page in [Google Translate](#) even if you have no intention to translate the language or check [Google's Cached page](#), but there are also ways around these to still cloak content behind them.

Index Pages Are Not Used as Typical KPIs

Key Performance Indicators (KPIs), which help measure the success of an SEO campaign, often revolve around organic search traffic and ranking. KPIs tend to focus on the goals of a business, which are tied to revenue.

An increase in indexed pages may increase the possible number of keywords you can rank for that can result in higher profits. However, the point of looking at indexed pages is mainly just to see whether search engines are able to crawl and index your pages properly.

Remember, your pages can't rank when search engines can't see, crawl, or index them.

A Decrease in Indexed Pages Isn't Always Bad

Most of the time, a decrease in indexed pages could mean a bad thing, but a fix to duplicate content, thin content, or low-quality content might also result in a decreased number of indexed pages, which is a good thing.

Learn how to evaluate your site by looking at these five possible reasons why your indexed pages are going down.

Chapter 6

An SEO Guide to HTTP Status Codes

SEJ
EBOOK

Written By
Brian Harnish
SEO Director, Site Objective



One of the most important assessments in any SEO audit is determining what hypertext transfer protocol status codes (or HTTP Status Codes) exist on a website.

These codes can become complex, often turning into a hard puzzle that must be solved before other tasks can be completed.

For instance, if you put up a page that all of a sudden disappears with a 404 not found status code, you would check server logs for errors and assess what exactly happened to that page.

If you are working on an audit, other status codes can be a mystery, and further digging may be required.

These codes are segmented into different types:

- 1xx status codes are informational codes.
- 2xx codes are success codes.
- 3xx redirection codes are redirects.
- 4xx are any codes that fail to load on the client side, or client error codes.
- 5xx are any codes that fail to load due to a server error.

1xx Informational Status Codes

These codes are informational in nature and usually have no real-world impact for SEO.

100 – Continue

Definition: In general, this protocol designates that the initial serving of a request was received and not yet otherwise rejected by the server.

SEO Implications: None

Real World SEO Application: None

101 - Switching Protocols

Definition: The originating server of the site understands, is willing and able to fulfill the request of the client via the Upgrade header field. This is especially true for when the application protocol on the same connection is being used.

SEO Implications: None

Real World SEO Application: None

102 – Processing

Definition: This is a response code between the server and the client that is used to inform the client side that the request to the server was accepted, although the server has not yet completed the request.

SEO Implications: None

Real World SEO Application: None

2xx Client Success Status Codes

This status code tells you that a request to the server was successful. This is mostly only visible server-side. In the real world, visitors will never see this status code.

SEO Implications: A page is loading perfectly fine, and no action should be taken unless there are other considerations (such as during the execution of a content audit, for example).

Real-World SEO Application: If a page has a status code of 200 OK, you don't really need to do much to it if this is the only thing you are looking at. There are other applications involved if you are doing a content audit, for example.

However, that is beyond the scope of this article, and you should already know whether or not you will need a content audit based on initial examination of your site.

How to find all 2xx success codes on a website via Screaming Frog:

There are two ways in Screaming Frog that you can find 2xx HTTP success codes: through the GUI, and through the bulk export option.

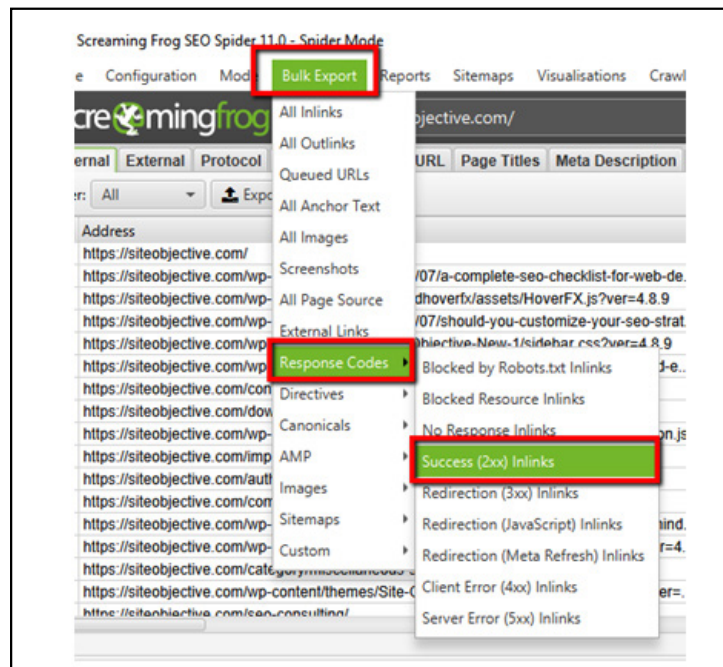
Method 1 – Through the GUI

1. Crawl your site using the settings that you are comfortable with.
2. All of your site URLs will show up at the end of the crawl.
3. Look for the Status Code column. Here, you will see all 200 OK, 2xx based URLs.

URL	Content	Status Code	Status	Indexability	Index
content/uploads/2017/07/a-complete-seo-checklist-for-web-de...	image/jpeg	200	OK	Indexable	
content/mu-plugins/1dhoverfx/assets/HoverFX.js?ver=4.8.9	application/javascript	200	OK	Indexable	
content/uploads/2017/07/should-you-customize-your-seo-strat...	image/jpeg	200	OK	Indexable	
content/themes/Site-Objective-New-1/sidebar.css?ver=4.8.9	text/css	200	OK	Indexable	
content/uploads/2017/07/how-to-audit-canonicalization-and-e...	image/jpeg	200	OK	Indexable	
content/uploads/2017/07/how-to-business-the-roi-of-seo/	text/html, charset=UTF-8	200	OK	Indexable	
content/mu-plugins/1donelevelmenu/assets/menu_selection.js	application/javascript	200	OK	Indexable	
content/uploads/2017/07/rover-your-website-conversions-with-split-testing/	text/html, charset=UTF-8	200	OK	Indexable	
content/uploads/2017/07/rover/admin/	text/html, charset=UTF-8	200	OK	Indexable	
content/uploads/2017/07/rover/competitor-analysis/	text/html, charset=UTF-8	200	OK	Indexable	
content/uploads/2017/07/rover/seo-facebook-live-brian-harnish-mind...	image/jpeg	200	OK	Indexable	
content/themes/Site-Objective-New-1/drop_menu2.css?ver=4.8.9	text/css	200	OK	Indexable	
content/uploads/2017/07/rover/miscellaneous-seo/	text/html, charset=UTF-8	200	OK	Indexable	
content/themes/Site-Objective-New-1/menu_menu1.css?ver=4.8.9	text/css	200	OK	Indexable	
content/uploads/2017/07/rover/consulting/	text/html, charset=UTF-8	200	OK	Indexable	

Method 2 – The Bulk Export Option

1. Crawl your site using the settings that you are comfortable with.
2. Click on Bulk Export
3. Click on Response Codes
4. Click on 2xx Success Inlinks



201 – Created

This status code will tell you that the server request has been satisfied and that the end result was that one or multiple resources were created.

202 – Accepted

This status means that the server request was accepted to be processed, but the processing has not been finished yet.

203 – Non-Authoritative Information

A transforming proxy modified a successful payload from the origin server's 200 OK response.

204 – No Content

After fulfilling the request successfully, no more content can be sent in the response payload body.

205 – Reset Content

This is similar to the 204 response code, except the response requires the client sending the request reset the document view.

206 – Partial Content

Transfers of one or more components of the selected page that corresponds to satisfiable ranges that were found in the range header field of the request. The server, essentially, successfully fulfilled the range request for said target resource.

207 – Multi-Status

In situations where multiple status codes may be the right thing, this multi-status response displays information regarding more than one resource in these situations.

3xx Redirection Status Codes

Mostly, 3xx Redirection codes denote redirects. From temporary to permanent. 3xx redirects are an important part of preserving SEO value.

That's not their only use, however. They can explain to Google whether or not a page redirect is permanent, temporary, or otherwise.

In addition, the redirect can be used to denote pages of content that are no longer needed.

301 – Moved Permanently

These are permanent redirects. For any site migrations, or other situations where you have to transfer SEO value from one URL to another on a permanent basis, these are the status codes for the job.

How Can 301 Redirects Impact SEO?

Google has said several things about the use of 301 redirects and their impact. John Mueller has [cautioned](#) about their use.

“So for example, when it comes to links, we will say well, it's this link between this canonical URL and that canonical URL- and that's how we treat that individual URL.

In that sense it's not a matter of link equity loss across redirect chains, but more a matter of almost usability and crawlability. Like, how can you make it so that Google can find the final destination as quickly as possible? How can you make it so that users don't have to jump through all of these different redirect chains. Because, especially on mobile, chain redirects, they cause things to be really slow.

If we have to do a DNS lookup between individual redirects, kind of moving between hosts, then on mobile that really slows things down. So that's kind of what I would focus on there.

Not so much like is there any PageRank being dropped here. But really, how can I make it so that it's really clear to Google and to users which URLs that I want to have indexed. And by doing that you're automatically reducing the number of chain redirects."

It is also important to note here that not all 301 redirects will pass 100 percent link equity. From Roger Montti's reporting:

"A redirect from one page to an entirely different page will result in no PageRank being passed and will be considered a soft 404."

John Mueller also mentioned previously:

"301-redirecting for 404s makes sense if you have 1:1 replacement URLs, otherwise we'll probably see it as soft-404s and treat like a 404."

The matching of the topic of the page in this instance is what's important. "the 301 redirect will pass 100 percent PageRank only if the redirect was a redirect to a new page that closely matched the topic of the old page."

302 – Found

Also known as temporary redirects, rather than permanent redirects. They are a cousin of the 301 redirects with one important difference: they are only temporary.

You may find 302s instead of 301s on sites where these redirects have been improperly implemented.

Usually, they are done by developers who don't know any better.

The other 301 redirection status codes that you may come across include:

300 – Multiple Choices

This redirect involves multiple documents with more than one version, each having its own identification. Information about these documents is being provided in a way that allows the user to select the version that they want.

303 – See Other

A URL, usually defined in the location header field, redirects the user agent to another resource. The intention behind this redirect is to provide an indirect response to said initial request.

304 – Not Modified

The true condition, which evaluated false, would normally have resulted in a 200 OK response should it have evaluated to true. Applies to GET or HEAD requests mostly.

305 – Use Proxy

This is now deprecated, and has no SEO impact.

307 – Temporary Redirect

This is a temporary redirection status code that explains that the targeted page is temporarily residing on a different URL. It lets the user agent know that it must NOT make any changes to the method of request if an auto redirect is done to that URL.

308 – Permanent Redirect

Mostly the same as a 301 permanent redirect.

4xx Client Error Status Codes

4xx client error status codes are those status codes that tell us that something is not loading – at all – and why.

While the error message is a subtle difference between each code, the end result is the same. These errors are worth fixing and should be one of the first things assessed as part of any website audit.

- Error 400 Bad Request
- 403 Forbidden
- 404 Not Found

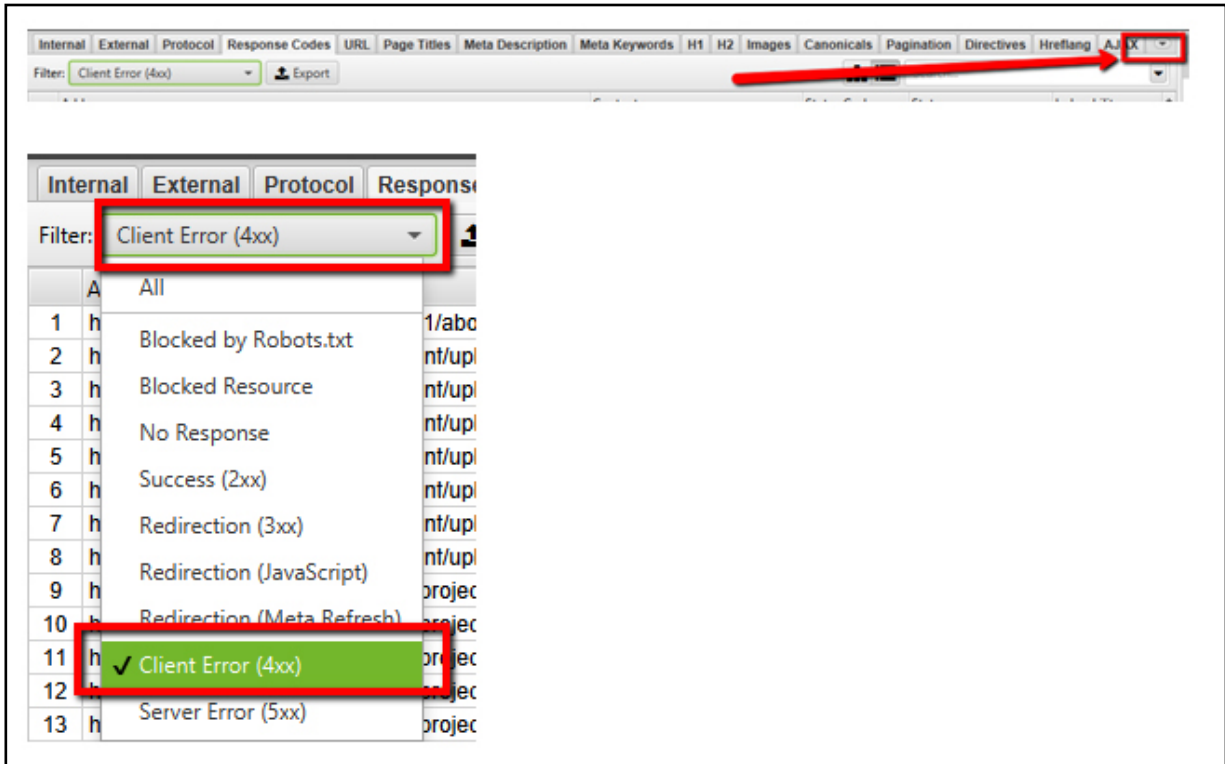
These statuses are the most common requests an SEO will encounter – the 400, 403 and 404 errors. These errors simply mean that the resource is unavailable and unable to load.

Whether it's due to a temporary server outage, or other reason, it doesn't really matter. What matters is the end result of the bad request – your pages are not being served by the server and is

There are two ways to find 4xx errors that are plaguing a site in Screaming Frog – through the GUI, and through bulk export.

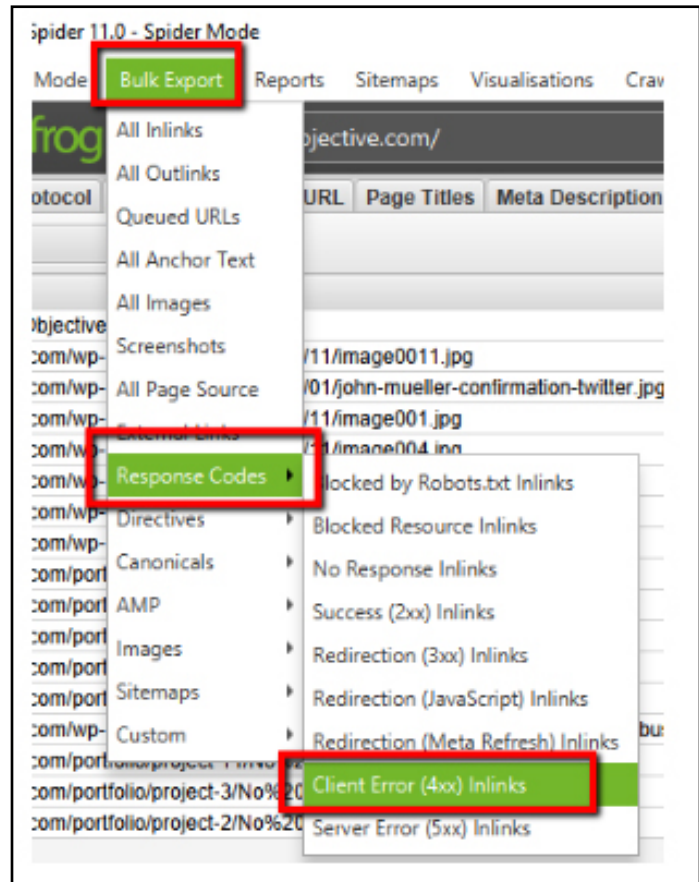
Screaming Frog GUI Method:

- 1.** Crawl your site using the settings that you are comfortable with.
- 2.** Click on the down arrow to the right.
- 3.** Click on response codes.
- 4.** Filter by Client Error (4xx).



Screaming Frog Bulk Export Method:

1. Crawl your site with the settings you are familiar with.
2. Click on Bulk Export.
3. Click on Response Codes.
4. Click on Client error (4xx) Inlinks.



These are other 4xx errors that you may come across, including:

- 401** - Unauthorized
- 402** - Payment Required
- 405** - Method Not Allowed
- 406** - Not Acceptable
- 407** - Proxy Authentication Required
- 408** - Request Timeout
- 409** - Conflict
- 410** - Gone
- 411** - Length Required
- 412** - Precondition Failed
- 413** - Payload Too Large
- 414** - Request-URI Too Long
- 415** - Unsupported Media Type
- 416** - Requested Range Not Satisfiable
- 417** - Expectation Failed
- 418** - I'm a teapot
- 421** - Misdirected Request
- 422** - Unprocessable Entity
- 423** - Locked
- 424** - Failed Dependency
- 426** - Upgrade Required
- 428** - Precondition Required
- 429** - Too Many Requests
- 431** - Request Header Fields Too Large
- 444** - Connection Closed Without Response
- 451** - Unavailable For Legal Reasons
- 499** - Client Closed Request

5xx Server Error Status Codes

All of these errors imply that there is something wrong at the server level that is preventing the full processing of the request.

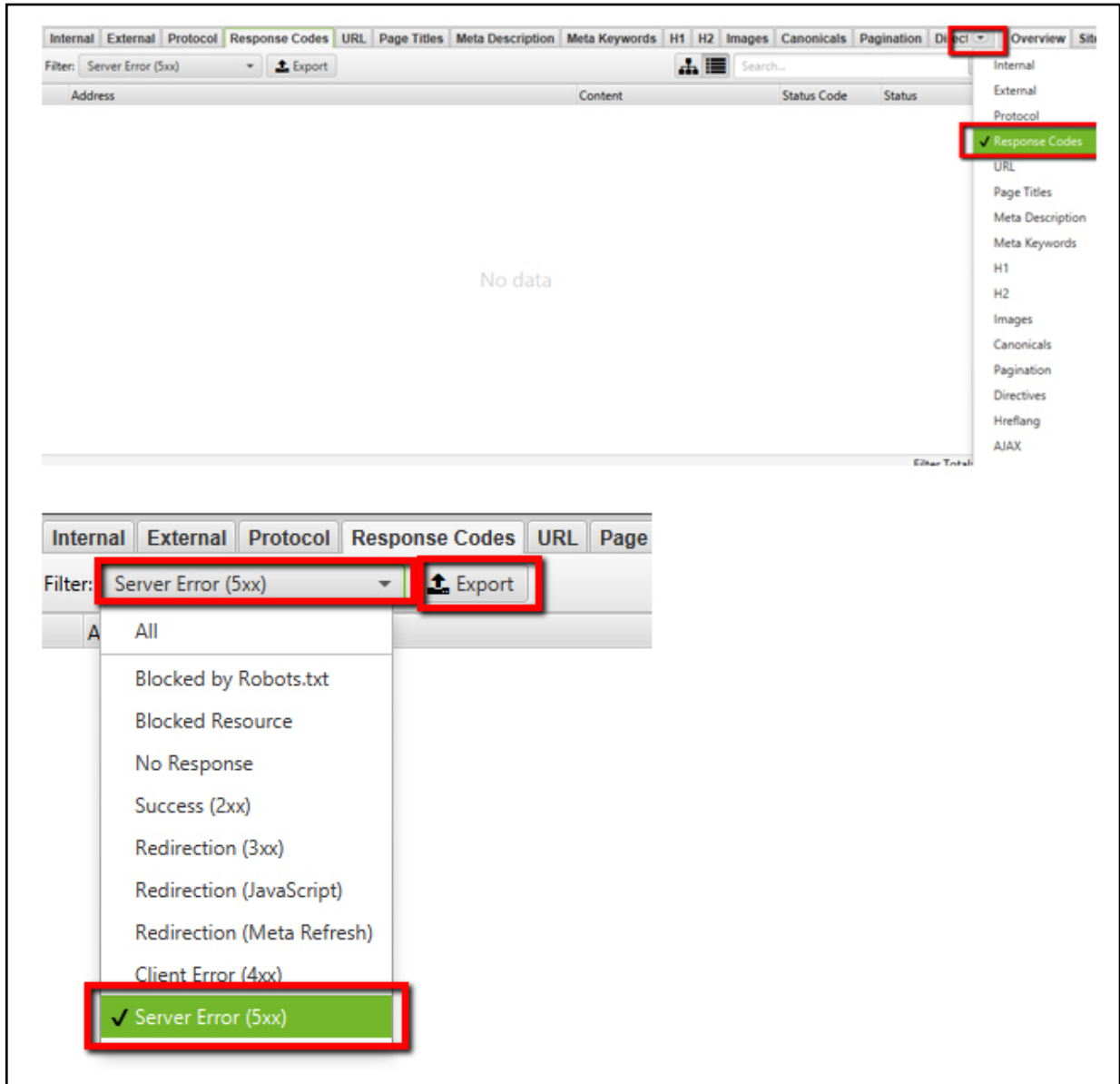
The end result will always (in most cases that serve us as SEOs) be the fact that the page does not load and will not be available to the client side user agent that is viewing it.

This can be a big problem for SEOs.

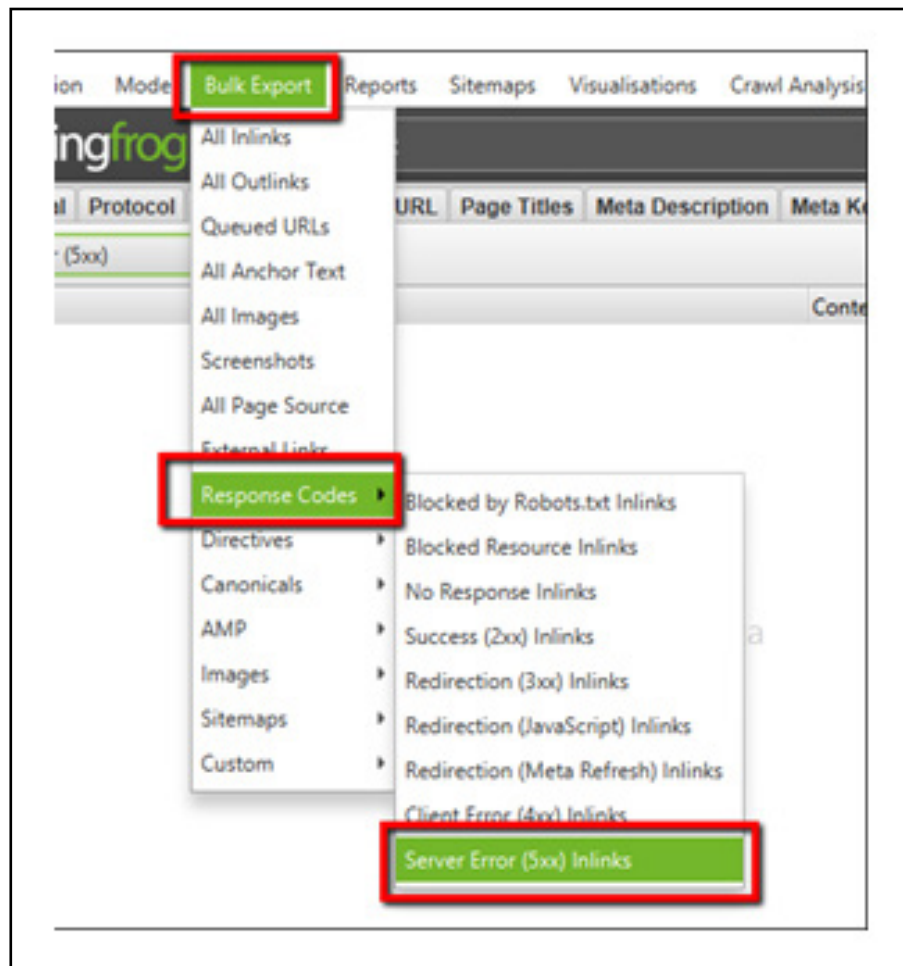
Again, using Screaming Frog, there are two methods you can use to get to the root of the problems being caused by 5xx errors on a website. A GUI method, and a Bulk Export method.

Screaming Frog GUI Method for Unearthing 5xx Errors

1. Crawl your site using the settings that you are comfortable with.
2. Click on the dropdown arrow on the far right.
3. Click on "response codes".
4. Click on Filter > Server Error (5xx)
5. Select Server Error (5xx).
6. Click on Export



Screaming Frog Bulk Export Method for Unearthing 5xx Errors



- 1.** Crawl your site using the settings you are comfortable with.
- 2.** Click on Bulk Export.
- 3.** Click on Response Codes.
- 4.** Click on Server Error (5xx) Inlinks.

This will give you all of the 5xx errors that are presenting on your site.

There are other 5xx http status codes that you may come across, including the following:

- 500** - Internal Server Error
- 501** - Not Implemented
- 502** - Bad Gateway
- 503** - Service Unavailable
- 504** - Gateway Timeout
- 505** - HTTP Version Not Supported
- 506** - Variant Also Negotiates
- 507** - Insufficient Storage
- 508** - Loop Detected
- 510** - Not Extended
- 511** - Network Authentication Required
- 599** - Network Connect Timeout Error

Making Sure That HTTP Status Codes Are Corrected On Your Site Is a Good First Step

When it comes to making a site that is 100 percent crawlable, one of the first priorities is making sure that all content pages that you want the search engines to know about are 100 percent crawlable. This means making sure that all pages are 200 OK.

Once that is complete, you will be able to move forward with more SEO audit improvements as you assess priorities and additional areas that need to be improved.

“A website’s work is never done” should be an SEO’s mantra. There is always something that can be improved on a website that will result in improved search engine rankings.

If someone says that their site is perfect, and that they need no further changes, then I have a \$1 million dollar bridge to sell you in Florida.



Chapter 7

404 vs. Soft 404 Errors: What's the Difference & How to Fix Both

SEJ
EBOOK

Written By
Benj Arriola
SEO Director, Myers Media Group



Every page that loads in a web browser has a response code included in the HTTP headers, which may or may not be visible on the web page itself.

There are many **different response codes** a server gives to communicate the loading-status of the page; one of the most well-known codes is the **404-response code**.

Generally, any code within 400 to 499 indicates that the page didn't load. The 404-response code is the only one that carries a specific meaning – that the page is actually gone and probably isn't coming back anytime soon.

What's a Soft 404 Error?

A soft 404 error isn't an official response code sent to a web browser. It's just a label Google adds to a page within their index. As Google crawls pages, it allocates resources carefully ensuring that no time is wasted by crawling missing pages which do not need to be indexed.

However, there are some servers that are poorly configured and their missing page loads a 200 code when it should display a 404-response code. If the invisible HTTP header displays a 200 code even if the web page clearly states that the page isn't found, the page might be indexed, which is a waste of resources for Google.

To combat this issue, Google notes the characteristics of 404 pages and attempts to discern whether the 404 page really is a 404 page. In other words, Google learned that if it looks like a 404, smells like a 404, and acts like a 404, then it's probably a genuine 404 page.

Potentially Misidentified as Soft 404

There are also cases wherein the page isn't actually missing, but certain characteristics have triggered Google to categorize it as a missing page.

Some of these characteristics include a small amount or lack of content on the page and having too many similar pages on the site.

These characteristics are also similar to the factors that the [Panda algorithm](#) tackles. The Panda update considers thin and duplicate content as negative ranking factors.

Therefore, fixing these issues will help avoid both soft 404s and Panda issues.

404 errors have two main causes:

- An error in the link, directing users to a page that doesn't exist.
- A link going to a page that used to exist and suddenly disappeared.

Linking Error

If the cause of the 404 is a linking error, you just have to fix the links.

The difficult part of this task is finding all the broken links on a site.

It can be more challenging for large, complex sites that have thousands or millions of pages. In instances like this, crawling tools come in handy. You can try using software such as Xenu, DeepCrawl, Screaming Frog, or Botify.

A Page That No Longer Exists

When a page no longer exists, you have two options:

- Restore the page if it was accidentally removed.
- 301 redirect it to the closest related page if it was removed on purpose.

First, you have to locate all the linking errors on the site. Similar to finding all errors in linking for a large scale website, you can use crawling tools. However, crawling tools may not find **orphaned pages**, which are pages that are not linked from anywhere within the navigational links or from any of the pages.

Orphaned pages can exist if they used to be part of the website, then after a website redesign, the link going to this old page disappeared, but external links from other websites might still be linking to them. To double check if these kinds of pages exist on your site, you can use a variety of tools.

Google Search Console

Search console will report 404 pages as Google's crawler goes through all the pages it can find. This can include links from other sites going to a page that used to exist on your website.

Google Analytics

You won't find a missing page report in Google Analytics by default. However, you can track them in a number of ways.

For one, you can create a custom report and segment out pages that have a page title mentioning Error 404 – Page Not Found.



Another way to find orphaned pages within Google Analytics is to create custom content groupings and to assign all 404 pages to a content group.

Site: Operator Search Command

Searching Google for “site:example.com” will list all pages of example.com that are indexed by Google. You can then individually check if the pages are loading or if they’re giving 404s.

To do this at scale, I like using WebCEO, which has a feature to run the site: operator not only on Google, but also on Bing, Yahoo, Yandex, Naver, Baidu, and Seznam.

Since all the search engines will only give you a subset, running it on multiple search engines can help give a larger list of pages of your site. This list can be exported and run on tools for a mass 404 check. I simply do this by adding all URLs as links within an HTML file and loading it on Xenu to massively check for 404 errors.

Other Backlink Research Tools

Backlink research tools like Majestic, Ahrefs, Moz Open Site Explorer, Sistrix, LinkResearchTools, and CognitiveSEO can also help.

Most of these tools will export a list of backlinks linking to your domain. From there, you can check all the pages that are being linked to and look for 404 errors.

How to Fix Soft 404 Errors

Crawling tools won't detect a soft 404 because it isn't really a 404 error. But you can use crawling tools to detect something else.

Here are a few things to find:

Thin Content: Some crawling tools not only report pages that have thin content, but also show a total word count. From there, you can sort URLs based on your content's number of words. Start with pages that have the least amount of words and evaluate whether the page has thin content.

Duplicate Content: Some crawling tools are sophisticated enough to discern what percentage of the page is template content. If the main content is nearly the same as many other pages, you should look into these pages and determine why duplicate content exists on your site.

Aside from the crawling tools, you can also use Google Search Console and check under crawl errors to find pages that are listed under soft 404s.

Crawling an entire site to find issues that cause soft 404s allows you to locate and correct problems before Google even detects them.

After detecting these soft 404 issues, you will need to correct them.

Most of the time, the solutions appear to be common sense. This can include simple things like expanding pages with thin content or replacing duplicate content with new and unique ones.

Throughout this process, here are a few things to consider:

Consolidate Pages: Sometimes thin content is caused by being too specific with the page topic, which can leave you with little to say. Merging several thin pages into one page can be more appropriate if the topics are related. Not only does this solve thin content issues, but it can fix duplicate content issues as well. For example, an e-commerce site selling shoes that come in different colors and sizes may have a different URL for each size and color combination. This leaves a large number of pages with content that is thin and relatively identical. The more effective approach is to put this all on one page instead and enumerate the options available.

Find Technical Issues That Cause Duplicate Content: Using even the simplest web crawling tool like Xenu (which doesn't look at content but only URLs, response codes, and title tags), you can still find duplicate content issues by looking at URLs. This includes things like www vs non-www URLs, http and https, with index.html and without, with tracking parameters and without, etc. A good summary of these common duplicate content issues found in URLs patterns can be found on [slide 6 of this presentation.](#)

Google Treats 404 Errors & Soft 404 Errors the Same Way

A soft 404 is not real 404 error, but Google will deindex those pages if they aren't fixed quickly. It is best to crawl your site regularly to see if 404 or soft 404 errors occur. Crawling tools should be a major component of your SEO arsenal.



404
PAGE NOT FOUND

Chapter 8

8 Tips to Optimize Crawl Budget for SEO

SEJ
EBOOK

Written By
Aleh Barysevich
Founder and CMO, SEO PowerSuite



When you hear the words “search engine optimization,” what do you think of?

My mind leaps straight to a list of SEO ranking factors, such as proper tags, relevant keywords, a clean sitemap, great design elements, and a steady stream of high-quality content.

However, **a recent article by my colleague**, Yauhen Khutarniuk, made me realize that I should be adding “crawl budget” to my list.

While many SEO experts overlook crawl budget because it’s not very well understood, Khutarniuk brings some compelling evidence to the table – which I’ll come back to later in this chapter – that crawl budget can, and should, be optimized.

This made me wonder: how does crawl budget optimization overlap with SEO, and what can websites do to improve their crawl rate?

First Things First – What Is a Crawl Budget?

Web services and search engines use web crawler bots, aka “spiders,” to crawl web pages, collect information about them, and add them to their index. These spiders also detect links on the pages they visit and attempt to crawl these new pages too.

Examples of bots that you’re probably familiar with include [Googlebot](#), which discovers new pages and adds them to the Google Index, or [Bingbot](#), Microsoft’s equivalent.

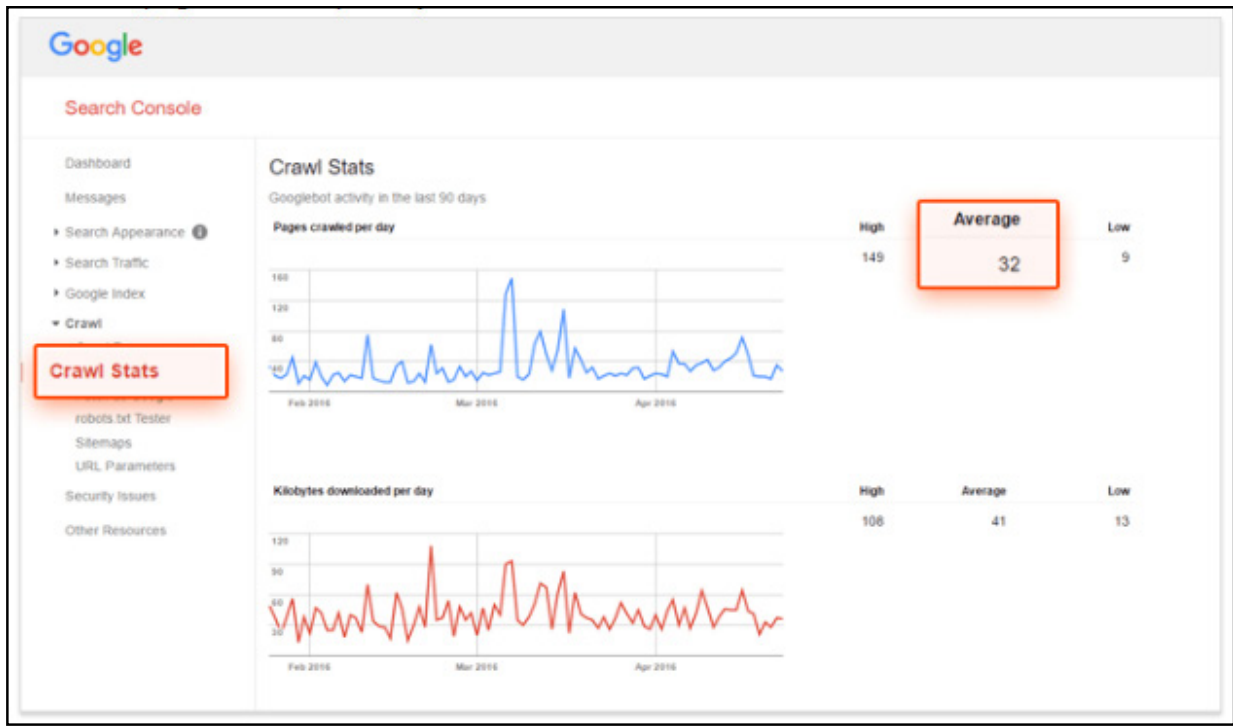
Most SEO tools and other web services also rely on spiders to gather information. For example, my company’s backlink index, SEO PowerSuite Backlink Index, is built using a spider called [BLEXBot](#), which crawls up to 7.1 billion web pages daily gathering backlink data.”

The number of times a search engine spider crawls your website in a given time allotment is what we call your “crawl budget.” So if Googlebot hits your site 32 times per day, we can say that your typical Google crawl budget is approximately 960 per month.

You can use tools such as [Google Search Console](#) and [Bing Webmaster Tools](#) to figure out your



website's approximate crawl budget. Just log in to Crawl > Crawl Stats to see the average number of pages crawled per day.



Is Crawl Budget Optimization the Same as SEO?

Yes – and no. While both types of optimization aim to make your page more visible and may impact your SERPs, SEO places a heavier emphasis on user experience, while spider optimization is entirely about appealing to bots.

So how do you optimize your crawl budget specifically? I've gathered the following nine tips to help you make your website as crawlable as possible.

How to Optimize Your Crawl Budget

1. Ensure Your Pages Are Crawlable

Your page is crawlable if search engine spiders can find and follow links within your website. You'll have to configure your .htaccess and robots.txt so that they don't block your site's critical pages.

You may also want to provide text versions of pages that rely heavily on rich media files, such as Flash and Silverlight.

Of course, the opposite is true if you do want to prevent a page from showing up in search results.

However, it's not enough to simply set your Robots.txt to "Disallow," if you want to stop a page from being indexed. [According to Google](#): "Robots.txt Disallow does not guarantee that a page will not appear in results."

If external information (e.g., incoming links) continue to direct traffic to the page that you've disallowed, Google may decide the page is still relevant.

In this case, you'll need to manually block the page from being indexed by using the noindex robots meta tag or the X-Robots-Tag HTTP header.

noindex meta tag: Place the following meta tag in the <head> section of your page to prevent most web crawlers from indexing your page:

```
noindex" />
```

X-Robots-Tag: Place the following in your HTTP header response to tell crawlers not to index a page:

```
X-Robots-Tag: noindex
```

Note that if you use noindex meta tag or X-Robots-Tag, you should not disallow the page in robots.txt, The page must be crawled before the tag will be seen and obeyed.

2. Use Rich Media Files Cautiously

There was a time when Googlebot couldn't crawl content like JavaScript, Flash, and HTML. Those times are mostly past (though Googlebot still struggles with Silverlight and some other files).

However, even if Google can read most of your rich media files, other search engines may not be able to, which means that you should use these files judiciously, and you probably want to avoid them entirely on the pages you want to be ranked.

You can find a full list of the [files that Google can index here](#).

3. Avoid Redirect Chains

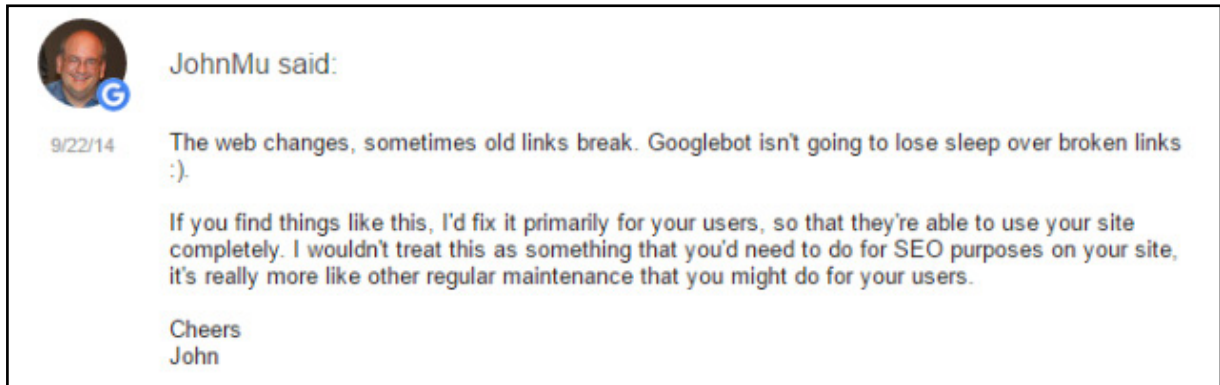
Each URL you redirect to wastes a little of your crawl budget.

When your website has long redirect chains, i.e., a large number of 301 and 302 redirects in a row, spiders such as Googlebot may drop off before they reach your destination page, which means that page won't be indexed.

Best practice with redirects is to have as few as possible on your website, and no more than two in a row.

4. Fix Broken Links

When asked whether or not broken links affect web ranking, Google's [John Mueller once said:](#)



If what Mueller says is true, this is one of the fundamental differences between SEO and Googlebot optimization, because it would mean that broken links do not play a substantial role in rankings, even though they greatly impede Googlebot's ability to index and rank your website.

That said, you should take Mueller's advice with a grain of salt – Google's algorithm has improved substantially over the years, and anything that affects user experience is likely to impact SERPs.

5. Set Parameters on Dynamic URLs

Spiders treat dynamic URLs that lead to the same page as separate pages, which means you may be unnecessarily squandering your crawl budget.

You can manage your URL parameters by going to your Google Search Console and clicking Crawl > Search Parameters.

From here, you can let Googlebot know if your CMS adds parameters to your URLs that doesn't change a page's content.

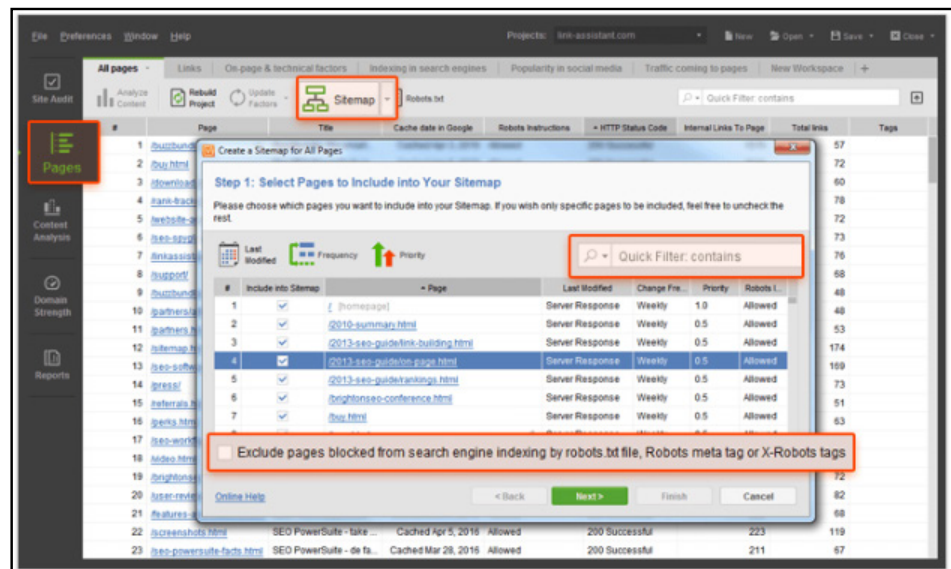
6. Clean Up Your Sitemap

XML sitemaps help both your users and spider bots alike, by making your content better organized and easier to find.

Try to keep your sitemap up-to-date and purge it of any clutter that may harm your site's usability, including 400-level pages, unnecessary redirects, non-canonical pages, and blocked pages.

The easiest way to clean up your sitemap is to use a tool like Website Auditor (disclaimer: my tool). You can use Website Auditor's XML sitemap generator to create a clean sitemap that excludes all pages blocked from indexing.

Plus, by going to Site Audit, you can easily find and fix all 4xx status pages, 301 and 302 redirects, and non-canonical pages.



7. Build External Links

Link building is still a hot topic – and I doubt it's going away anytime soon. As SEJ's [Anna Crowe elegantly put it:](#)

“Cultivating relationships online, discovering new communities, building brand value – these small victories should already be imprints on your link-planning process. While there are distinct elements of link building that are now so 1990s, the human need to connect with others will never change.”

Now, in addition to Crowe's excellent point, we also have evidence from Yauhen Khutarniuk's experiment that external links closely correlate with the number of spider visits your website receives.

In his experiment, he used our tools to measure all of the internal and external links pointing to every page on 11 different sites. He then analyzed crawl stats on each page and compared the results. This is an example of what he found on just one of the sites he analyzed:

While the data set couldn't prove any conclusive connection between internal links and crawl rate, Khutarniuk did find an overall “strong correlation (0,978) between the number of spider visits and the number of external links.”

8. Maintain Internal Link Integrity

While Khutarniuk's experiment proved that internal link building doesn't play a substantial role in crawl rate, that doesn't mean you can disregard it altogether.

A well-maintained site structure makes your content easily discoverable by search bots without wasting your crawl budget. A well-organized internal linking structure may also improve user experience – especially if users can reach any area of your website within three clicks.

Making everything more easily accessible in general means visitors will linger longer, which may improve your SERPs.



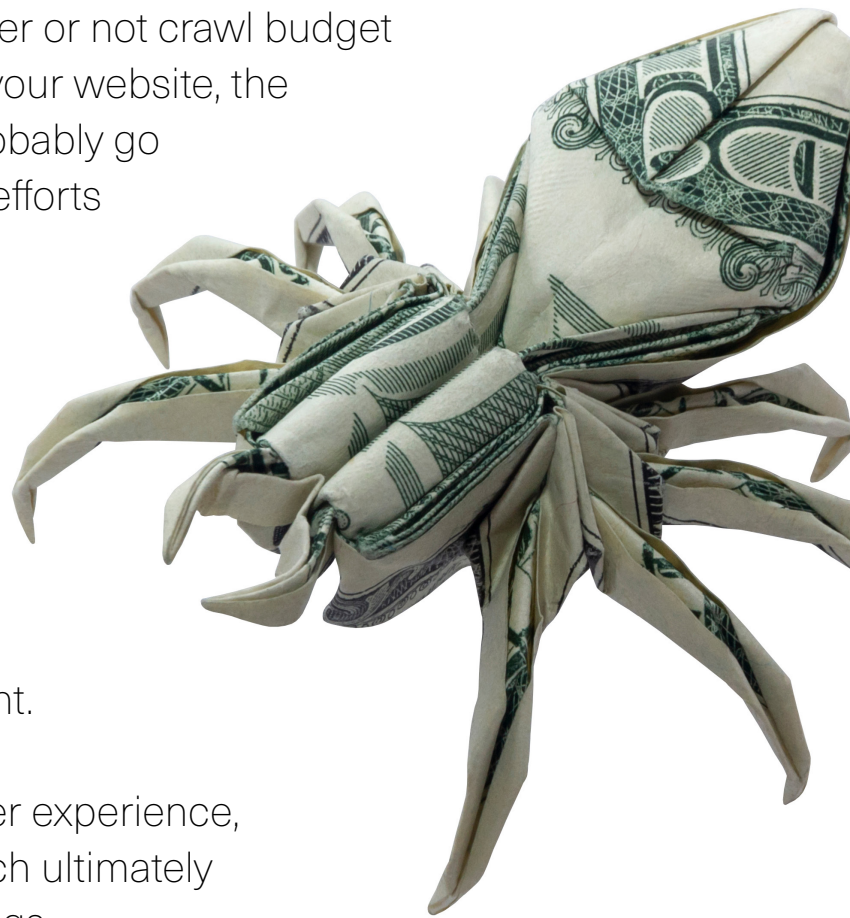
Conclusion: Does Crawl Budget Matter?

By now, you've probably noticed a trend in this article – the best-practice advice that improves your crawlability tends to improve your searchability as well.

So if you're wondering whether or not crawl budget optimization is important for your website, the answer is YES – and it will probably go hand-in-hand with your SEO efforts anyway.

Put simply, when you make it easier for Google to discover and index your website, you'll enjoy more crawls, which means faster updates when you publish new content.

You'll also improve overall user experience, which improves visibility, which ultimately results in better SERPs rankings.



Chapter 9

How to Improve Your Website Navigation: 7 Essential Best Practices

SEJ
EBOOK

Written By
Benj Arriola
SEO Director, Myers Media Group



Website navigation, when done right, is great for your users and your SEO performance.

Good website navigation makes it easy for your visitors to find what they want and for search engines to crawl.

The result: more conversions and greater search visibility.

But how do you actually do it? By using these website navigation best practices.

What is Website Navigation?

Website navigation (a.k.a., internal link architecture) are the links within your website that connect your pages. The primary purpose of website navigation is to help users easily find stuff on your site. Search engines use your website navigation to discover and index new pages. Links help search engines to understand the content and context of the destination page, as well as the relationships between pages.

Users come first. This is the underlying objective of website navigation you must always remember.

Satisfy users first. Make navigation easy. Then, optimize for search engines without hurting the user experience.

If you more basic information on website navigation, you'll find these SEJ posts helpful:

- [Internal Linking Guide to Boost Your SEO](#) by Syed Balkhi
- [Your Essential Guide to Internal Content Linking](#) by Julia McCoy

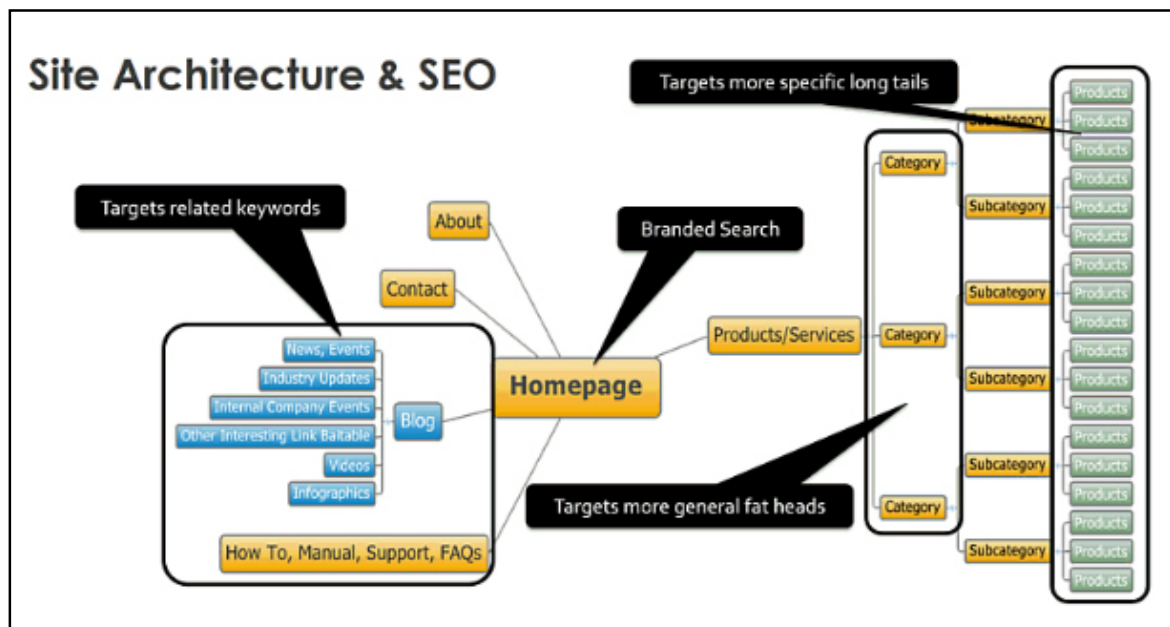
The remainder of this post will maintain a broader focus on website navigation best practices, outlining various internal linking situations that can cause issues for your website visitors and search engines. This topic will be especially relevant and important for anyone working on large websites.

Website Navigation & Content Hierarchies

When searching for a specific page within a book, you can simply read through the table of contents or the index. When you walk around the grocery store, the aisles are labeled with general section categories and more subcategories are listed on the shelves themselves. Both provide an efficient way to navigate through a lot of content.

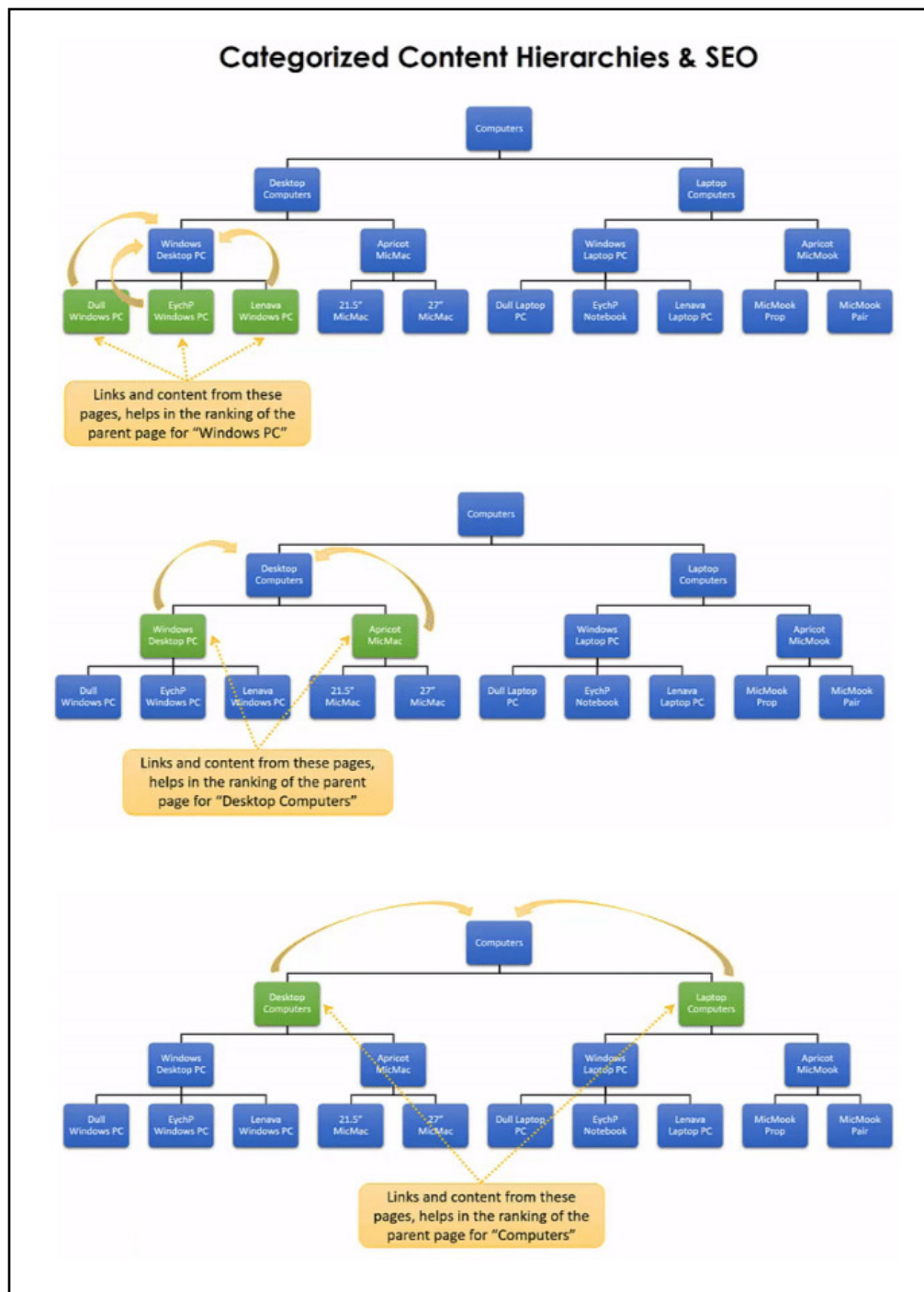
Content hierarchies exist to simplify the process of locating content. When a mass amount of content exists, it can be broken down into a few broad categories.

Within those broad categories, you can create even narrower classifications; this builds differing hierarchical levels that users can easily navigate. Utilizing content hierarchies organizes pages of a website in a way that makes sense to the user and the search engine.



Importance of Content Hierarchies & Website Navigation

The categorization and sub-categorization of content help pages improve in rank for general head terms and for specific long-tail terms.



Problems Caused by Content Hierarchies

Categorization of content and building hierarchies create content silos, like clusters of closely related topics. Google will crawl different pages at different rates, following links from different sites.

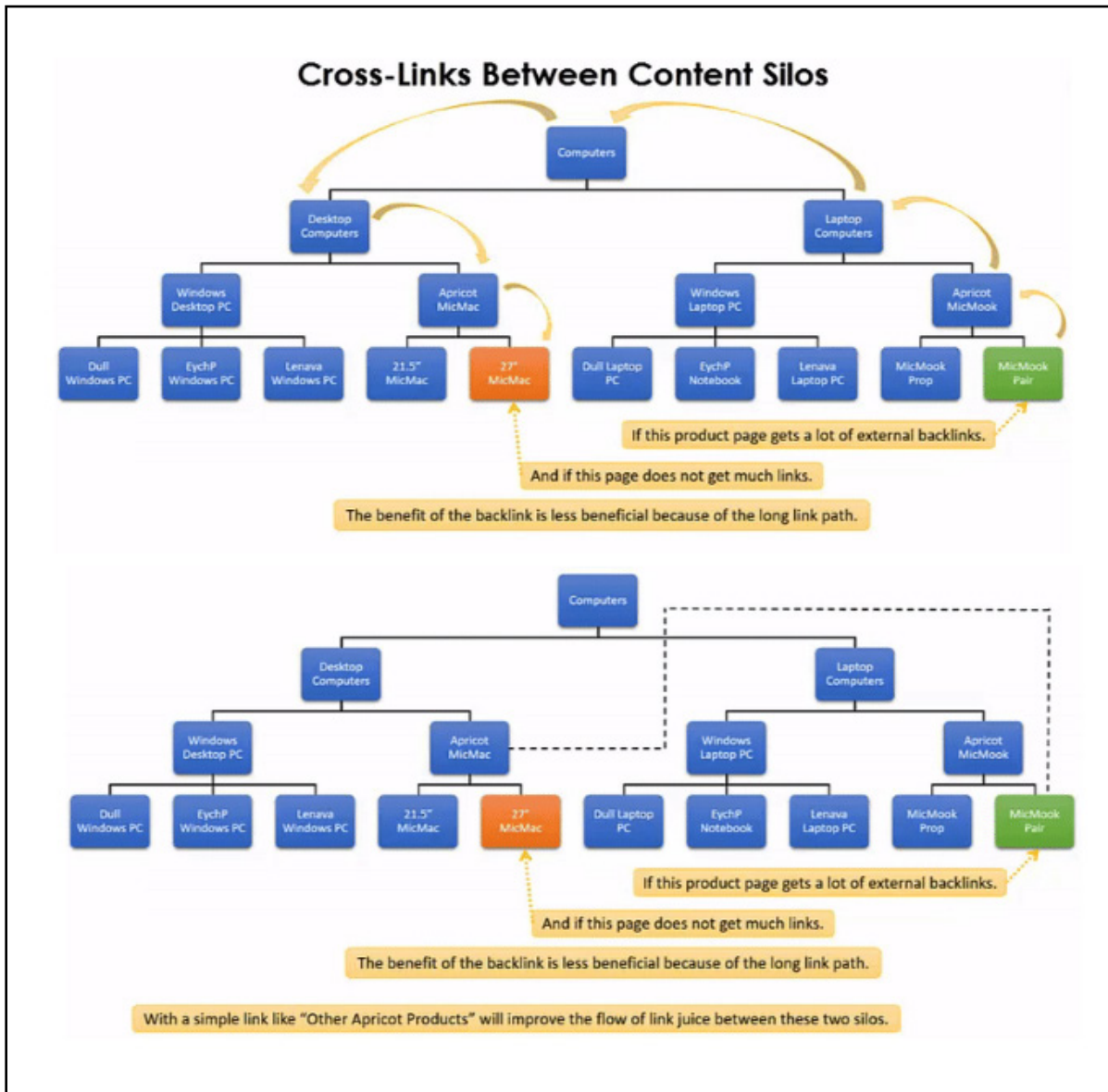
Some content silos are more popular than others. These pages may get more external links and traffic than others and, as a result, earn more prominent positions in organic search.

When content is too siloed and fails to get links and traffic, it might not perform as well – even if your other content silos perform extremely well. The content hierarchies can isolate certain popular page clusters that may be located too deep within the site.

This is where horizontal linking comes into play.

As much as link relevancy helps in ranking, the lack of cross-linking between content silos can be detrimental to your overall rankings. There are always ways to create relationships that horizontally link categories to one another.

The fact that all pages belong to the same website already indicates that these pages are not completely irrelevant to each other.



Action Items: Linking Between Content Categories

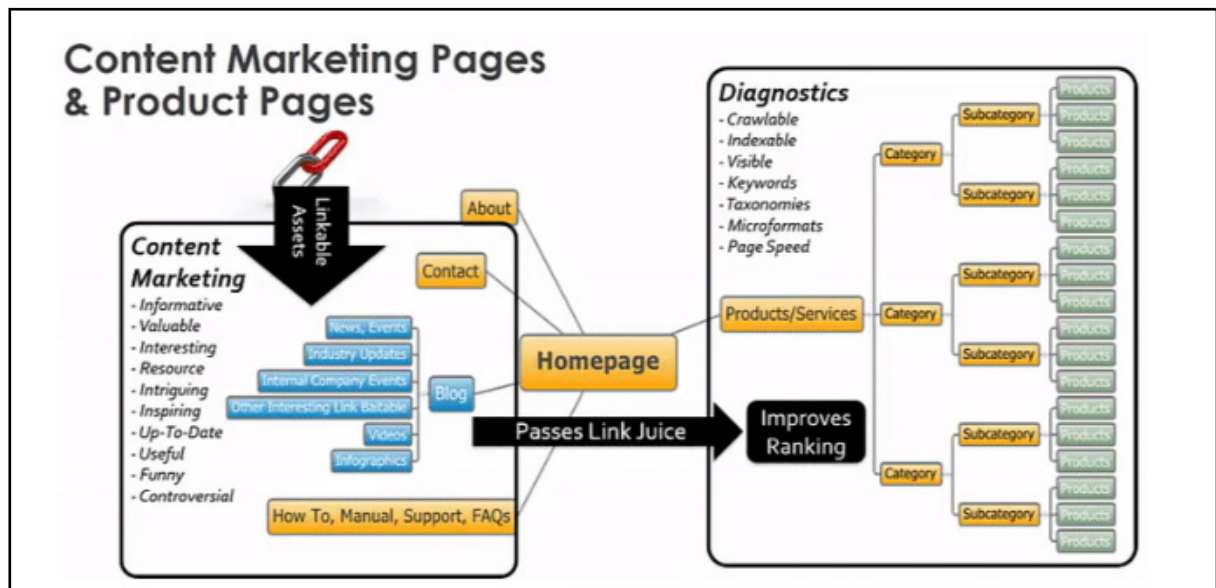
- Categorize content in a way that forms category hierarchies that make sense to the user and interlink these pages properly, going up and down the hierarchy. These are the majority of the links.
- Create cross-linking between pages that are under different categories but still have similarities.

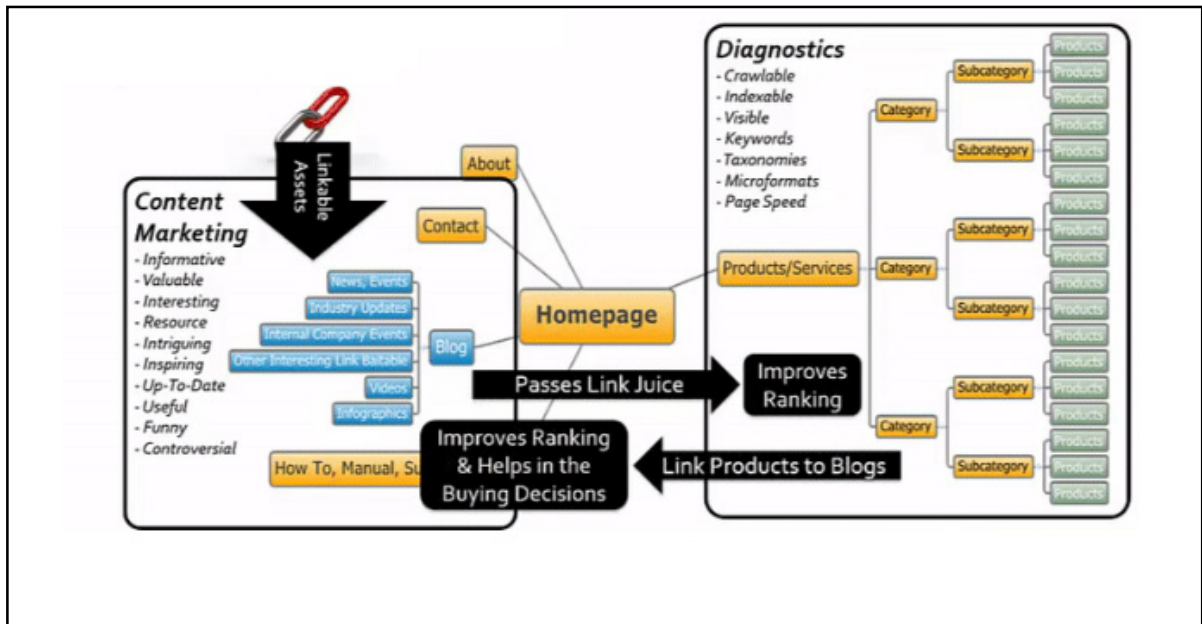
Links Between Product & Content Marketing Pages

Companies selling more than one product or service will do everything mentioned above on categorizing the pages, creating content silos, and interlinking them.

However, many SEO teams and content teams also create assets that are designed to be compelling and shareable. Oftentimes, this comes in the form of a blog, with posts containing links to specific products and services.

Blog posts can be useful because they direct more traffic toward product pages. However, many sites fail to link the product pages back to the blog pages. Using this type of horizontal linking helps inform users about your product or service and increases your SEO performance.





Action Items: Linking Between Product and Content Pages

- Product pages should also link back to related content marketing pages. This may include blog posts, FAQs, and product manuals.

Website Navigation Using JavaScript Effects

Occasionally, links and web pages are written in JavaScript. This is a problem because search engines have difficulty locating internal links that are created in JavaScript.

Although Google has improved in recent years in terms of reading JavaScript, SEO specialists have concluded that results are inconsistent. Other search engines still have no capabilities when it

comes to reading JavaScript. This means your internal linking could be completely lost when search engines crawl your content.

The SEO world is divided over whether using JavaScript is practical. On one hand, some SEO experts avoid JavaScript altogether. On the other hand, web designers and usability experts claim that JavaScript is essential to the user experience. I believe there is a middle ground where JavaScript can be used while avoiding any SEO issues.

Links That Display and Hide Content Already on the Page

JavaScript can be used to display and hide certain content on a page without actually changing the page you are on. When this happens, all of your content is pre-loaded to the page.

In this case, search engines are still able to crawl all of your content, even when some of it is hidden. This is only successful when the amount of content that is hidden remains minor; it can become problematic when the entire page changes but the URL remains the same.

Problems arise because of the fact that when you hide too much content within one URL, it dilutes the content focus of what that page is all about. A completely different topic should have its own page.



Action Items: Links That Display and Hide Content

- For small amounts of content, remove the anchor tag and replace with a JavaScript onclick event handler.
 - Use CSS to control the cursor and change from an arrow to a hand pointer.
- For large amounts of content, including single-page parallax scrolling websites, not all content should be pre-loaded.
 - Only pre-load content directly related to the URL. For all anchor tags, there should be an href value and an onclick setting.
 - This href value leads to a new URL that only pre-loads the content related to this new URL.
 - The onclick function will prevent the new URL from loading but will allow content from the destination URL to load.
 - Use the pushState function to update the URL even if that page did not load.

A more in-depth presentation of how this can be specifically implemented on websites is explained well in this presentation done at seoClarity in 2016. It specifically talks about [AngularJS, a popular JavaScript framework, and its SEO issues and solutions.](#) However, the lessons here are also applicable to almost any JavaScript framework.

Using Tracking Parameters in the URL

Usability experts and conversion optimization specialists track user behavior in different ways.

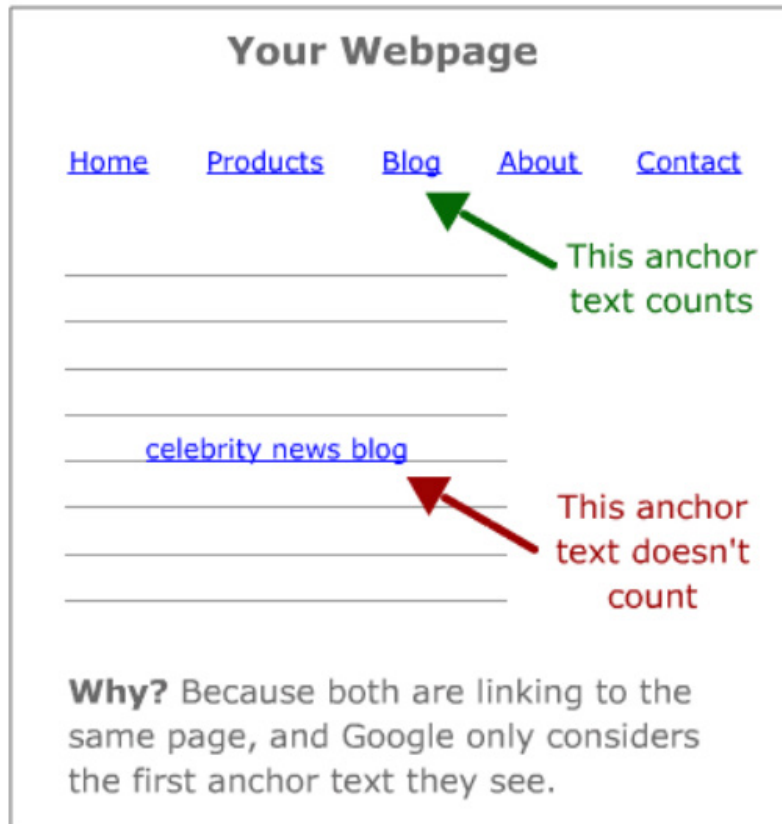
Sometimes, this involves using tracking parameters in URLs within the site. This causes duplicate content issues due to linking to different URLs that have the exact same content. This can be resolved in a number of ways.

Action Items: Tracking Parameters in URLs

- Avoid using tracking parameters in the URL. Instead, track these by using JavaScript tracking onclick event handlers on links that will pass the same tracking parameters. If using Google Analytics, this can be done with event tracking.
- Always using a self-referencing canonical tag is a good practice to have to avoid many kinds of duplicate content issues.

The First Link Priority

in search engine crawling where only the first link is considered and the duplicate link is disregarded. This has been discussed in forums and [tested in 2008](#) by a number of people, including Rand Fishkin and myself.



First Link Priority as Illustrated on Moz, by Rand Fishkin

A few things worth mentioning:

- In 2014 Matt Cutts, former head of Google's spam team, [said this is no longer an issue.](#) I have yet to test this again and I haven't seen any other SEO professionals test this recently.
- When this was first tested and detected to be an issue, the HTML version was 4.1, XHTML 1.1 was on the rise, and HTML 5 did not yet exist. Today, HTML 5 exists with tags like <header>, <article>, and <sidebar>. Maybe this time Google treats links in the header, sidebar, and article tags.

SEO Issues That Arise From the First Link Priority

Top-bar navigation and left side-bar often comes first within the source code before the main content. Additionally, navigational elements in these menus often have short anchor text. They tend to be less keyword focused and more design focused.

Links within the main content of a page have a tendency to be more keyword focused, with surrounding content that supports the keyword. They are also more flexible in length, with longer, more specific anchor text; this longer text increases the variety of keywords that a page can potentially rank for. However, because of first link priority issues, these links are often overlooked by search engines.

Action Items: First Link Priority Issue

- Consider code order. Prioritize the main content before the sidebar and top bar navigation. CSS can be used to control float direction, from left to right or right to left to make the sidebar navigation load after the main content. The top bar navigation can be controlled with absolute positioning.

Handling Navigation in Large Websites

For large websites (those with hundreds of thousands or millions pages), website navigation can be a huge challenge. The natural site navigation within categorized menus generally links to all pages of the site, and an XML sitemap can help index all pages. However, the lack of cross-linking between content silos can create distance between pages.

On a large site, it can be difficult to identify all possible links between product pages and the corresponding product marketing pages. Some sections of large sites may not be receiving much of the link love they need from other pages. Additionally, other issues like the first link priority and issues with JavaScript could be hard to detect across millions of pages.

Here are three solutions to these challenges:

1. Delegate to Different Departments

Large companies have proportionately large websites with multiple employees belonging to different departments. Many departments may correspond to different sections of the website.

Make sure that everyone involved in maintaining the different website sections abides by the same SEO principles and practices. Then, distribute the labor in optimizing navigation across the whole website.

2. Use Tools or Build Tools

Automation always makes manual processes more scalable. Unless you have your own proprietary tool, there may not be a single tool to identify and fix all issues mentioned above.

Crawling tools like Xenu, Screaming Frog, DeepCrawl, or Botify can analyze your existing links, determine the issues, and provide a description of the site architecture. If you want to visualize the site architecture, tools like DynoMapper and PowerMapper can help achieve this.

Link research tools like Moz's Open Site Explorer, Ahrefs, Majestic, Sistrix, LRT, and CognitiveSEO can analyze which pages get the most backlinks externally then add cross-links from these pages leading to more important pages of the site. The proprietary tool we use automates the process of crawling the page and determining which pages link to one another.

3. Use a Phased Approach

Large websites don't always have large teams behind them to distribute the work of optimizing pages. If there is a lack of resources, you can create your own tools to ease this process. If these tools do not provide the help you need, then consider a phased approach. This entails working on one section at a time with an optimization schedule. This is a day-by-day process and may take longer, but relying on metrics like organic search traffic will help you determine what to optimize first.

7 Key Takeaways

- **Users come first:** Your website navigation should satisfy users first. Then, optimize your navigation for SEO performance. Never compromise the user experience.
- **Cross-linking between content silos:** Content relevancy between pages is important for ranking, which comes naturally in a well-categorized, hierarchical site architecture. However, this can have limitations when it lacks cross-linking between content silos where some pages are just too deep or too far away from receiving a good amount of link juice from other sources.
- **Blogs to products, products to blogs:** Create high-quality content that is helpful and relevant to your target audience. If these blog posts help in a product buying decision, then link to the blog post from the specific product page(s).

- **Tracking parameters:** Avoid using them; use the onClick event handler on links for tracking purposes. It is always safe to have a self-referencing canonical tag.
- **JavaScript links:** Avoid using JavaScript to write content and links. If there is no way around it, there are methods to make it work.
- **First link priority:** Ideally, main content comes first. Next, is the sidebar, followed by the top bar. Lastly, handle the footer. Further testing is needed to determine if this is really still a valid concern, but it doesn't hurt to stick to this method.
- **Huge websites:** Thousands to millions of pages are hard to do all of the above. Delegate to a team, automate tasks by using tools, or handle the issues one at a time.

Chapter 10

HTTP or HTTPS? Why You Need a Secure Site

SEJ
EBOOK

Written By
Jenny Halasz
President, JLH Marketing



When Google first started encouraging sites to go to HTTPS in May 2010, many webmasters scoffed at the idea.

After all, HTTPS was only for sites that have transactions or which collect personal information, right?

Then on August 6, 2014, **Google announced that they would be showing a preference for HTTPS** sites in search results. This led SEOs all over the world to declare that HTTPS was now mandatory, and a ranking factor.

Finally, Google amended its advice on May 13, 2015. They stated that HTTPS was not actually a ranking factor, just that when it came to certain types of queries, they'd show a preference for it. HTTPS was a "tiebreaker". Google doubled down on this on September 15 of that year.

Webmasters breathed a collective sigh of relief, as their SEOs and marketing directors stopped pushing HTTPS so hard. After all, migrating to HTTPS is a lot of work!

It requires that all of the former pages be redirected, that all images and other linked file types be secure, and back then, it could even slow down the server response time a bit as that "handshake" verification took place (this is no longer true).

Many SEOs Didn't Believe in HTTPS at First

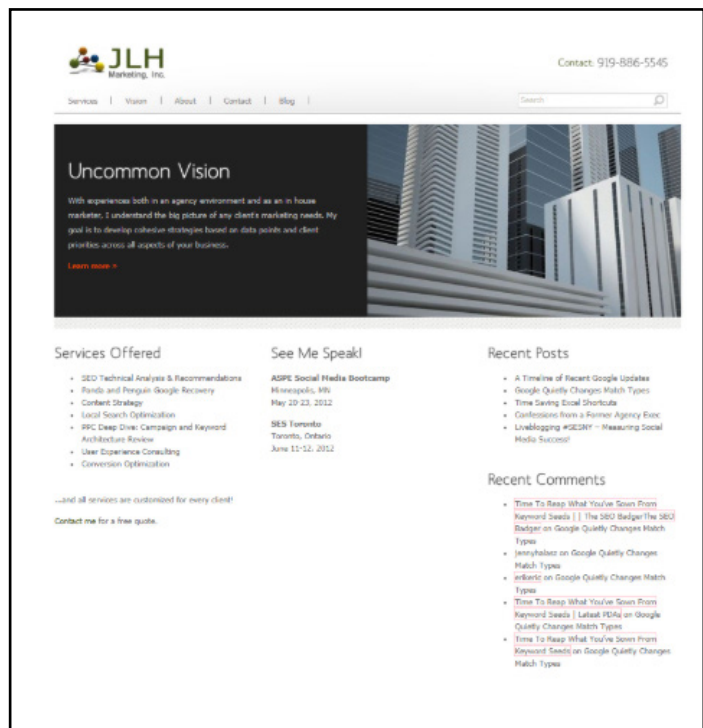
As all this was happening, I went on speaking about SEO, always indicating that I felt HTTPS was not that important unless you were collecting personal information or credit card numbers through your site.

But in 2012, I attended a conference where I learned something that would change the way I felt about HTTPS forever.

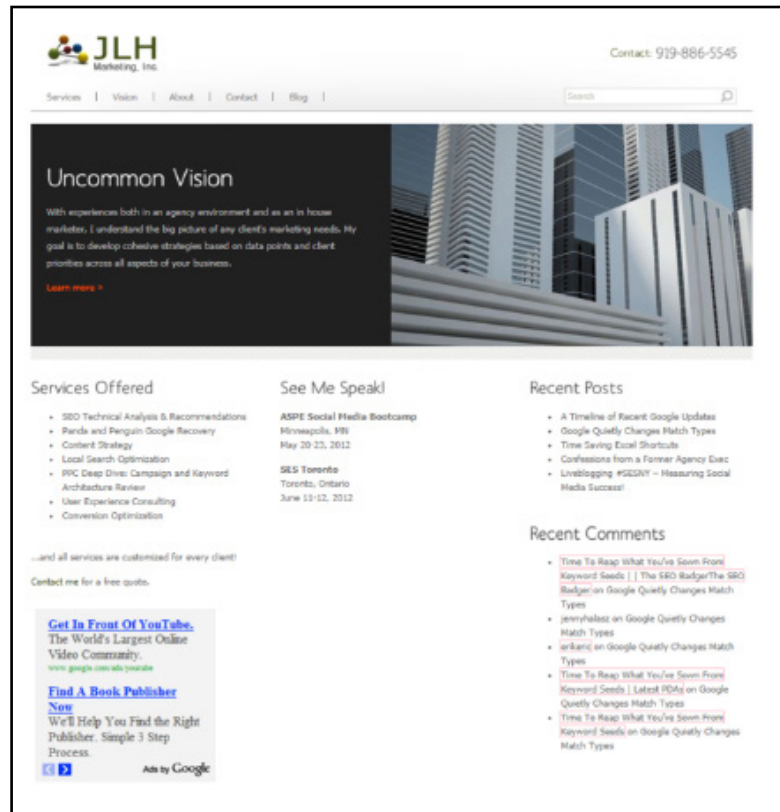
Every time I tell people this tidbit, they are surprised to learn about it. Are you ready for it?

Public Wi-Fi networks can insert advertising on your site if your site is not HTTPS.

Still not sure why that's a big deal? Here's what my website looked like back in 2012. It was not secure:



Here's what my website looked like that day my opinion on HTTPS changed forever.



Hint: The difference is the AdSense block in the lower left corner

You see, I was connected to the free Wi-Fi network provided by my hotel. I saw these ads show up on my website and immediately went into a tailspin... I could not understand how I could have ads on my site!

I didn't use AdSense; I had never added any ad code on my site. But there it was, right there in the HTML! I dug around in the code, thinking for sure that I'd been hacked.

Finally, I called the tech support number on the notepad by the phone:

“Hello Tech Support for XYZ Hotel Wi-Fi”

“Hi, can you tell me why I’m seeing ads on websites that I typically never see ads on?”

“Yes ma’am. The hotel uses Google AdSense to defray the cost of the free Wi-Fi service. The ads are dynamically inserted in applicable websites.”

I hung up the phone in shock. Really? The network could change what appeared in the code?

I tested a few other sites. Sure enough, there was my son’s pre-school. With an ad for a Las Vegas hotel in the bottom left corner – same place the ad on my site had been.

I checked a few others... the local police station... with an ad for a nearby restaurant.

The nearby mall had an ad for skin care products not sold in any of the stores at the mall.

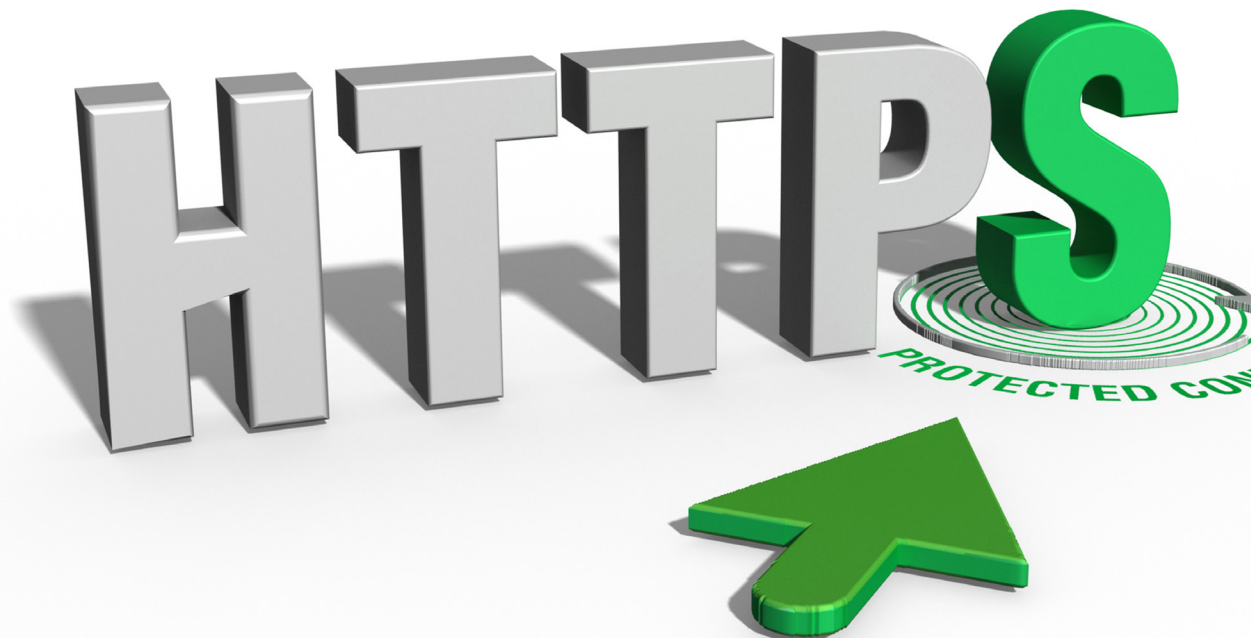
That’s when I realized that this had some serious possible consequences.

What if an ad for a steak restaurant started showing up on a site of vegan recipes? That would be completely off-brand and could potentially lose them a visitor.

Not to mention the other nefarious things people could potentially do to an insecure site.

I researched and realized that the protocol was what made this possible.

Without a public/private key pair (as is provided in HTTPS), an intermediary could easily intercept and steal or change any information before it got to its destination (the browser).



You Need to Go to HTTPS

This means that without that secure connection, any network that is between the source host and the destination host can change what the destination host gets.

If you don't understand why that's important, check out this trace route (tracert) from my home network to Google.com. Each one of these entries is a separate "hop" or server.

Without HTTPS, any one of these servers could change what Google delivered to my browser as a result (Google is HTTPS so that wouldn't happen).

```
Tracing route to google.com [172.217.14.100]
over a maximum of 30 hops:
  0  0 ms   0 ms   0 ms  ubnt [192.168.1.1]
  1  17 ms  22 ms  29 ms  104-174-100-140-1-100.res.nrc.com [174.109.140.1]
  2  25 ms  18 ms  20 ms  104-174-101-145-104-101ad.res.nrc.com [174.109.145.14]
  3  34 ms  14 ms  23 ms  104-174-101-141-104-101.res.nrc.com [174.109.141.104]
  4  23 ms  40 ms  61 ms  104-174-101-141-104-101.res.nrc.com [174.109.141.104]
  5  44 ms  55 ms  52 ms  104-174-101-141-104-101.res.nrc.com [174.109.141.104]
  6  82 ms  32 ms  32 ms  66.109.5.137
  7  32 ms  40 ms  30 ms  ix-ge-13-0.tor01.com.sbbn.net [216.6.103.103]
  8  70 ms  73 ms  34 ms  if-ge-1-3.tor01.com.sbbn.net [216.6.103.2]
  9  38 ms  38 ms  30 ms  72.14.170.10
 10  38 ms  51 ms  62 ms  108.170.249.1
 11  33 ms  61 ms  44 ms  216.239.59.11
 12  72 ms  41 ms  64 ms  216.239.59.11
 13  59 ms  33 ms  60 ms  216.239.59.11
 14  38 ms  34 ms  84 ms  209.85.248.103
 15  36 ms  35 ms  55 ms  108.170.249.103
 16  35 ms  48 ms  38 ms  108.170.249.103
 17  54 ms  76 ms  82 ms  at11111111-11-11-11-100.net [172.217.14.100]

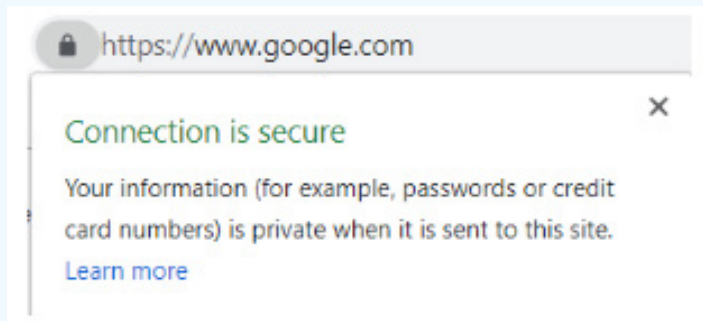
Trace complete.
```

For the safety and security of your visitors, your network, and yourself, you need to make your site HTTPS.

It really doesn't matter if your site is just a brochure site. Even if you don't collect any emails or have any login screens, you still need to migrate your site to HTTPS.

TL;DR? More Reasons You Should Switch to HTTPS

- **Protect Your Users' Information.** Make sure their data is protected as it passes through all of those hops to get to you.
- **Get the Lock Icon in the Browser Window.** It looks like this:



- **You Have to Have it to Implement AMP.** AMP technology only works on a secure server. AMP's creators designed it that way on purpose.
- **Protect Your Brand.** There's a lot more that can be inserted in websites beyond ads. Think pr0n, pills, and gambling.

- **Better Analytics Data.** HTTPS connections don't allow data from HTTP connections to be sent through HTTPS channels. If your site is not HTTPS, you can lose referrer data and other information from secure sites that link to your site.
- **Many Applications, Third Parties, and Browser Service Workers Will Not Support HTTP Sites.** If your site is not secure, you will have problems installing, creating and even using many third-party tools and scripts.
- **It's a Tie-Breaker for Google Ranking.** All things being equal, Google will choose to rank sites that are HTTPS before sites that are HTTP.

Ready to Switch to HTTPS?

We highly recommend this guide: [HTTP to HTTPS Migration: The Ultimate Stress-Free Guide.](#)

Chapter 11

How to Improve Page Speed for More Traffic & Conversions

SEJ
EBOOK

Written By
Jeremy Knauff
CEO, Spartan Media



Page speed is a critical factor in digital marketing today.

It has a significant impact on:

- How long visitors stay on your site.
- How many of them convert into paying customer.
- How much you pay on a CPC basis in paid search.
- Where you rank in organic search.

Unfortunately, most websites perform poorly when it comes to page speed, and that has a direct negative impact on their revenue.

There is an almost infinite number of things we can spend our days doing as digital marketers, and there's never enough time to do them all. As a result, some things get pushed to the back burner.

One of the things that seem to get pushed back most often is optimizing page speed. This is easy to understand because most people don't truly comprehend the importance of this often overlooked detail, so they don't see the value in investing time and money to improve it by a few seconds or less.

What may seem like an inconsequential amount of time to some marketers, including those who focus solely on search engine optimization, has been proven to be monumental by data from industry giants all the way down to our own analytics data.

I'll assume that you're like me and you want to maximize your results, and of course, your revenue, right? Then let's get started in making your website faster than greased snot! (That's quite a visual, isn't it?)

1. Ditch the Budget Web Hosting

We're all trying to save money these days, after all, those subscriptions to Raven, SEMrush, Moz, and all the other tools we use on a daily basis add up quickly. It's almost like having an extra kid.

One way a lot of people try to save money is by choosing the kind of cheap shared **hosting** that crams as many websites as they can fit onto a server, much like a bunch of clowns piling into a single car. Performance be damned!

Sure, your website will be available most of the time as it would with most any web host, but it will load so bloody slowly that your visitors will leave frustrated without ever converting into buyers.

"But it's barely noticeable!" these bargain shoppers insist. Here's the thing — it might be barely noticeable to you because it's your baby and you love it.

But everyone else only wants to get in and get out of your website as quickly as possible.

People want to be on your site for just long enough to do what they came to do, whether that means to get an answer, buy a product, or some other specific objective. If you slow them down even a little bit, they will be likely to hate their experience and leave without converting.



Think about it like this:

Most people love their own kids unconditionally. But someone else's kid screaming, throwing things, disrupting their night out at a restaurant? They hate that kid. It's the same with your website.

How Much of a Difference Does It Really Make?

According to a study conducted by [Amazon](#), a difference of just 100ms — a unit of time that a human can't even perceive, was enough to reduce their sales by 1 percent. [Walmart](#) found similar results.

If that tiny unit of time has that much direct impact on sales, what kind impact do you think an extra second or more will have?

But it doesn't stop there because how quickly (or slowly) your website loads also has an impact on organic search ranking and pay-per-click costs. In other words, if your website loads slowly, you should expect your competitors who have invested in this critical area to eat your lunch.

Bottom line: skip the budget web hosting. If they are selling it like a commodity (based mainly on price) then they'll treat their customers like a commodity too.

There are a lot of web hosts that are optimized for speed, particularly for WordPress websites, and some of them are priced similarly to the budget options. So ask around, do some testing, and invest in a web host that will give you the performance to satisfy both your visitors and Google.

2. Reduce HTTP Calls

Every file needed for a webpage to render and function, such as HTML, CSS, JavaScript, images, and fonts require a separate HTTP request. The more requests made, the slower that page will load.

Now if you're anything like most of the people I talk to, you're probably thinking "Oh, I don't need to worry about that, Jeremy. I know what I'm doing and I don't add a bunch of bloated garbage into my website!"

That may be partially true. You may not add a bunch of bloated garbage to your website, but for 90 percent+ of the websites that I encounter — it's still there anyway.

That bloat isn't there because the Bloat Fairy snuck it in while you were sleeping. It's there because a majority of web designers, regardless of skill or experience, don't make page speed a priority. The sad truth is that most don't even know how.

Here's where the problem starts:

Most themes load one or more CSS files and several JavaScript files. Some, such as JQuery or FontAwesome, are usually loaded remotely from another server, which dramatically increases the time it takes a page to load.

This becomes even more problematic when you consider the additional CSS and JavaScript files added by plugins. It's easy to end up with half a dozen or more HTTP requests just from CSS and JavaScript files alone.

When you factor in all of the images on a page, which each require a separate HTTP request, it quickly gets out of hand.

- Merge JavaScript files into one file.
- Merge CSS files into one file.
- Reduce or eliminate plugins that load their own JavaScript and/or CSS files. In some cases, as with Gravity Forms, you have the option to disable them from being loaded.
- Use sprites for frequently used images.
- Use a font like FontAwesome or Ionic Icons instead of image files wherever possible because then only one file needs to be loaded.

3. Include the Trailing Slash

Omitting the trailing slash on links pointing to your website, whether from external sources (link building efforts) or from within your own website, has an adverse impact on speed.

Here's how:

When you visit a URL without the trailing slash, the web server will look for a file with that name. If it doesn't find a file with that name, it will then treat it as a directory and look for the default file in that directory.

In other words, by omitting the trailing slash, you're forcing the server to execute an unnecessary 301 redirect. While it may seem instantaneous to you, it does take slightly longer, and as we've already established, every little bit adds up.

`https://example.com` (this is bad)

or

`https://example.com/services` (this is also bad)

vs

`https://example.com/` (this is good)

or

`https://example.com/services/` (this is also good)

4. Enable Compression

Enabling GZIP compression can significantly reduce the amount of time it takes to download your HTML, CSS, JavaScript files because they are downloaded as much smaller compressed files, which are then decompressed once they get to the browser.

Don't worry — your visitors won't have to do anything extra because all modern browsers support GZIP and automatically process it for all HTTP requests already.

5. Enable Browser Caching

With browser caching enabled, the elements of a webpage are stored in your visitors' browser so the next time they visit your site, or when they visit another page, their browser can load the page without having to send another HTTP request to the server for any of the cached elements.

Once the first page has been loaded and its elements are stored in the user's cache, only new elements need to be downloaded on subsequent pages. This can drastically reduce the number of files that need to be downloaded during a typical browsing session.

6. Minify Resources

Minifying your CSS and JavaScript files removes unnecessary white space and comments to reduce the file size, and as a result, the time it takes to download them.

Fortunately, this doesn't have to be a manual process because there are several tools available online to convert a file into a smaller, minified version of itself.

There are also several plugins available for WordPress that will replace the links in your website head for your regular CSS and JavaScript files with a minified version of them without modifying your original files, **including popular caching plugins such as:**

- W3 Total Cache
- WP Super Cache
- WP Rocket

It may take a bit of effort to get the settings just right because minification can often break CSS and JavaScript, so once you've minified everything, be sure to test your website thoroughly.

7. Prioritize Above-the-Fold Content

Your website can appear to the visitor to load more quickly if it's coded to prioritize above-the-fold content — in other words, the content that is visible before a visitor scrolls.

This means ensuring that any elements that appear above the fold are also as near the beginning of the HTML code so the browser can download and render them first.

It's also critical to include any CSS and JavaScript that are required to render that area inline rather than in an external CSS file.

8. Optimize Media Files

Because mobile devices with high-quality cameras are common and modern content management systems such as WordPress makes it convenient to upload images, many people simply shoot a photo and upload it without realizing that, often, the image is at least four times bigger than is necessary.

This slows your website down considerably — especially for mobile users.

Optimizing the media files on your website has the potential to improve your page speed tremendously, and doing so is relatively easy, so it's a good investment of your time.

Optimizing Images

- Opt for the ideal format. JPG is perfect for photographic images, while GIF or PNG are best for images with large areas of solid color. 8-bit PNG files are for images without an alpha channel (transparent background) and 24-bit files are for images with an alpha channel.
- Ensure images are properly sized. If an image is displayed at 800 pixels wide on your website, there is no benefit to using a 1600 pixels wide image.
- Compress the image file. Aside from being the top image editing program, Adobe Photoshop has awesome image compression capabilities and starts at \$9.99/month. You can also use free WordPress plugins – such as [WWW Image Optimizer](#), [Imsanity](#), and [TinyJPG](#) – that automatically compress uploaded images.

Optimizing Video

- Choose the ideal format. MP4 is best in most cases because it produces the smallest file size.
- Serve the optimal size (dimensions) based on visitors' screen size. Eliminate the audio track if the video is used in the background as a design element.
- Compress the video file. I use Adobe Premiere most of the time, but Camtasia is a solid choice too.
- Reduce the video length.

- Consider uploading videos to YouTube or Vimeo instead of serving them locally and use their iframe embedding code. You shouldn't stop there though because that only scratches the surface.
- To truly optimize the media on your website, you need to serve the appropriately-sized images based on the screen size rather than simply resizing them.
- There are two ways to handle this, based on the implementation of an image.

Images within the HTML of your website can be served using src set, which enables the browser to select, download, and display the appropriate image based on the screen size of the device a visitor is using.

Images placed via CSS – typically as background images, can be served using media queries to select the appropriate image based on screen size of the device a visitor is using.

9. Utilize Caching & CDNs

Caching enables your web server to store a static copy of your webpages so they can be delivered more quickly to a visitor's browser, while a CDN allows those copies to be distributed to servers all over the world so that a visitor's browser can download them from the server closest to their location. This improves page speed dramatically.

Chapter 12

7 Ways a Mobile-First Index Impacts SEO

SEJ
EBOOK

Written By
Roger Montti
Search Engine Journal



If you don't like change, then the Internet is not for you.

Google is constantly changing how they're indexing and ranking sites. It's realistic to expect more changes on the way.

I've identified seven insights about a **mobile-first index** and how that may influence rankings and SEO.

1. Mobile-First Informational Needs Are Changing

It may be inappropriate to generalize what kind of content is best for a mobile-first index. Every search query is different and how it is ranked in Google can be different.

Here is a sample of a few kinds of queries:

- Long tail queries
- Informational queries (what actor starred in...)
- Local search queries
- Transactional queries
- Research queries
- “How do I” queries?
- Conversational Search
- Personal Search

Personal Search & Conversational Search in Mobile

Personal Search and Conversational Search are the latest evolution in how people search. It is driven by mobile searches.

The way people search has changed because they are searching on phones. This must be taken into consideration when creating your search strategy.

Personal Search

According to Google's page on [Personal Searches](#):

"Over the past two years, we've seen an increase in searches that include highly personal and conversational language—using words like "me," "my," and "I."

- 60% + Growth in mobile searches for "__ for me" in the past two years.
- 80% + Growth in mobile searches for "__ should I __" in the past two years."

According to Google, Personal Searches fall into three categories:

- Solving a problem
- Getting things done
- Exploring around me

Conversational Search

Conversational search is a reference to the use of natural language in search queries. This means that users are literally speaking to their devices and expecting a natural response.

This is another change in how people search that is changing how we must think of content when creating content.

Many publishers, including Search Engine Journal, have experienced an increase in traffic by refashioning existing content to better meet the needs of mobile users.

According to Google's web page on [Conversational Search](#):

1. Mobile searches for “do I need” have grown over 65%.

For example, “how much do I need to retire,” “what size generator do I need,” and “how much paint do I need.”

2. Mobile searches for “should I” have grown over 65%.

For example, “what laptop should I buy,” “should I buy a house,” “what SPF should I use,” and “what should I have for dinner.”

3. Mobile searches starting with “can I” have grown over 85%.

For example, “can I use paypal on amazon,” “can I buy stamps at walmart,” and “can I buy a seat for my dog on an airplane.”

Mobile Search Trends Drive Content Relevance Trends

The above kinds of queries for both personal and conversational search are trending upwards and represent a meaningful change in what people are looking for. Content should adapt to that.

Each kind search query can be answered by a different kind of web page, with different content length, with different needs for diagrams, maps, depth, and so on.

One simply cannot generalize and say that Google prefers short form content because that's not always what mobile users prefer.

Thinking in terms of what most mobile users might prefer for a specific query is a great start.

But the next step involves thinking about the problem that a specific search query is trying to solve and what the best solution for most users is going to be.

Then crafting a content-based response that is appropriate for that situation.

And as you'll read below, for some queries the most popular answer might vary according to time. For some queries, a desktop optimal content might be appropriate.

2. Satisfy the Most Users

Identifying the problem users are trying to solve can lead to multiple answers.

If you look at the SERPs you will see there are different kinds of sites. Some might be review sites, some might be informational, some might be educational.

Those differences are indications that there multiple problems users are trying to solve. What's helpful is that Google is highly likely to order the SERPs according to the most popular user intent, the answer that satisfies the most users.

So if you want to know which kind of answer to give on a page, take a look at the SERPs and let the SERPs guide you.

Sometimes this means that most users tend to be on mobile and short-form content works best.

Sometimes it's fifty/fifty and most users prefer in-depth content or multiple product choices or fewer product choices.

Don't be afraid of the mobile index. It's not changing much.

It's simply adding an additional layer, to understand which kind of content satisfies the typical user (mobile, laptop, desktop, combination) and the user intent.

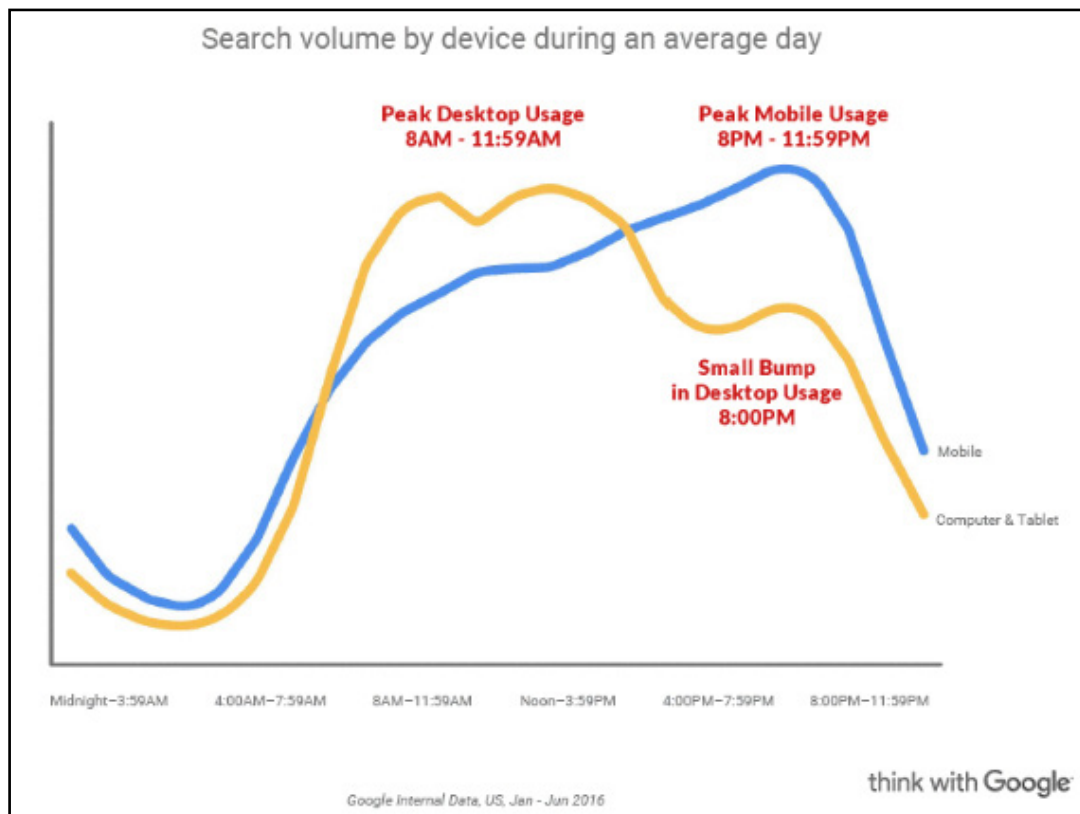
It's just an extra step to understanding who the most users are and from there asking how to satisfy them, that's all.

3. Time Influences Observed User Intent

Every search query demands a specific kind of result because the user intent behind each query is different. Mobile adds an additional layer of intent to search queries.

In a Think with Google publication about [how people use their devices](#) (PDF), Google stated this:

“The proliferation of devices has changed the way people interact with the world around them. With more touchpoints than ever before, it’s critical that marketers have a full understanding of how people use devices so that they can be here and be useful for their customers in the moments that matter.”



Time plays a role in how the user intent changes.

The time of day that a query is made can influence what device that user is using, which in turn says something about that users needs in terms of speed, convenience, and information needs.

Google's research from the above-cited document states this: "Mobile leads in the morning, but computers become dominant around 8 a.m. when people might start their workday. Mobile takes the lead again in the late afternoon when people might be on the go, and continues to increase into the evening, spiking around primetime viewing hours."

This is what I mean when I say that Google's mobile index is introducing a new layer of what it means to be relevant. It's not about your on-page keywords being relevant to what a user is typing.

A new consideration is about how your web page is relevant to someone at a certain time of day on a certain device and how you're going to solve the most popular information need at that time of day.

Google's March 2018 official mobile-first announcement

stated it like this:

"We may show content to users that's not mobile-friendly or that is slow loading if our many other signals determine it is the most relevant content to show."

What signals is Google looking at? Obviously, the device itself could be a signal.

But also, according to Google, time of day might be a signal because not only does device usage fluctuate during the day but the intent does too.

4. Defining Relevance in a Mobile-First Index

Google's focus on user intent 100 percent changes what the phrase "relevant content" means, especially in a mobile-first index.

People on different devices search for different things. It's not that the mobile index itself is changing what is going to be ranked.

The user intent for search queries is constantly changing, sometimes in response to Google's ability to better understand what that intent is.

Some of those core algorithm updates could be changes related to how Google understands what satisfies users.

You know how SEOs are worrying about click-through data? They are missing an important metric. CTR is not the only measurement tool search engines have.

Do you think CTR 100 percent tells what's going on in a mobile-first index? How can Google understand if a SERP solved a user's problem if the user does not even click through?

That's where a metric similar to [Viewport Time](#) comes in. Search engines have been using variations of Viewport Time to understand mobile users.

Yet the SEO industry is still wringing its hands about CTR. **Ever feel like a piece of the ranking puzzle is missing? This is one of those pieces.**

Google's understanding of what satisfies users is constantly improving. And that impacts the rankings. How we provide the best experience for those queries should change, too.

An important way those solutions have changed involves understanding the demographics of who is using a specific kind of device.

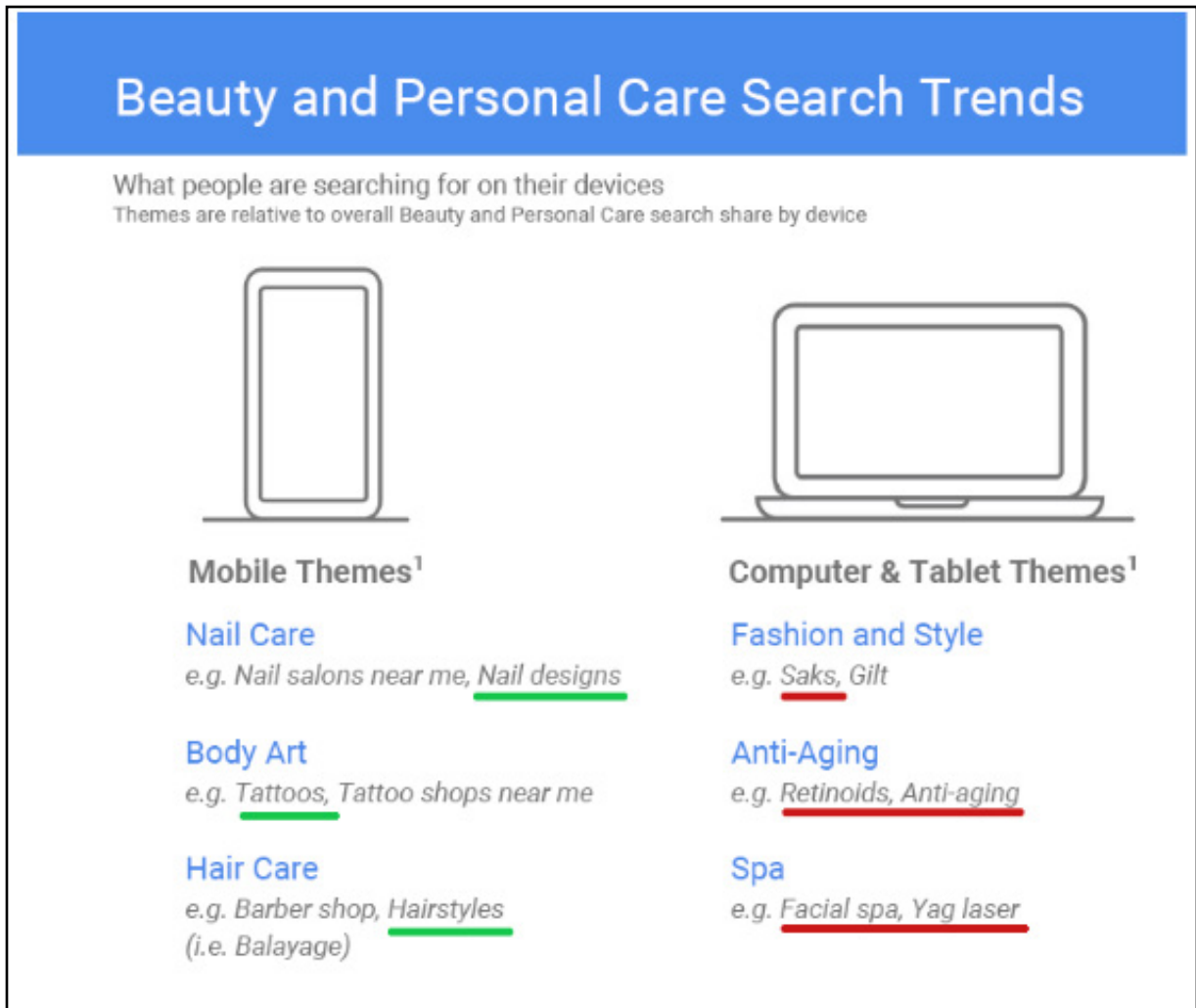
What does it mean when someone asks a question on one device versus another device?

One answer is that the age group might influence who is asking a certain question on a certain device.

For example, Google shared the following [insights about mobile and desktop users](#) (PDF). Searchers in the Beauty and Health niche search for different kinds of things according to device.

Examples of top beauty and health queries on mobile devices are for topics related to tattoos and nail salons.

Examples of Beauty and Health desktop queries indicate an older user because they're searching for stores like Saks and beauty products such as anti-aging creams.



It's naïve to worry about whether you have enough synonyms on your page. That's not what relevance is about.

Relevance is not about keyword synonyms. Relevance is often about problem-solving at certain times of day and within specific devices to specific age groups.

You can't solve that by salting your web page with synonyms.

5. Mobile First Is Not About User-Friendliness

An important quality of the mobile-first index is convenience when satisfying a user intent.

- Does the user intent behind the search query demand a quick answer or a shorter answer?
- Does the web page make it hard to find the answer?
- Does the page enable comparison between different products?

Now answer those questions by adding the phrase, on mobile, on a tablet, on a desktop and so on.



6. Would a Visitor Understand Your Content?

Google can know if a user understands your content. Users vote with their click and viewport time data and quality raters create another layer of data about certain queries.

With enough data Google can predict it what a user might find useful. This is where machine learning comes in.

Here's what Google says about machine learning in the context of User Experience (UX):

“Machine learning is the science of making predictions based on patterns and relationships that've been automatically discovered in data.”

If content that is difficult to read is a turn-off, that may be reflected in what sites are ranked and what sites are not.

If the topic is complex and a complex answer solves the problem then that might be judged the best answer.

I know we're talking about Google but it's useful to understand the state of the art of search in general.

Microsoft published a fascinating study about teaching a machine to predict what a user will find interesting. The paper is titled, [**Predicting Interesting Things in Text.**](#)

This research focused on understanding what made content interesting and what caused users to keep clicking to another page.

In other words, it was about training a machine to understand what satisfies users.

Here's a synopsis:

“We propose models of “interestingness”, which aim to predict the level of interest a user has in the various text spans in a document. We obtain naturally occurring interest signals by observing user browsing behavior in clicks from one page to another. We cast the problem of predicting interestingness as a discriminative learning problem over this data.

We train and test our models on millions of real world transitions between Wikipedia documents as observed from web browser session logs. On the task of predicting which spans are of most interest to users, we show significant improvement over various baselines and highlight the value of our latent semantic model.”

In general, I find good results with content that can be appreciated by the widest variety of people.

This isn't strictly a mobile-first consideration but it is increasingly important in an Internet where so people of diverse backgrounds are accessing a site with multiple intents multiple kinds of devices.

Achieving universal popularity becomes increasingly difficult so it may be advantageous to appeal to the broadest array of people in a mobile-first index.

7. Google's Algo Intent Hasn't Changed

Looked at a certain way, it could be said that Google's desire to show users what they want to see has remained consistent.

What has changed is the users' age, what they desire, when they desire it and what device they desire it on. So the intent of Google's algorithm likely remains the same.

The mobile-first index can be seen as a logical response to how users have changed. It's backwards to think of it as Google forcing web publishers to adapt to Google.

What's really happening is that web publishers must adapt to how their users have changed.

Ultimately that is the best way to think of the mobile-first index. Not as a response to what Google wants but to approach the problem as a response to the evolving needs of the user.

Chapter 13

The Complete Guide to Mastering Duplicate Content Issues

SEJ
EBOOK

Written By
Stoney G deGeyter
VP Search and Advertising, The Karcher Group



In the SEO arena of website architecture, there is little doubt that eliminating duplicate content can be one of the hardest fought battles.

Too many content management systems and piss-poor developers build sites that work great for displaying content but have little consideration for how that content functions from a search-engine-friendly perspective.

And that often leaves damaging duplicate content dilemmas for the SEO to deal with.

There are two kinds of duplicate content, and both can be a problem:

- Onsite duplication is when the same content is duplicated on two or more unique URLs of your site. Typically, this is something that can be controlled by the site admin and web development team.
- Offsite duplication is when two or more websites publish the exact same pieces of content. This is something that often cannot be controlled directly but relies on working with third-parties and the owners of the offending websites.

Why Is Duplicate Content a Problem?

The best way to explain why duplicate content is bad is to first tell you why unique content is good.

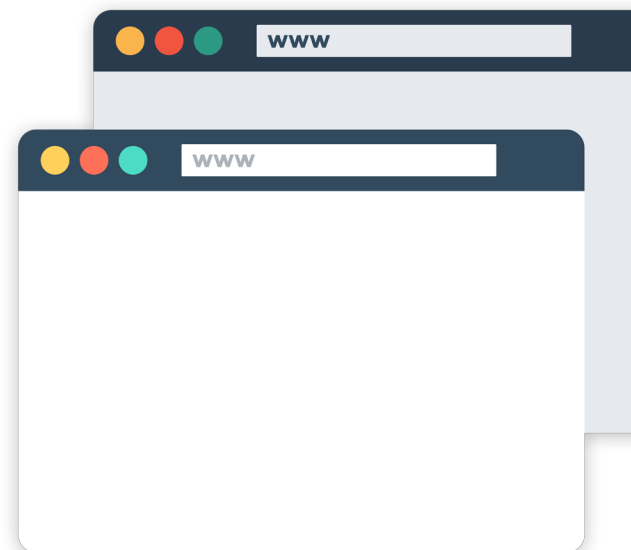
Unique content is one of the best ways to set yourself apart from other websites. When the content on your website is yours and yours alone, you stand out. You have something no one else has.

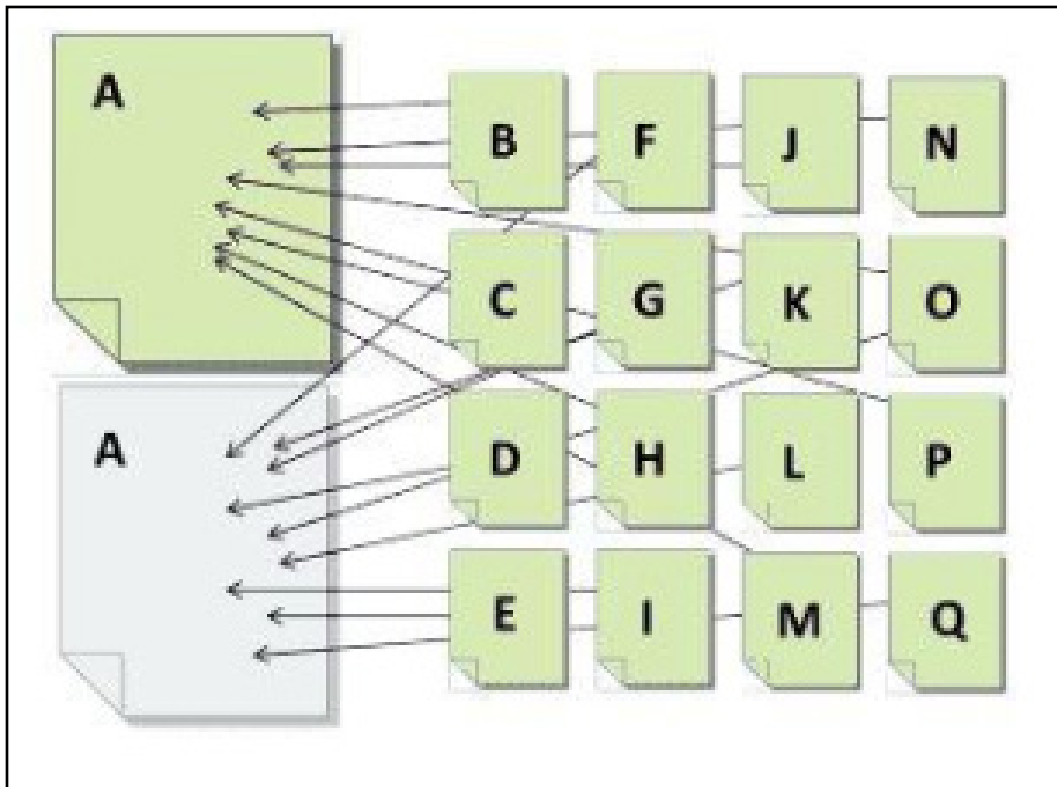
On the other hand, when you use the same content to describe your products or services or have content republished on other sites, you lose the advantage of being unique.

Or, in the case of onsite duplicate content, individual pages lose the advantage of being unique.

Look at the illustration below. If A represents content that is duplicated on two pages, and B through Q represents pages linking to that content, the duplication causes a split the link value being passed.

Now imagine if pages B-Q all linked to only on page A. Instead of splitting the value each link provides, all the value would go to a single URL instead, which increases the chances of that content ranking in search.





Whether onsite or offsite, all duplicate content competes against itself. Each version may attract eyeballs and links, but none will receive the full value it would get if it were the sole and unique version.

However, when valuable and unique content can be found on no more than a single URL anywhere on the web, that URL has the best chance of being found based on it being the sole collector of authority signals for that content.

Now, having that understanding, let's look at the problems and solutions for duplicate content.

Offsite Duplicate Content

Offsite duplication has three primary sources:

- Third-party content you have republished on your own site. Typically, this is in the form of generic product descriptions provided by the manufacturer.
- Your content that has been republished on third-party sites with your approval. This is usually in the form of article distribution or perhaps reverse article distribution.
- Content that someone has stolen from your site and republished without your approval. This is where the content scrapers and thieves become a nuisance.

Let's look at each.

Content Scrapers & Thieves

Content scrapers are one of the biggest offenders in duplicate content creation. Spammers and other nefarious perpetrators build tools that grab content from other websites and then publish it on their own.

For the most part, these sites are trying to use your content to generate traffic to their own site in order to get people to click their ads. (Yeah, I'm looking at you, Google!)

Unfortunately, there isn't much you can do about this other than to submit a [**copyright infringement report**](#) to Google in hopes that it will be removed from their search index. Though, in some cases, submitting these reports can be a full-time job.

Another way of dealing with this content is to ignore it, hoping Google can tell the difference between a quality site (yours) and the site the scraped content is on. This is hit and miss as I've seen scraped content rank higher than the originating source.

What you can do to combat the effects of scraped content is to utilize absolute links (full URL) within the content for any links pointing back to your site. Those stealing content generally aren't in the business of cleaning it up so, at the very least, visitors can follow that back to you.

You can also try adding a canonical tag back to the source page (a good practice regardless). If the scrapers grab any of this code, the canonical tag will at least provide a signal for Google to recognize you as the originator.

Article Distribution

Several years ago, it seemed like every SEO was republishing their content on "ezines" as a link building tactic. When Google cracked down on content quality and link schemes, republishing fell by the wayside.

But with the right focus, it can be a solid marketing strategy. Notice, I said "marketing" rather than "SEO" strategy.

For the most part, any time you're publishing content on other websites, they want the unique rights to that content.

Why? Because they don't want multiple versions of that content on the web devaluing what the publisher has to offer.

But as Google has gotten better about assigning rights to the content originator (better, but not perfect), many publishers are allowing content to be reused on the author's personal sites as well.

Does this create a duplicate content problem? In a small way, it can, because there are still two versions of the content out there, each potentially generating links.

But in the end, if the number of duplicate versions is limited and controlled, the impact will be limited as well. In fact, the primary downside lands on the author rather than the secondary publisher.

The first published version of the content will generally be credited as the canonical version. In all but a few cases, these publishers will get more value from the content over the author's website that republishes it.

Generic Product Descriptions

Some of the most common forms of duplicated content comes from product descriptions that are reused by each (and almost every) seller.

A lot of online retailers sell the exact same products as thousands of other stores. In most cases, the product descriptions are provided by the manufacturer, which is then uploaded into each site's database and presented on their product pages.

While the layout of the pages will be different, the bulk of the product page content (product descriptions) will be identical.

Now multiply that across millions of different products and hundreds of thousands of websites selling those products, and you can wind up with a lot of content that is, to put it mildly, not unique.

How does a search engine differentiate between one or another when a search is performed?

On a purely content-analysis level, it can't. Which means the search engine must look at other signals to decide which one should rank.

One of these signals is links. Get more links and you can win the bland content sweepstakes.

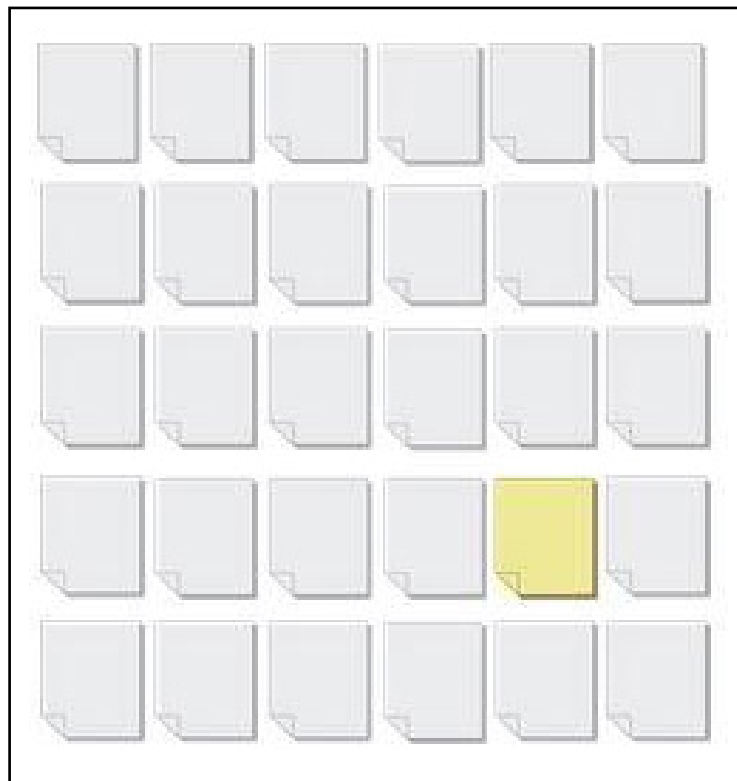
But if you're up against a more powerful competitor, you may have a long battle to fight before you can catch them in the link building department. Which brings you back to looking for another competitive advantage.

The best way to achieve that is by taking the extra effort to write unique descriptions for each product. Depending on the number

of products you offer, this could end up being quite a challenge, but in the end, it'll be well worth it.

Take a look at the illustration below. If all the gray pages represent the same product with the same product descriptions, the yellow represents the same product with a unique description.

If you were Google, which one would you want to rank higher?



Any page with unique content is going to automatically have an inherent advantage over similar but duplicate content. That may or may not be enough to outrank your competition, but it surely is the baseline for standing out to not just Google, but your customers as well.

Onsite Duplicate Content

Technically, Google treats all duplicate content the same, so onsite duplicate content is really no different than offsite.

But onsite is less forgivable because this is one type of duplication that you can actually control. It's shooting your SEO efforts in the proverbial foot.

Onsite duplicate content generally stems from bad site architecture. Or, more likely, bad website development!

A strong site architecture is the foundation for a strong website. When developers don't follow search-friendly best practices, you can wind up losing valuable opportunity to get your content to rank due to this self-competition.

There are some who argue against the need for good architecture, citing Google propaganda about how Google can "figure it out." The problem with that is that it relies on Google figuring things out.

Yes, Google can determine that some duplicate content should be considered one and the same, and the algorithms can take this into account when analyzing your site, but that's no guarantee they will.

Or another way to look at it is that just because you know someone smart doesn't necessarily mean they'll be able to protect you from your own stupidity! If you leave things to Google and Google fails, you're screwed.

Now, let's dive into some common onsite duplicate content problems and solutions.

The Problem: Product Categorization Duplication

Far too many ecommerce sites suffer from this kind of duplication. This is frequently caused by content management systems that allow you to organize products by category, where a single product can be tagged in multiple categories.

That in itself isn't bad (and can be great for the visitor), however in doing so, the system generates a unique URL for each category in which a single product shows up in.

Let's say you're on a home repair site and you're looking for a book on installing bathroom flooring.

You might find the book you're looking for by following any of these navigation paths:

- Home > flooring > bathroom > books
- Home > bathroom > books > flooring
- Home > books > flooring > bathroom

Each of these is a viable navigation path, but the problem arises when a unique URL is generated for each path:

- <https://www.myfakesite.com/flooring/bathroom/books/fake-book-by-fake-author>
- <https://www.myfakesite.com/bathroom/books/flooring/fake-book-by-fake-author>
- <https://www.myfakesite.com/books/flooring/bathroom/fake-book-by-fake-author>

I've seen sites like this create up to ten URLs for every single product turning a 5k product website into a site with 45k duplicate pages. That is a problem.

If our example product above generated ten links, those links would end up being split three ways. Whereas, if a competitor's page for the same product got the same ten links, but to only a single URL, which URL is likely to perform better in search? The competitor's! Not only that, but search engines limit their crawl bandwidth so they can spend it on indexing unique and valuable content.

When your site has that many duplicate pages, there is a strong chance the engine will stop crawling before it even gets a fraction of your unique content indexed.

This means hundreds of valuable pages won't be available in search results and those that are indexed are duplicates competing against each other.

The Solution: Master URL Categorizations

One fix to this problem is to only tag products for a single category rather than multiples. That solves the duplication issue, but it's not necessarily the best solution for the shoppers since it eliminates the other navigation options for finding the product(s) they want. So, scratch that one off the list.

Another option is to remove any type of categorization from the URLs altogether.

This way, no matter the navigation path used to find the product, the product URL itself is always the same, and might look something like this:

- <https://www.myfakesite.com/products/fake-book-by-fake-author>

This fixes the duplication without changing how the visitor is able to navigate to the products. The downside to this method is that you lose the category keywords in the URL. While this provides a small benefit to the totality of SEO, every little bit can help.

If you want to take your solution to the next level, getting the most optimization value possible while keeping the user experience at the same time, build an option that allows each product to be assigned to a “master” category, in addition to others.

When a master category is in play, the product can continue to be found through the multiple navigation paths, but the product page is accessed by a single URL that utilizes the master category.

That might make the URL look something like this:

- <https://www.myfakesite.com/flooring/fake-book-by-fake-author> OR
- <https://www.myfakesite.com/bathroom/fake-book-by-fake-author> OR
- <https://www.myfakesite.com/books/fake-book-by-fake-author>

This latter solution is the best overall, though it does take some additional programming. However, there is one more relatively easy “solution” to implement, but I only consider it a band-aid until a real solution can be implemented.

Band-Aid Solution: Canonical Tags

Because the master-categorization option isn’t always available to out of the box CMS or ecommerce solutions, there is an alternative option that will “help” solve the duplicate content problem.

This involves preventing search engines from indexing all non-canonical URLs. While this can keep duplicate pages out of the search index, it doesn’t fix the issue of splitting the page’s authority. Any link value sent to a non-indexable URL will be lost.

The better band-aid solution is to utilize canonical tags. This is similar to selecting a master category but generally requires little, if any, additional programming.

You simply add a field for each product that allows you to assign a canonical URL, which is just a fancy way of saying, “the URL you want to show up in search.”

The canonical tag looks like this:

- `<link rel="canonical" href="https://www.myfakesite.com/books/fake-book-by-fake-author" />`

Despite the URL the visitor is on, the behind-the-scenes canonical tag on each duplicate URL would point to a single URL.

In theory, this tells the search engines not to index the non-canonical URLs and to assign all other value metrics over to the canonical version. This works most of the time, but in reality, the search engines only use the canonical tag as a “signal.” They will then choose to apply or ignore it as they see fit.

You may or may not get all link authority passed to the correct page, and you may or may not keep non-canonical pages out of the index. I always recommend implementing a canonical tag, but because it’s unreliable, consider it a placeholder until a more official solution can be implemented.

The Problem: Redundant URL Duplication

One of the most basic website architectural issues revolves around how pages are accessed in the browser.

By default, almost every page of your site can be accessed using a slightly different URL. If left unchecked, each URL leads to the exact same page with the exact same content.



Considering the home page alone, it can likely be accessed using four different URLs:

- `http://site.com`
- `http://www.site.com`
- `https://site.com`
- `https://www.site.com`

And when dealing with internal pages, you can get an additional version of each URL by adding a trailing slash:

- `http://site.com/page`
- `http://site.com/page/`
- `http://www.site.com/page`
- `http://www.site.com/page/`
- Etc.

That's up to eight alternate URLs for each page! Of course, Google should know that all these URLs should be treated as one, but which one?

The Solution: 301 Redirects & Internal Link Consistency

Aside from the canonical tag, which I addressed above, the solution here is to ensure you have all alternate versions of the URLs redirecting to the canonical URL.

Keep in mind, this isn't just a home page issue. The same issue applies to every one of your site URLs. Therefore, the redirects implemented should be global.

Be sure to force each redirect to the canonical version. For instance, if the canonical URL is <https://www.site.com>, each redirect should point there.

Many make the mistake of adding additional redirect hops that might look like this:

- Site.com > <https://site.com> > <https://www.site.com>
- Site.com > www.site.com > <https://www.site.com>

Instead, the redirects should look like this:

- <http://site.com> > <https://www.site.com/>
- <http://www.site.com> > <https://www.site.com/>
- <https://site.com> > <https://www.site.com/>
- <https://www.site.com> > <https://www.site.com/>
- <http://site.com/> > <https://www.site.com/>
- <http://www.site.com/> > <https://www.site.com/>
- <https://site.com/> > <https://www.site.com/>

By reducing the number of redirect hops you speed up page load, reduce server bandwidth, and have less that can go wrong along the way.

Finally, you'll need to make sure all internal links in the site point to the canonical version as well.

While the redirect should solve the duplicate problem, redirects can fail if something goes wrong on the server or implementation side of things.

If that happens, even temporarily, having only the canonical pages linked internally can help prevent a sudden surge of duplicate content issues from popping up.

The Problem: URL Parameters & Query Strings

Years ago, the usage of session IDs created a major duplicate content problem for SEOs.

Today's technology, however, has made session IDs all but obsolete, but another problem has arisen that is just as bad, if not worse: URL parameters.

Parameters are used to pull fresh content from the server, usually based on one or more filter or selections being made.

The two examples below show alternate URLs for a single URL: `site.com/shirts/`.

The first shows the shirts filtered by color, size, and style, the second URL shows shirts sorted by price, then a certain number of products to show per page:

- Site.com/shirts/?color=red&size=small&style=long_sleeve
- Site.com/shirts/?sort=price&display=12

Based on these filters alone, there are three viable URLs that search engines can find.

But the order of these parameters can change based on the order in which they were chosen, which means you might get several more accessible URLs like this:

- Site.com/shirts/?size=small&color=red&style=long_sleeve
- Site.com/shirts/?size=small&style=long_sleeve&color=red
- Site.com/shirts/?display=12&sort=price

And this:

- Site.com/shirts/?size=small&color=red&style=long_sleeve&display=12&sort=price
 - Site.com/shirts/?display=12&size=small&color=red&sort=price
 - Site.com/shirts/?size=small&display=12&sort=price&color=red&style=long_sleeve
- Etc.

You can see that this can produce a lot of URLs, most of which will not pull any type of unique content. Of the parameters above, the only one you might want to write sales content for is the style. The rest, not so much.

The Solution: Parameters for Filters, Not Legitimate Landing Pages

Strategically planning your navigation and URL structure is critical for getting out ahead of the duplicate content problems.

Part of that process includes understanding the difference between having a legitimate landing page and a page that allows visitors to filter results. And then be sure to treat these accordingly when developing the URLs for them.

Landing page (and canonical) URLs should look like this:

- Site.com/shirts/long-sleeve/
- Site.com/shirts/v-neck/
- Site.com/shirts/collared/

And the filtered results URLs would look something like this:

- Site.com/shirts/long-sleeve/?size=small&color=red&display=12&sort=price
- Site.com/shirts/v-neck/?color=red
- Site.com/shirts/collared/?size=small&display=12&sort=price&color=red

With your URLs built correctly, you can do two things:

- Add the correct canonical tag (everything before the “?” in the URL).
- Go into Google Search Console and tell Google to ignore all such parameters.

If you consistently use parameters only for filtering and sorting content, you won't have to worry about accidentally telling Google not to crawl a valuable parameter... because none of them are.

But because the canonical tag is only a signal, you must complete step two for best results. And remember this only affects Google. You have to do the same with Bing.

Pro Developer Tip: Search engines typically ignore everything to the right of a pound “#” symbol in the URL.

If you program that into every URL prior to any parameter, you won't have to worry about the canonical being only a band-aid solution:

- Site.com/shirts/long-sleeve/#?size=small&color=red&display=12
- &sort=price
- Site.com/shirts/v-neck/#?color=red
- Site.com/shirts/

collared/#?size=small&display=12&sort=price&color=red
If any search engine were to access the URLs above, they would only index the canonical part of the URL and ignore the rest.

The Problem: Ad Landing Page & A/B Test Duplication

It's not uncommon for marketers to develop numerous versions of similar content, either as a landing page for ads, or A/B/multivariate testing purposes.

This can often get you some great data and feedback, but if those pages are open for search engines to spider and index, it can create duplicate content problems.

The Solution: NoIndex

Rather than use a canonical tag to point back to the master page, the better solution here is to add a noindex meta tag to each page to keep them out of the search engines' index altogether.

Generally, these pages tend to be orphans, not having any direct links to them from inside the site. But that won't always keep search engines from finding them.

The canonical tag is designed to transfer page value and authority to the primary page, but since these pages should not be collecting any value, keeping them out of the index is preferred.

When Duplicate Content Isn't (Much Of) a Problem

One of the most common SEO myths is that there is a duplicate content penalty.

There isn't. At least no more than there is a penalty for not putting gas in your car and letting it run empty.

Google may not be actively penalizing duplicate content, but that doesn't mean there are not natural consequences that occur because of it.

Without the threat of penalty, that gives marketers a little more flexibility in deciding which consequences they are willing to live with.

While I would argue that you should aggressively eliminate (not just band-aid over) all on-site duplicate content, offsite duplication may actually create more value than consequences.

Getting valuable content republished off-site can help you build brand recognition in a way that publishing it on your own can't.

That's because many offsite publishers have a bigger audience and a vastly larger social reach.

Your content, published on your own site may reach thousands of eyeballs, but published offsite it might reach hundreds of thousands.

Many publishers do expect to maintain exclusive rights to the content they publish, but some allow you to repurpose it on your own site after a short waiting period. This allows you to get the additional exposure while also having the opportunity to build up your own audience by republishing your content on your site at a later date.

But this type of article distribution needs to be limited in order to be effective for anyone. If you're shooting your content out to hundreds of other sites to be republished, the value of that content diminishes exponentially.

And typically, it does little to reinforce your brand because the sites willing to publish mass duplicated content are of little value to begin with.

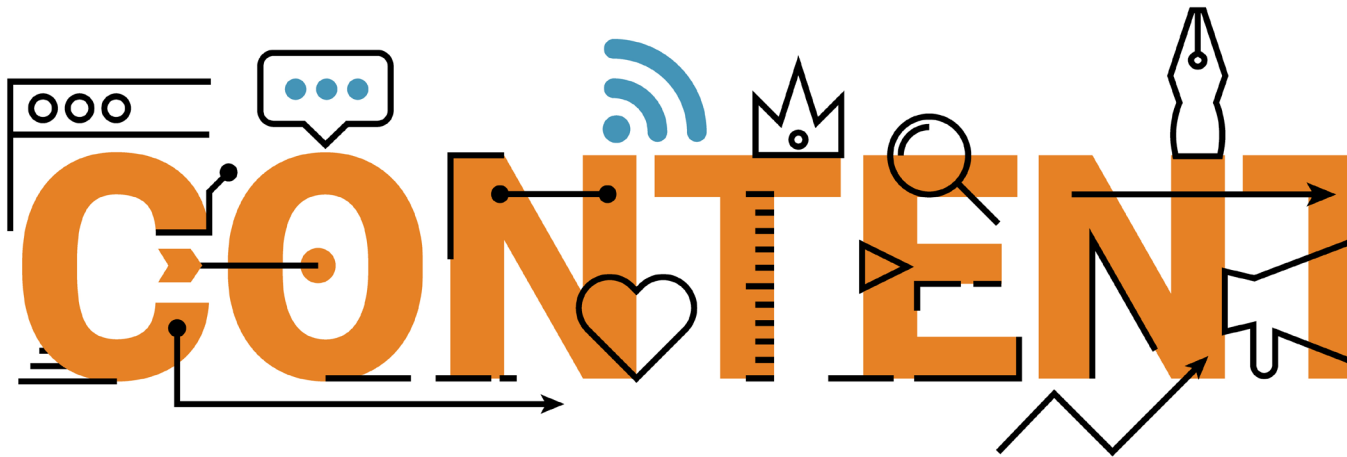
In any case, weigh the pros and cons of your content being published in multiple places.

If duplication with a lot of branding outweighs the smaller authority value you'd get with unique content on your own site, then, by all means, pursue a measured republishing strategy.

But the keyword there is measured. What you don't want to be is the site that only has duplicate content. At that point, you begin to undercut the value you're trying to create for your brand.

By understanding the problems, solutions and, in some cases, value, of duplicate content, you can begin the process of eliminating the duplication you don't want and pursuing the duplication you do.

In the end, you want to build a site that is known for strong, unique content, and then use that content to get the highest value possible.

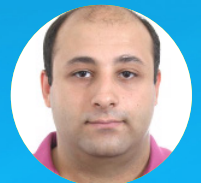


Chapter 14

A Technical SEO Guide to Redirects

SEJ
EBOOK

Written By
Vahan Petrosyan
Lead Developer, Search Engine Journal



Websites change structure, delete pages and often move from one domain to another.

Handling **redirects** correctly is crucial in order to avoid losing rankings and help search engines understand the changes you have done.

Redirects have a status code starting with number three (i.e., 3XX). There are **100 different** possible status codes but only a **few are implemented** to carry certain information.

In this guide, we will cover 3XX redirects relevant to SEO.

301: Moved Permanently

This well-known redirect indicates to a client* that the resource was changed to another location and that it should use the new URL for future requests. When search engines see a 301 redirect, they pass the old page's ranking to the new one.

Before making a change, you need to be careful when deciding to use a 301 redirect. This is because if you change your mind later and decide to remove the 301 redirect, your old URL may not rank anymore.

Even if you swap the redirects, it will not help you get the old page back to its previous ranking position. So the main thing to remember is that there's no way to undo a 301 redirect.

*(*For beginners who may get confused with generic name client is used instead of browser since not only browsers are able to browse URLs but also search engine bots which are not browsers.)*

307: Temporary Redirect

In HTTP 1.1, a 301 redirect means the resource is temporarily moved and the client should use the **original resource's** URL for future requests.

For SEO, this means the client should follow a redirect but search engines should not update their links in the SERPs to the new, temporary page.

In a 307 redirect, PageRank is not passed from the original resource to the new one – contrary to a 301 redirect.

302: Found

This means that the resource a client is looking for was found on another URL in the HTTP 1.1 version but was temporarily moved in HTTP 1.0.

302 vs. 307

In almost all cases, 302 and 307 redirects will be **treated the same.** But a 302 status code doesn't necessarily mean the client must follow a redirect and it is not considered an error if it decides to stay there.

Modern clients will most likely follow the new destination but some old clients may incorrectly stay on the same URL.

Contrary to a 302 status code, the 307 status code guarantees that the request method will not be changed. For instance, the GET request must continue to GET and POST to POST.

With a 302 status code, some old or buggy clients may change the method which may cause unexpected behavior.

For temporary redirects, you can use either 302 or 307 – but I do prefer 307.

For routine redirect tasks, 301 (permanent redirect) and 307 (temporarily redirect) status codes should be used depending on what type of change you are implementing on your website. On both cases, the syntax of redirects doesn't change.

You may handle redirect via server config files .htaccess on Apache, example.conf file on Ngix, or via plugins if you are using WordPress. In all instances, they have the same syntax for writing redirect rules. They differ only with commands used in configuration files.

For example, redirect on Apache will look like this:

Options +FollowSymlinks

RewriteEngine on

RedirectMatch 301 ^/oldfolder/ /newfolder/

(you can read about symlinks [here](#)). and on Ngix servers like

rewrite ^/oldfolder/ /newfolder/ permanent;

The commands used to tell servers status code of redirect and the action command differ. For instance:

- Servers status code of redirect: “301” vs. “permanent”
- Action command: “RedirectMatch” vs. “rewrite”.

But the syntax of the redirect (^/oldfolder/ /newfolder/) is the same for both.

On Apache, make sure on your server mod_rewrite and mod_alias modules (which are responsible for handling redirects) are enabled. Since the most widely spread server types is Apache, here are examples for .htaccess apache files. Make sure that the .htaccess file has these two lines

Options +FollowSymlinks

RewriteEngine on

above the redirect rules and put the rules below them.

For understanding the examples below you may refer table below on [RegExp](#) basics.

*	zero or more times
+	One or more times
.	any single character
?	Zero or one time
^	Start of the string
\$	End of the string
a b	OR operadn " " a or b
(z)	remembers the match to be used when calling \$1

Redirect Single URL

The most common and widely used type of redirect that is used when deleting pages or changing page URLs. For instance, say you changed URL from /old-page/ to /new-page/. The redirect rule would be:

```
RewriteRule ^old-page(/?|/.*)$ /new-page/ [R=301,L]
```

OR

```
RedirectMatch 301 ^/old-page(/?|/.*)$ /new-page/
```

The only difference between the two methods is that the first one uses Apache mod_rewrite module and the second one uses mod_alias. It can be done using both methods.

Regular expression “^” means URL must start with “/old-page” while “(/?|/.*)\$” indicates that **anything that follows “/old-page/” with slash “/” or without exact match must be redirected to /new-page/.**

We could also use “(*)” ie. “^/old-page(*)”, but the problem is, if you have another page with a similar URL like /old-page-other/, it will also be redirected when we only want to redirect /old-page/.

The following URLs will match and directed to new page

<i>/old-page/</i>	<i>/new-page/</i>
<i>/old-page</i>	<i>/new-page/</i>
<i>/old-page/?utm_source=facebook.com</i>	<i>/new-page/?utm_source=facebook.com</i>
<i>/old-page/child-page/</i>	<i>/new-page/</i>

It will redirect any variation of page URL to new one.

Redirect All Except

Let's say we have bunch of URLs like `/category/old-subcategory-1/`, `/category/old-subcategory-2/`, `/category/final-subcategory/` and want to merge all subcategories into `/category/final-subcategory/`. We need here "all except" rule

```
RewriteCond %{REQUEST_URI} !/category/final-subcategory/  
RewriteCond %{REQUEST_FILENAME} !-f  
RewriteRule ^(category)/. /category/final-subcategory/ [R=301,L]
```

Here, we want to redirect all under `/category/` on the fifth line except if it is `/category/final-subcategory/` on the fourth line. We also have "!" rule on the fourth line which means to ignore any file like images, CSS or javascript files.

Otherwise, if we have some assets like `/category/image.jpg` it will be also redirected to `/final-subcategory/` and cause a page break.

Directory Change

In case you did a category restructuring and want to move everything under the old directory to the new one, you can use the rule below.

```
RewriteRule ^old-directory$ /new-directory/ [R=301,NC,L]  
RewriteRule ^old-directory/(.*)$ /new-directory/$1 [R=301,NC,L]
```

I used `$1` in the target to tell the server that it should remember everything in the URL that follows `/old-directory/` (i.e., `/old-directory/subdirectory/`) and pass it (i.e., `/subdirectory/`) onto the destination. As a result, it will be redirected to `/new-directory/subdirectory/`.

I used two rules: one case with no trailing slash at the end and the other one with a trailing slash.

I could combine them into one rule using `(/?.*)$` RegExp at the end, but it would cause problems and add `“//”` slash to the end of URL when the requested URL with no trailing slash has a query string (i.e., `“/old-directory?utm_source=facebook”` would be redirected to `“/new-directory//?utm_source=facebook”`).

Remove a Word from URL

Let's say you have 100 URLs in your website with city name “chicago” and want to remove it.

Example, for the URL `http://yourwebiste.com/example-chicago-event/`, the redirect rule would be:

```
RewriteRule ^(.*)-chicago-(.*) http://%{SERVER_NAME}/$1-$2  
[NC,R=301,L]
```

If the example URL is in the form `http:// yourwebiste.com/example/chicago/event/`, then redirect will be:

```
RewriteRule ^(.*)/chicago/(.*) http://%{SERVER_NAME}/$1/$2  
[NC,R=301,L]
```

Canonicalization

Having canonical URLs is the most important part of SEO.

If it is missing, you might endanger your website with duplicate content issues because search engines treat URLs with “www” and “non-www” versions as different pages with the same content.

Therefore, it is mandatory to make sure you run website only with only one version you choose.

If you want to run your website with “www” version, use this rule:

```
RewriteCond %{HTTP_HOST} ^yourwebsite\.com [NC]
RewriteRule ^(.*)$ http://www.yourwebsite.com/$1 [L,R=301]
```

For a “non-www” version:

```
RewriteCond %{HTTP_HOST} ^www\.yourwebsite\.com [NC]
RewriteRule ^(.*)$ http://yourwebsite.com/$1 [L,R=301]
```

Trailing slash is also part of canonicalization since URLs with a slash at the end or without are also treated differently.

```
RewriteCond %{REQUEST_FILENAME} !-f
RewriteRule ^.*[^/])$ /$1/ [L,R=301]
```

This will make sure /example-page is redirected to /example-page/. You may choose to remove the slash instead of adding then you will need the other rule below:

```
RewriteCond %{REQUEST_FILENAME} !-d
RewriteRule ^.*)/$ /$1 [L,R=301]
```

HTTP to HTTPS Redirect

After Google's initiative to encourage website owners to use SSL, **migrating to HTTPS** is one of the commonly used redirects that almost every website has.

The rewrite rule below can be used to force HTTPS on every website.

```
RewriteCond %{HTTP_HOST} ^yourwebsite\.com [NC,OR]  
RewriteCond %{HTTP_HOST} ^www\.yourwebsite\.com [NC]  
RewriteRule ^(.*)$ https://www.yourwebsite.com/$1 [L,R=301,NC]
```

Basically, you can combine www or non-www version redirect into one HTTPS redirect rule using this.

Redirect from Old Domain to New

This is also one of the most used redirects when you decide to do rebranding and you need to change domain. The rule below redirects old-domain.com to new-domain.com

```
RewriteCond %{HTTP_HOST} ^old-domain.com$ [OR]  
RewriteCond %{HTTP_HOST} ^www.old-domain.com$  
RewriteRule (.*)$ http://www.new-domain.com/$1 [R=301,L]
```

It uses two cases: one with "www" version of URLs and another "non-www" because any page for historical reasons may have incoming links to both versions.

Most site owners use WordPress and may not need to use .htaccess file for redirects but use plugin instead. Handling redirects by using plugins may be a little different from what we discussed above and you may need to read their documentation in order to be able to handle RegExp correctly for specific plugin.

From existing ones I would recommend free plugin called Redirection which has many parameters to control redirect rules and many useful docs.

Redirect Bad Practices

1. Redirecting All 404 Broken URLs to the Home Page

This case often happens when you are lazy to investigate all of your 404 URLs and map them to the appropriate landing page.

According to Google, they are still all treated as 404s.



If you have too many pages like this, you should consider creating beautiful 404 pages and engage users to browse further or find something other than what they were looking for by displaying a search option.

It is strongly recommended by Google that redirected page content should be equivalent to the old page. Otherwise, such redirect may be considered as soft 404 and you will lose the rank of that page.

2. Wrong Mobile Page Specific Redirects

If you have different URLs for desktop and mobile websites (i.e., “yoursite.com” for desktop and “m.yoursite.com” for mobile), you should make sure to redirect users to the appropriate page of the mobile version.

Correct: “yoursite.com/sport/” to “m.yoursite.com/sport/”

Wrong: “yoursite.com/sport/” to “m.yoursite.com”

Also, you have to make sure that if one page is a 404 on desktop, it should also be a 404 on mobile.

If you have no mobile version for a page, you can avoid redirecting to mobile version and keep them on the desktop page.

3. Using Meta Refresh

It is possible to do redirect using meta refresh tag like example below:

```
<meta http-equiv="refresh" content="0;url=http://yoursite.com/new-page/" />
```

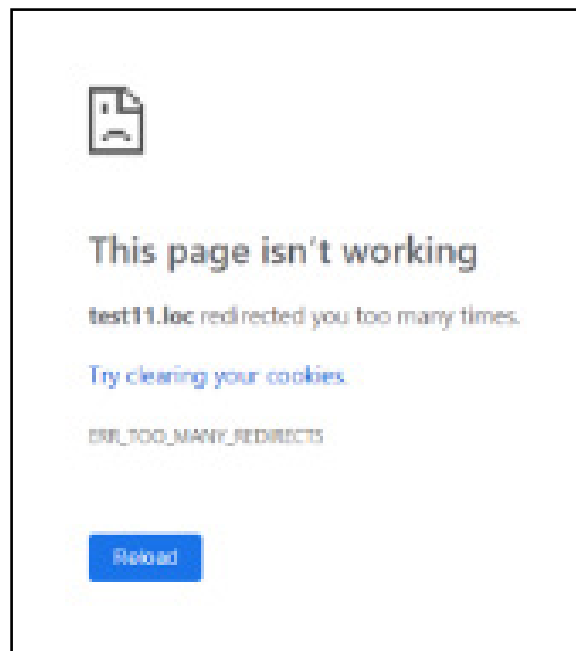
If you insert this tag in /old-page/ it will redirect the user immediately to /new-page/. This redirect is not prohibited by Google but they clearly don't recommend using it.



According to John Mueller, search engines may not be able to recognize that type of redirect properly. The same is also true about JavaScript redirects.

4. Too Many Redirects

This message displays when you have a wrong regular expression setup and it ends up in an infinite loop.



Usually, this happens when you have a redirects chain.

Let's say you redirected page1 to page2 a long time ago. Now you might have forgotten that page1 is redirected and decide to redirect page2 to page1 again.

As a result, you will end up with a rule like this:

```
RewriteRule ^page1 /page2 [R=301,NC,L]  
RewriteRule ^page2 /page1 [R=301,NC,L]
```

This will create an infinite loop and produce the error shown above.



Chapter 15

SEO-Friendly Pagination: A Complete Best Practices Guide

SEJ
EBOOK

Written By
Jes Scholz

International Digital Director, Ringier



Site pagination is a wily shapeshifter.

It's used in contexts ranging from displaying items on category pages, to article archives, to gallery slideshows and forum threads.

For SEO professionals, it isn't a question of if you'll have to deal with pagination, it's a question of when.

At a certain point of growth, websites need to split content across a series of component pages for **user experience** (UX).

Our job is to help search engines crawl and understand the relationship between these URLs so they index the most relevant page.

Over time, the SEO best practices of pagination handling have evolved. Along the way, many myths have presented themselves as facts. But no longer.

This article will:

- **[Debunk the myths around how pagination hurts SEO.](#)**
- **[Present the optimal way to manage pagination.](#)**
- **[Review misunderstood or subpar methods of pagination handling.](#)**
- **[Investigate how to track the KPI impact of pagination.](#)**

But, before I dig into these details. It's important to note that pagination isn't for ranking purposes, but it still has value.

How Pagination Can Hurt SEO

You've probably read that pagination is bad for SEO.

However, in most cases, this is due to a lack of correct pagination handling, rather than the existence of pagination itself.

Let's look at the supposed evils of pagination and how to overcome the SEO issues it could cause.

Pagination Causes Duplicate Content

Correct if pagination has been improperly implemented, such as having both a "View All" page and paginated pages without a correct `rel=canonical` or if you have created a `page=1` in addition to your root page.

Incorrect when you have SEO friendly pagination. Even if your H1 and meta tags are the same, the actual page content differs. So it's not duplication.



Joost de Valk
@jdevalk

@JohnMu do you agree that people can safely ignore the duplicate meta description warning in Google Search Console for paginated archive URLs?

John
@JohnMu



Yep, that's fine. It's useful to get feedback on duplicate titles & descriptions if you accidentally use them on totally separate pages, but for paginated series, it's kinda normal & expected to use the same.

Pagination Creates Thin Content

Correct if you have split an article or photo gallery across multiple pages (in order to drive ad revenue by increasing pageviews), leaving too little content on each page.

Incorrect when you put the desires of the user to easily consume your content above that of banner ad revenues or artificially inflated pageviews. Put a **UX-friendly amount of content** on each page.

Pagination Dilutes Ranking Signals

Correct if pagination isn't handled well as it can cause internal link equity and other ranking signals, such as backlinks and social shares, to be split across pages.

Incorrect when rel="prev" and rel="next" link attributes are used on paginated pages, so that Google knows to consolidate the ranking signals.

Pagination Uses Crawl Budget

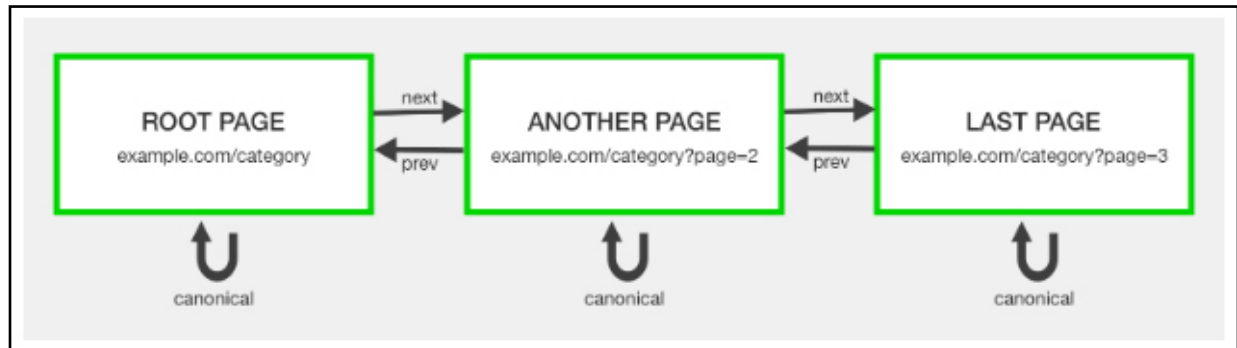
Correct if you're allowing Google to crawl paginated pages. And there are some instances where you would want to use that budget.

For example, for Googlebot to travel through paginated URLs to consolidate ranking signals and to reach deeper content pages.

Often incorrect when you set **Google Search Console** pagination parameter handling to "Do not crawl" or set a robots.txt disallow, in the case where you wish to conserve your crawl budget for more important pages.

Managing Pagination According to SEO Best Practices

Use rel="next" & rel="prev" Link Attributes



You should indicate the relationship between component URLs in a paginated series with rel="next" and rel="prev" attributes.

Google recommends this option, noting they take this markup as "a strong hint" that you would like the pages to be treated "as a logical sequence, thus consolidating their linking properties and usually sending searchers to the first page."

Practically, this means rel="next" / "prev" are treated as signals rather than directives. They won't always prevent paginated pages from being displayed in search results. But such an occurrence would be rare.

Complement the rel="next" / "prev" with a self-referencing rel="canonical" link. So /category?page=4 should rel="canonical" to /category?page=4.

This is the recommended approach by Google, as pagination changes the page content and so is the master copy of that page. If the URL has additional parameters, include these in the rel="prev" / "next" links, but don't include them in the rel="canonical".

For example:

```
<link rel="next" href="https://www.example.com/  
category?page=2&order=newest" />  
<link rel="canonical" href="https://www.example.com/  
category?page=2" />
```

Doing so will indicate a clear relationship between the pages, without sending ranking signals to non-SEO relevant parameter-based URLs and preventing the potential of duplicate content.

Common errors to avoid:

- Placing the link attributes in the <body> content. They're only supported by search engines within the <head> section of your HTML.
- Adding a rel="prev" link to the first page (a.k.a. the root page) in the series or a rel="next" link to the last. For all other pages in the chain, both link attributes should be present.
- Beware of your root page canonical URL. Chances are on ?page=2, rel=prev should link to the canonical, not a ?page=1.

The <head> code of a four-page series will look something like this:

- One pagination tag on the root page, pointing to the next page in series.
 - `<link rel="next" href="https://www.example.com/category?page=2">`
 - `<link rel="canonical" href="https://www.example.com/category">`
- Two pagination tags on page 2.
 - `<link rel="prev" href="https://www.example.com/category">`
 - `<link rel="next" href="https://www.example.com/category?page=3">`
 - `<link rel="canonical" href="https://www.example.com/category?page=2">`
- Two pagination tags on page 3.
 - `<link rel="prev" href="https://www.example.com/category?page=2">`
 - `<link rel="next" href="https://www.example.com/category?page=4">`
 - `<link rel="canonical" href="https://www.example.com/category?page=3">`
- One pagination tag on page 4, the last page in the paginated series.
 - `<link rel="prev" href="https://www.example.com/category?page=3">`
 - `<link rel="canonical" href="https://www.example.com/category?page=4">`

Modify Paginated Pages Titles & Meta Descriptions

Although the `rel="next"` and `rel="prev"` attributes should, in most cases, cause Google to return the root page in the SERPs, you can further encourage this and prevent “Duplicate meta descriptions” or “Duplicate title tags” warnings in Google Search Console with an easy modification to your code.

If the root page has the formula:

BMW Cars for Sale in London | Brand Name
<https://www.domain.cctld/cars/bmw/london> ▼
Find the latest cars for sale by owner or from a trusted dealer in London. Compare prices, features & photos. Contact sellers today.

The successive paginated pages could have the formula:

Result Page 2 for BMW Cars for Sale in London | Brand Name
<https://www.domain.cctld/cars/bmw/london?page=2> ▼
21 - 40 (out of 524) cars for sale by owner or from a trusted dealer in London. Compare prices, features & photos. Contact sellers today.

These paginated URL page titles and meta description are purposefully suboptimal to dissuade Google from displaying these results, rather than the root page.

Don't Include Paginated Pages in XML Sitemaps

While `rel="next"` / `rel="prev"` pagination URLs are technically indexable, they aren't an SEO priority to spend crawl budget on.

As such, they don't belong in your XML sitemap.

Handle Pagination Parameters in Google Search Console

If you have a choice, run pagination via a parameter rather than a static URL. For example:

example.com/category?page=2 over example.com/category/page-2

You can then configure the parameter in Google Search Console to “Paginates” and at any time change the signal to Google to crawl “Every URL” or “No URLs”, based on how you wish to use your crawl budget. No developer needed!



Misunderstood, Outdated or Plain Wrong SEO Solutions to Paginated Content

Do Nothing

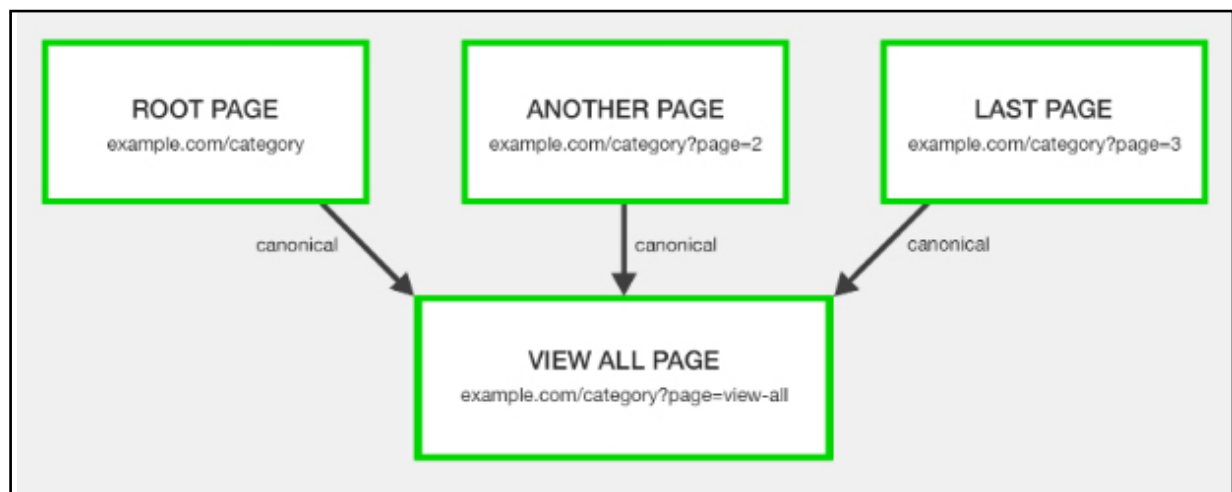


Google says they do “a good job returning the most relevant results to users, regardless of whether content is divided into multiple pages” and recommends you can handle pagination by doing nothing.

While there is a core of truth to this statement, by doing nothing you’re gambling with your SEO.

There’s always value in giving clear guidance to crawlers how you want them to index and display your content.

Canonicalize to a View All Page



The last option recommended by Google is a View All page. This version should contain all the component page content on a single URL.

Additionally, the paginated pages should all rel="canonical" to the View All page to consolidate ranking signals.

The argument here is that searchers prefer to view a whole article or list of categories items on a single page, as long as it's fast loading and easy to navigate.

So if your paginated series has an alternative View All version that offers the better user experience, Google will favor this page for inclusion in the search results as opposed to a relevant segment page of the pagination chain.

Which raises the question – why do you have paginated pages in the first place?

Let's make this simple.

If you can provide your content on a single URL while offering a good user experience, there is no need for pagination or a View All version.

If you can't, for example, a category page with thousands of products would be ridiculously large and take too long to load, then paginate with rel="next" / "prev". View All is not the best option as it would not offer a good user experience.

Using both rel="next" / "prev" and a View All version gives no clear mandate to Google and will result in confused crawlers.

Don't do it.

Canonicalize to the First Page



A common mistake is to point the rel="canonical" from all paginated results to the root page of the series.

Some ill-informed SEO people suggest this as a way to consolidate authority across the set of pages to the root page, but this is unnecessary when you have rel="next" and rel="prev" attributes.

Incorrect canonicalization to the root page runs the risk of misdirecting search engines into thinking you have only a single page of results.

Googlebot then won't index pages that appear further along the chain, nor acknowledge the signals to the content linked from those pages.

You don't want your detailed content pages dropping out of the index because of poor pagination handling.

Google is clear on the requirement. Each page within a paginated series should have a self-referencing canonical, unless you use a View All page.

Use the rel=canonical incorrectly and chances are Googlebot will just ignore your signal.

Noindex Paginated Pages



A classic method to solve pagination issues was a robots noindex tag to prevent paginated content from being indexed by search engines.

Relying solely on the noindex tag for pagination handling will result in ranking signals from your component pages not being consolidated. Clearly inferior SEO to using rel="next" / "prev".

But as the rel="next" / "prev" method allows search engines to index pagination pages, I've also seen some SEO folks advising to add "extra security" with a noindex tag.

This is unnecessary. Only in rare circumstances would Google choose to return a paginated page in the SERPs. The benefits are, at best, theoretical.

But what you may not be aware of is that a long-term noindex on a page will eventually lead **Google to nofollow the links** on that page. So, again, it could potentially cause content linked from the paginated pages to be removed from the index.

Pagination & Infinite Scrolling



A newer form of pagination handling is by infinite scroll, where content is pre-fetched and added directly to the user's current page as they scroll down.

Users may appreciate this, but Googlebot? Not so much.

Googlebot doesn't emulate behavior like scrolling to the bottom of a page or clicking to load more. Meaning without help, search engines can't effectively crawl all of your content.

To be SEO-friendly, convert your infinite scroll page to an equivalent paginated series that is accessible even with JavaScript disabled.

As the user scrolls, use JavaScript to adapt the URL in the address bar to the component paginated page.

Additionally, implement a `pushState` for any user action that resembles a click or actively turning a page. You can check out this functionality in the [demo created by John Mueller](#).

Essentially, you're still implementing the SEO best practice recommended above, you are just adding additional user experience functionality on top.

Discourage or Block Pagination Crawling



Some SEO pros recommend avoiding the issue of pagination handling altogether by simply blocking Google from crawling paginated URLs.

In such a case, you would want to have well-optimized XML sitemaps to ensure pages linked via pagination have a chance to be indexed.

There are three ways to do this:

- The messy way: Add nofollow to all links that point towards paginated pages.
- The cleaner way: Use a robots.txt disallow.
- The no dev needed way: Set paginated page parameter to “Paginates” and for Google to crawl “No URLs” in Google Search Console.

By using one of these methods to discourage search engines from crawling paginated URLs you:

- Stop search engines from consolidating ranking signals of paginated pages.
- Prevent the passing of internal link equity from paginated pages down to the destination content pages.
- Hinder Google's ability to discover your destination content pages. The obvious upside is that you save on crawl budget.

There is no clear right or wrong here. You need to decide what is the priority for your website.

Personally, if I were to prioritize crawl budget, I would do so by using pagination handling in Google Search Console as it has the optimum flexibility to change your mind.

Tracking the KPI Impact of Pagination

So now you know what to do, how do you track the effect of optimization pagination handling?

Firstly, gather benchmark data to understand how your current pagination handling is impacting SEO.

Sources for KPIs can include:

- Server log files for the number of paginated page crawls.
- Site: search operator (for example `site:example.com inurl:page`) to understand how many paginated pages Google has indexed.
- Google Search Console Search Analytics Report filtered by pages containing pagination to understand the number of impressions.
- Google Analytics landing page report filtered by paginated URLs to understand on-site behavior.

If you see an issue getting search engines to crawl your site pagination to reach your content, you may want to [change the pagination links.](#)

Once you have launched your best practice pagination handling, revisit these data sources to measure the success of your efforts.

Chapter 16

What is Schema Markup & Why It's Important for SEO

SEJ
EBOOK

Written By
Chuck Price
Founder, Measurable SEO



What is Schema Markup?

Schema markup, found at [Schema.org](https://schema.org), is a form of microdata. Once added to a webpage, schema markup creates an enhanced description (commonly known as a rich snippet), which appears in search results.

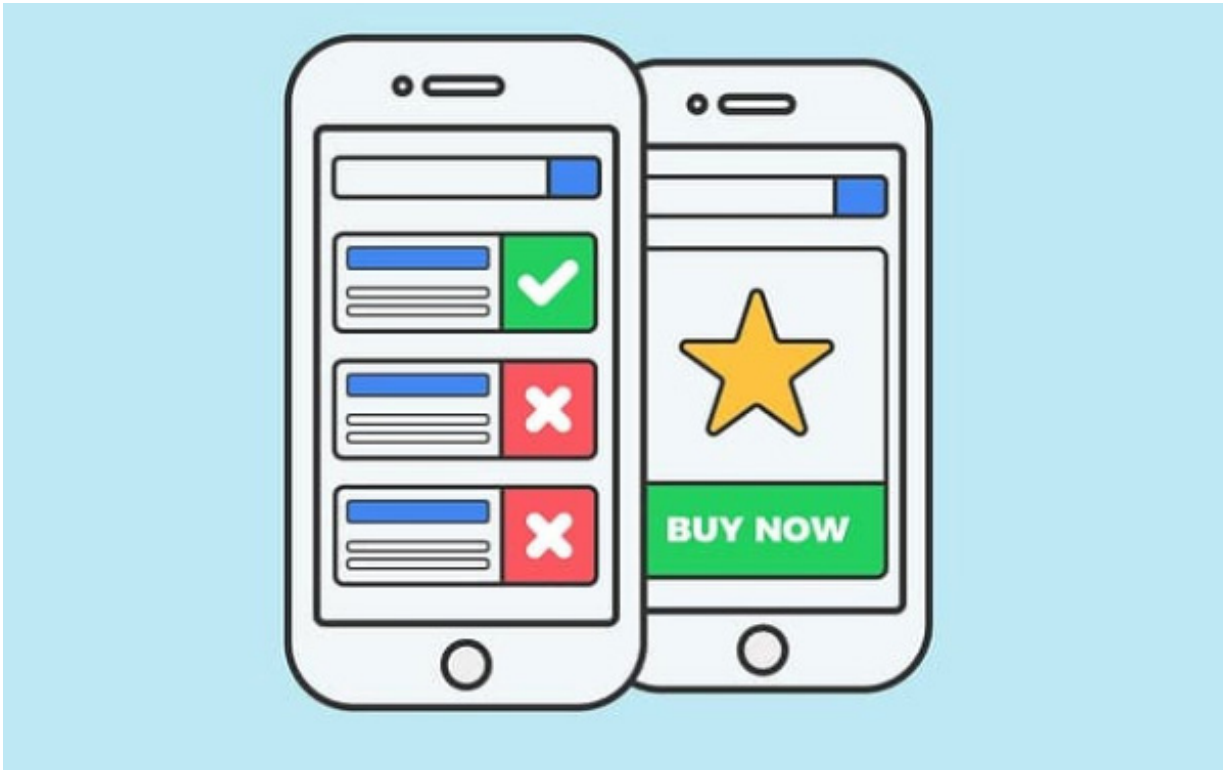
Top search engines – including Google, Yahoo, Bing, and Yandex – first started collaborating to create Schema.org, back in 2011.

Schema markup is especially important in the age of [Hummingbird](#) and [RankBrain](#). How a search engine interprets the context of a query will determine the quality of a search result.

Schema can provide context to an otherwise ambiguous webpage.

Via [Schema.org](https://schema.org):

“Most webmasters are familiar with HTML tags on their pages. Usually, HTML tags tell the browser how to display the information included in the tag. For example, `<h1>Avatar</h1>` tells the browser to display the text string “Avatar” in a heading 1 format. However, the HTML tag doesn’t give any information about what that text string means—“Avatar” could refer to the hugely successful 3D movie, or it could refer to a type of profile picture—and this can make it more difficult for search engines to intelligently display relevant content to a user.”



Does Schema Improve Your Search Rankings?

There is no evidence that microdata has a direct affect on organic search rankings.

Nonetheless, rich snippets do make your webpages appear more prominently in SERPs. This improved visibility has been shown to **improve click-through rates.**

According to a **study** by acmqe, less than one-third of Google's search results include a rich snippet with Schema.org markup. This exposes a huge opportunity for the rest. Very few things in SEO, today, can move the dial quickly. This can.

What Is Schema Used For?

- [Businesses and organizations](#)
- [Events](#)
- [People](#)
- [Products](#)
- [Recipes](#)
- [Reviews](#)
- [Videos](#)

Above are some of the most popular uses of schema. However, there's a good chance that if you have any sort of data on your website, it's going to have an associated [itemscope](#), [itemtype](#) and [itemprop](#).

FIGURE 4A: MAJOR SITES THAT HAVE PUBLISHED SCHEMA.ORG

CATEGORY	SITES
News	nytimes.com, guardian.com, bbc.co.uk
Movies	imdb.com, rottentomatoes.com, movies.com
Jobs / Careers	careerjet.com, monster.com, indeed.com
People	linkedin.com, pinterest.com, familysearch.org, archives.com
Products	ebay.com, alibaba.com, sears.com, cafepress.com, sulit.com, fotolia.com
Video	youtube.com, dailymotion.com, frequency.com, vinebox.com
Medical	cvs.com, drugs.com
Local	yelp.com, allmenus.com, urbanspoon.com
Events	wherevent.com, meetup.com, zillow.com, eventful.com
Music	last.fm, myspace.com, soundcloud.com

Adding Schema to Your Webpages

Using Microdata

Microdata is a set of tags that aims to make annotating HTML elements with machine-readable tags much easier. Microdata is a great place for beginners to start because it's so easy to use.

However, the one downside to using microdata is that you have to mark every individual item within the body of your webpage. As you can imagine, this can quickly get messy.

Before you begin to add schema to your webpages, you need to figure out the 'item type' of the content on your webpage.

For example, does your web content focus on food? Music? Tech? Once you've figured out the item type, you can now determine how you can tag it up.

Let's look at an example. Let's say that you own a store that sells high-quality routers. If you were to look at the source code of your homepage you would likely see something akin to this:

```
<div>
```

```
<h1>TechHaven</h1>
```

```
<h2>The best routers you'll find online!</h2>
```

```
<p>Address:</p>
```

```
<p>459 Humpback Road</p>
```

```
<p>Rialto, Ca</p>
```

```
<p>Tel: 909 574 3903</p>
```

```
<p><a href="http://www.techhaven.com/menu">Click here to  
view our best routers!</a></p>  
<p>We're open: </p>  
<p>Mon-Sat 8am – 10:30pm</p>  
<p>Sun: 2pm – 8pm</p>  
</div>
```

Once you dive into the code, you'll want to find the portion of your webpage that talks about what your business has to offer. In this example, that data can be found between the two `<div>` tags.

Now, at the top you can add in:

```
<div itemscope>
```

By adding this tag, we're stating that the HTML code contained between the `<div>` blocks are identifying a specific item.

Next, we have to identify what that item is by using the 'itemtype' attribute to identify the type of item our webpage is about (tech).

```
<div itemscope itemtype="http://schema.org/tech">
```

An item type comes in the form of a URL (such as `http://schema.org/tech`). Let's say, for example, that your site revolved around beauty products instead of technology. Your item type URL might look like this:

```
http://schema.org/beauty.
```

To make things easier you can browse a [list of item types](#) here, plus you can view extensions to identify the specific entity that you're looking for. Keep in mind that this list is not all encompassing,

so there is a possibility that you won't find the item type for your specific niche.

Tracking back to the tech page, you want to tag the part of the webpage that contains the name of the business. You can do this between the <h1> tags.

Now, we'll be using the 'itemprop' tag, which labels the properties of an item:

```
<h1 itemprop="name">Tech Haven</h1>
```

You can apply these tags to the rest of the page now. When using tags to identify item properties, it's not necessary to tag the entire line, just the one portion the property is making reference to.

For example, if you have a line that says Address: 1234 w sunshine blvd, then you only need to apply the tags around the address itself and nothing else.

```
<h2 itemprop="description">The best routers you'll find online!</h2>
```

```
<p>Address:</p>
```

```
<span itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
```

```
<p itemprop="streetAddress">459 Humpback Road </p>
```

```
<p itemprop="addressLocality">Rialto, Ca</p></span>
```

```
<p>Tel: <span itemprop="telephone">909 574 3903</span></p>
```

```
<p><a itemprop="menu" href="http:// http://www.techhaven.com/menu ">Click here to view our tasty range of dishes!</a></p>
```

```
<p>We're open:</p>
```

```
<p itemprop="openingHours">Mon-Sat 8am – 10:30pm</p>
```



```
<p itemprop="openingHours">Sun: 2pm – 8pm</p>  
</div>
```

This code may look complicated, but schema.org provides examples on how to use the different item types, so you can actually see what the code is supposed to do. Don't worry, you won't be left out in the cold trying to figure this out on your own!

If you're still feeling a little intimidated by the code, [Google's Structured Data Markup Helper](#) makes it super easy to tag your webpages.

To use this amazing tool, just select your item type, paste in the URL of the target page or the content you want to target, and then highlight the different elements so that you can tag them.

Using RDFa

RDFa is an acronym for Resource Description Framework in Attributes. Essentially, RDFa is an extension to HTML5 and it was designed to aid users in marking up structured data.

RDFa is considered to be a W3C recommendation, meaning that it is a web standard, and it can be used to chain structured data vocabularies together. This is especially useful if you want to add structured data that stretches beyond the limits of Schema.org.

You can breathe a sigh of relief. RDFa isn't much different from Microdata.

Similar to microdata, RDFa tags incorporate with the preexisting HTML code in the body of your webpage. For the sake of familiarity, we'll look at the tech website once again as an example.

The HTML for your tech site would likely look like this before it was modified:

```
<div>
<h1>Tech Haven</h1>
<h2>The best routers online!</h2>
<p>Address:</p>
<p>459 Humpback Road </p>
<p>Rialto, Ca</p>
<p>Tel: 909 574 3903</p>
<p><a href="http://www.techhaven.com/menu">Click here
to view our best routers!</a></p>
<p>We're open:</p>
<p>Mon-Sat 8:00am – 10:30pm</p>
<p>Sun: 2pm – 8pm</p>
</div>
```

To begin, you want to ensure that the vocabulary that you're using is Schema.org and that the webpage in question is making reference to a technology page.

For this example, you can search for "technology" on Schema.org to learn how to tag different elements. Typically, you'll find examples near the bottom of the page that will show you how to use them in practice.

Simply click on the RDFa tab to view specific RDFa examples. Next, you need to use the vocab tag combined with the URL `http://schema.org` to identify the vocabulary for the markup. To identify the page type, use the `typeof` tag. Unlike microdata, which uses a URL to identify types, RDFa uses one or more words to classify types.

```
<div vocab="http://schema.org/" typeof="technology">
```

If you wish to identify a property further than you should use the typeof attribute.

For example, if you wish to further expand upon an address property you can use "PostalAddress" like so:

```
<div property="address" typeof="PostalAddress">
```

Comparing microdata and RDFa side by side, the typeof attribute is the equivalent of the itemtype attribute found in Microdata. Furthermore, the propertyattribute would be the equivalent to the itemprop attribute.

For further explanation, you can visit Schema.org to check lists and view examples. You can find which kinds of elements are defined as properties, and which are defined as types.

Going back to our earlier example, the tech page would look like this after it has been appropriately tagged:

```
<h2 property="description">The best routers on the internet!</h2>
<p>Address:</p>
<div property="address" typeof="PostalAddress">
<p property="streetAddress">459 Humpback Road</p>
<p property="addressLocality">Rialto, Ca</p>
</div>
<p>Tel: <span property="telephone">909 574 3903</span></p>
<p><a property="menu" href="http://www.techhaven/menu">Click here to view our best routers!</a></p>
<p>We're open:</p>
<p property="openingHours">Mon-Sat 8am – 10:30pm</p>
<p property="openingHours">Sun: 2pm – 8pm</p>
</div>
```

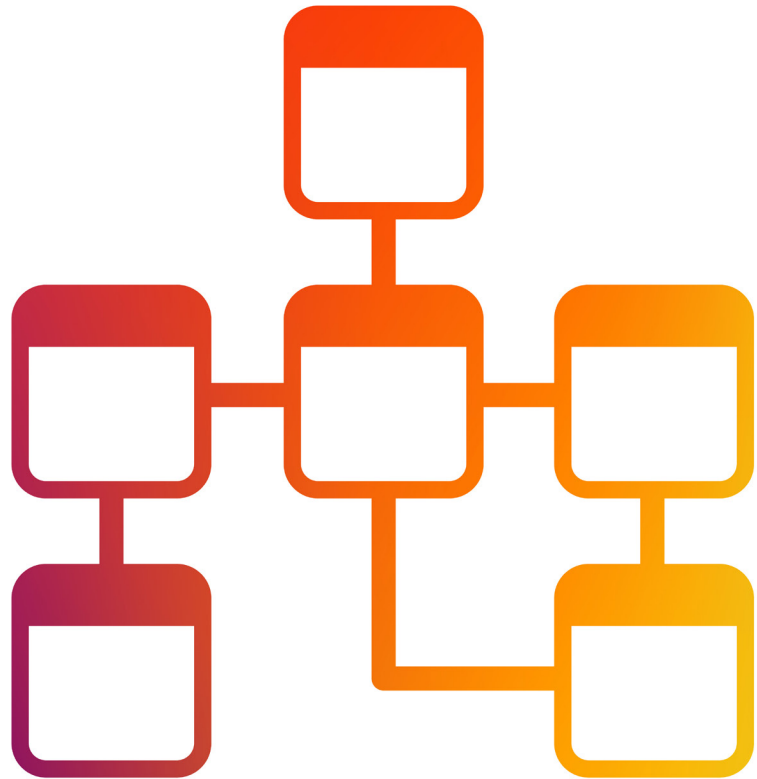
To aid you, every page on Schema.org provides examples on how to properly apply tags. Of course, you can also fall back on Google's [Structured Data Testing Tool](#).

Conclusion

Hopefully, any fears that you might have had when you heard the word "Schema" or "Structured Data" have been put to rest.

Schema is much easier to apply than it seems and it's a best practice that you need to incorporate into your webpages.

The work may seem tedious. But given time and effort, you'll be certain to reap the benefits from your labor.



Chapter 17

Faceted Navigation: Best Practices for SEO

SEJ
EBOOK

Written By
Natalie Hoben
Digital Marketing Specialist, Forthea
Interactive



When it comes to large websites, such as e-commerce sites with thousands upon thousands of page, the importance of things like crawl budget cannot be understated.



Building a website with an **organized architecture** and smart **internal linking strategy** is key for these types of sites.

However, doing that properly oftentimes involves new challenges when trying to accommodate various attributes that are a common theme with e-commerce (sizes, colors, price ranges, etc.).

Faceted navigation can help solve these challenges on large websites.

However, faceted navigation must be well thought out and executed properly so that both users and search engine bots remain happy.

What is Faceted Navigation?

To begin, let's dive into what faceted navigation actually is.

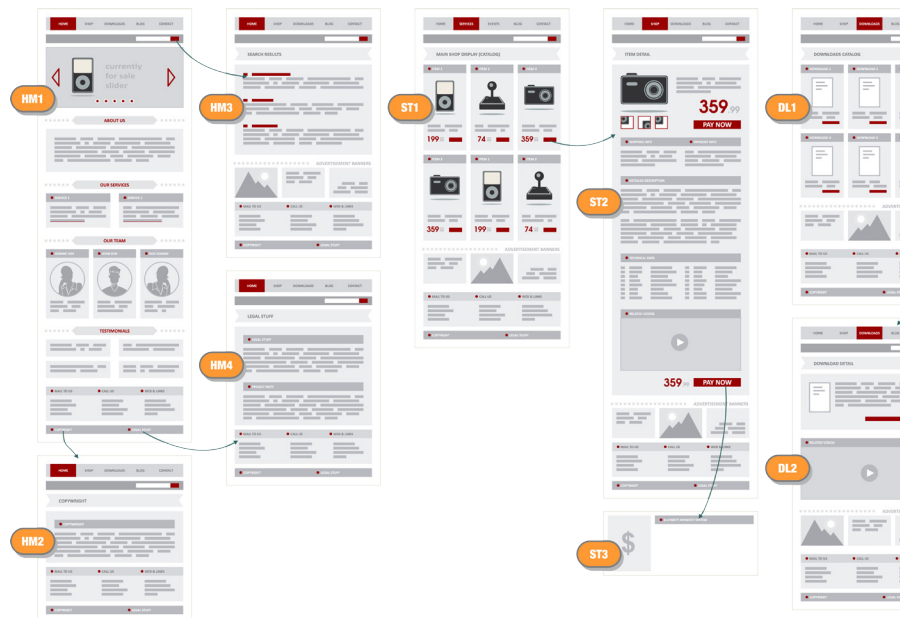
Faceted navigation is, in most cases, located on the sidebars of an e-commerce website and has multiple categories, files, and facets.

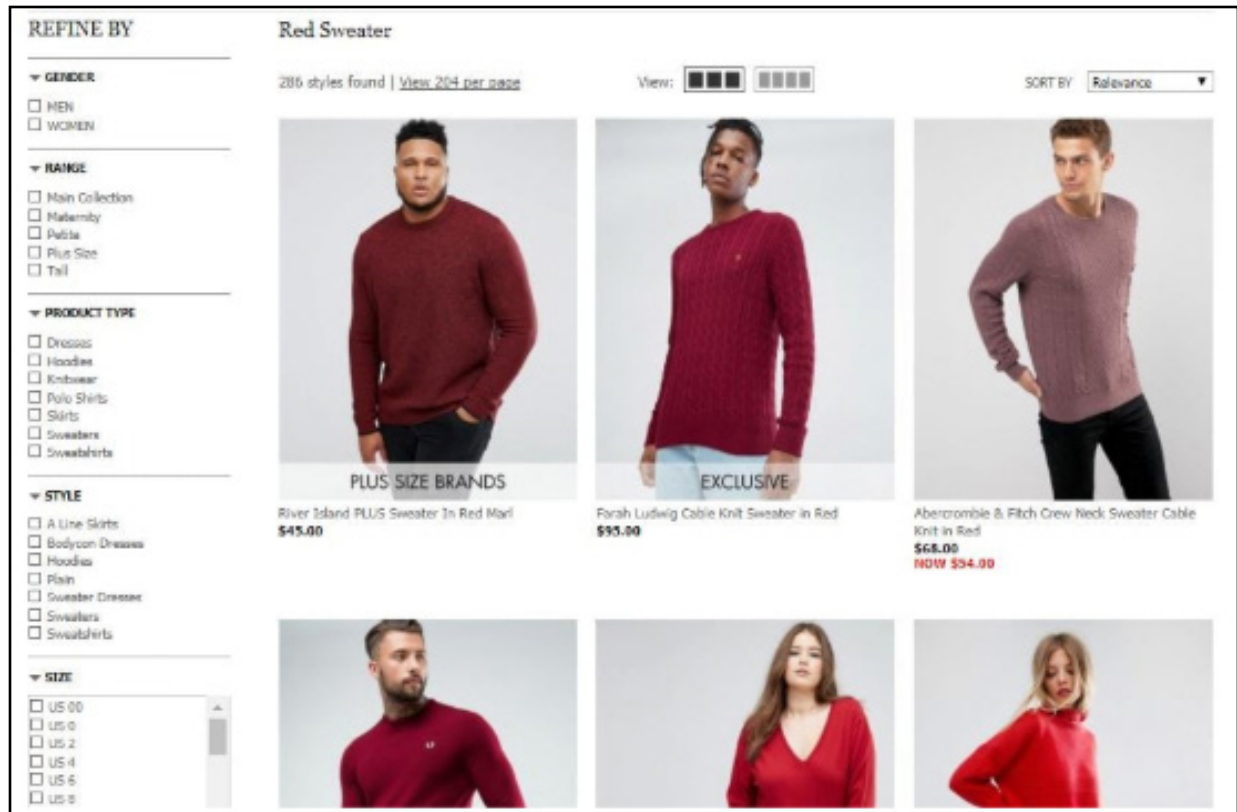
It essentially allows people to customize their search based on what they are looking for on the site.

For example, a visitor may want a purple cardigan, in a size medium, with black trim.

Facets are indexed categories that help to narrow down a production listing and also function as an extension of a site's main categories.

Facets, in their best form, should ideally provide a unique value for each selection and, as they are indexed, each one on a site should send relevancy signals to search engines by making sure that all critical attributes appear within the content of the page.





Filters are utilized to sort items with a listings page.

While the user can use this to narrow down what they are looking for, the actual content on the page remains the same.

This can potentially lead to multiple URLs creating duplicate content, which is a concern for SEO.

There are a few potential issues that faceted navigation can create that can negatively affect SEO.

The main three issues boil down to:

- Duplicate content.
- Wasted crawl budget.
- Diluted link equity.

As different parameters are created, they can quickly multiply. The number of incredibly-related pieces of content continues to grow significantly and different links may be going to all of these different versions of a page, which can dilute link equity, and thus affect the page's ranking ability.

In order to be able to make sure that search engine crawlers aren't wasting valuable crawl budget on pages that have little to no value, you need to take certain steps.

That starts with preventing search engine bots from crawling certain multi-selected facets, such as "color" or "size".

When trying to determine how to solve this faceted navigation conundrum, there are a few solutions that are implementable. Which one to use, however, will rely heavily on what parts of the site should be indexed.

Noindex

Noindex tags can be implemented to inform bots of which pages not to include in the index. This method will remove pages of the index, however, there will still be crawl budget spent on them and link equity that is diluted.

For example, if you wished to include a page for “red sweaters” in the index, but did not want “red sweaters under \$50” in the index, then a `noindex` tag to the second result would exclude it.

Bots would still be able to find and crawl the page, though, and this causes crawl budget to be wasted.

The pages would also still get wasted link equity.

Robots.txt

A disallow can be implemented for certain sections of a site.

The advantage to this solution is that it’s fast and customizable.

However, the disallow is merely a directive for Google, and they do not have to abide by it.

In addition, link equity may be hindered from flowing to different parts of the site.

```
User-agent: *  
Disallow: /*noindex=1
```

For example, we could disallow red sweaters under \$50 in the robots file, instructing Google to not visit a page with the `>$50` parameter. However, if any follow links pointing to any URL with that parameter in it existed, Google could still possible index it.

Canonicalization

Canonical tags allow you to instruct Google that a group of alike pages has a preferred version.

Link equity can be consolidated into the chosen preferred page utilizing this method. However, crawl budget will still be wasted.

Canonical tags can also be ignored by search engines, so this solution should be used along with another.

For example, `/red-sweaters?under-50/` could have the canonical URL set as `/red-sweaters/`. Google would attribute the authority and link equity to the canonicalized page, but crawl budget would still be wasted.

AJAX

When it comes to using AJAX to solve faceted navigation issues, the main positive benefit is that a different, new URL is not generated when a visitor visits a page and selects a filter.

JavaScript hosted client-side takes care of the entire process. No web server is needed.

In order to ensure that this method is effective, it is necessary that a crawl path is accessible to the particular pages that are important to get into rankings.

The `pushState` method of the HTML5 history API and server configuration that responds to these requests with HTML rendered server-side can help ensure that AJAX can fully work its magic and keep SEO in a healthy state.

Google Search Console

This should ideally be a last resort option.

It is a decent temporary solution while adjustments are being made to the navigation. This is because it only instructs Google on how a site should be crawled, instead of correcting the issue.

By navigating to the URL parameters tool in [Google Search Console](#), you can choose what effect each parameter has on the page and how Google should treat those pages.

Other Ways to Get the Most out of Faceted Navigation

- Implement pagination with rel="next" and rel="prev" in order to group indexing properties, from pages to a series as a whole.
- Each page needs to link to children pages and parent. This can be done with breadcrumbs.
- Use rigorous URL facet ordering so that duplication problems do not arise.
- Prevent clicks when no items are present for the filter.
- Only use canonical URLs in sitemaps.
- Facets should always be presented in a unified, logical manner (i.e., alphabetical order).

- Don't rely solely on one "fix" if it doesn't take care of indexing, link dilution, and crawl. For example, noindex and nofollow tags do not help with crawl budget. Same with configuring parameters in Google Search Console.
- If a particular combination of facets occurs that receive a good amount of traffic, consider allowing indexation.

Conclusion

Although faceted navigation can be great for UX, it can cause a multitude of problems for SEO.

Duplicate content, wasted crawl budget, and diluted link equity can all cause severe problems on a site.

It is crucial to carefully plan and implement the necessary methods available in order to avoid any many issues down the line when it comes to faceted navigation.

Chapter 18

Understanding JavaScript Fundamentals: Your Cheat Sheet

SEJ
EBOOK

Written By
Rachel Costello
Technical SEO & Content Manager,
DeepCrawl

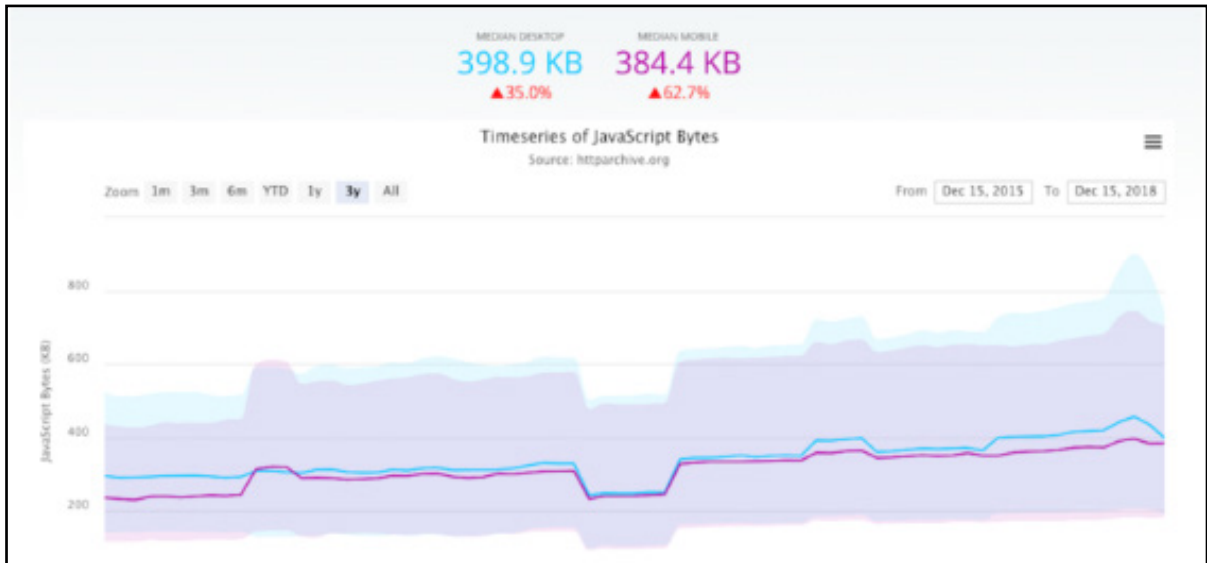


JavaScript is a complex topic that can be extremely difficult to get a handle on.

However, it has never been more important to understand it because JavaScript is becoming increasingly prevalent on the websites that we manage.

As the modern web continues to evolve, **JavaScript usage continues to rise.**

SEO professionals may long for times gone by when websites were static and coded only in HTML and CSS. However, engaging websites often require interactivity, which is usually powered by JavaScript.



The number of JavaScript bytes across the entire web has increased by 35 percent on desktop and 62.7 percent on mobile over the last three years.

As Google Webmaster Trends Analyst John Mueller put it: JavaScript is **“not going away.”**

This programming language is all around us, so we should get better acquainted with it. Let's be proactive and learn more about JavaScript rather than fearing it.



There is often a misconception that JavaScript is solely for developers to worry about.

I would argue that this isn't the case, as it can cause a problem for anyone who wants customers and search engines to be able to access their website's content.

If you aren't completely familiar with JavaScript, or even have absolutely no idea what it is or does, don't worry.

I've put together a glossary of the key terms and fundamental concepts you should know to help you get started on your journey of discovery.

What Is JavaScript?

JavaScript is a programming language that allows you to implement complex features on a website, such as dynamic elements or interactivity.

JavaScript is executed once the information from the HTML and CSS in the source code has been parsed and constructed.

The JavaScript will then trigger any events or variables specified within it, the Document Object Model (DOM) will be updated, and, finally, the JavaScript will be rendered in the browser.

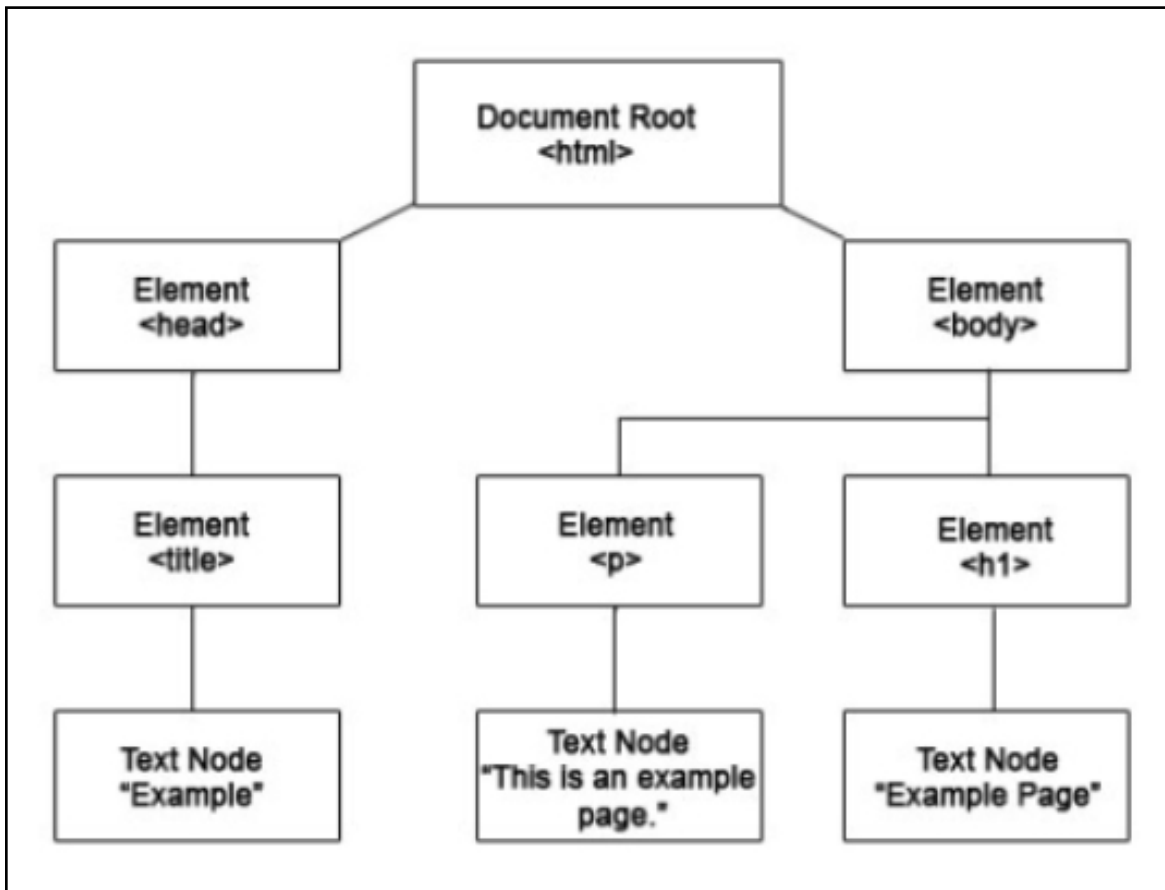
The HTML and CSS will often form the foundations of a page's structure, and any JavaScript will make the final tweaks and alterations.



Document Object Model (DOM)

The **Document Object Model (DOM)** is created when a page is loaded, and it is made up of nodes and objects which map out all of the different elements and attributes on a page.

The page is mapped out in this way so that other programs can modify and manipulate the page in terms of its structure, content, and styling.



Altering the elements of a page's DOM is possible through using a language like JavaScript.

ECMAScript

ECMAScript (ES) is a scripting language that was created to standardize the use of JavaScript code.

Different editions of ECMAScript are released when the language is updated and tweaked over time, such as ES5 and ES6 (which is also referred to as ES2015).

Transpiling

A **transpiler** is a tool that transforms source code into a different programming language. The concept is a bit like Google Translate, but for code.

You can convert a particular source language into a different target language, for example, JavaScript to C++ or Python to Ruby.

With regard to JavaScript rendering particularly, a transpiler is often recommended for transforming ES6 into ES5 because Google currently uses an old version of Chrome for rendering which doesn't yet support ES6.



Chrome 41

When rendering pages, Google uses a [web rendering service](#) which is based on Chrome 41. This means that Google's rendering engine supports the same features and functionalities of that particular version of Chrome.

When you consider that the most up-to-date version is Chrome 71, you can see that many versions have been launched since [Chrome 41 went live in 2015](#), and all of these versions came with new features. This is why Google's rendering service currently supports ES5 rather than the later ES6 version of the language.

Single-page Application (SPA)

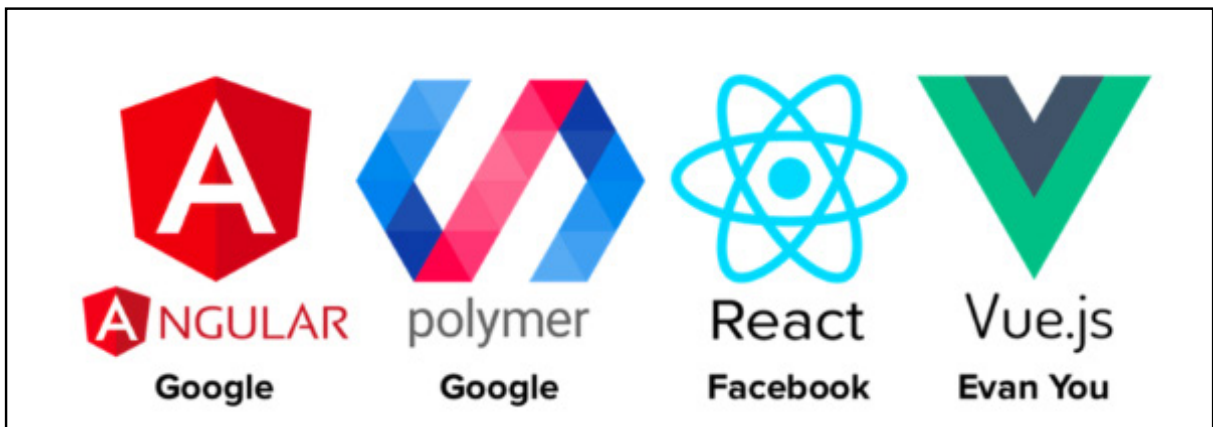
A **single-page application (SPA)** is a website or web app that dynamically re-writes and re-renders a page as a user interacts with it, rather than making separate requests to the server for new HTML and content.

JavaScript frameworks can be used to support the dynamically changing elements of SPAs.

Angular, Polymer, React & Vue

These are all different types of JavaScript frameworks.

- Angular and Polymer were developed by Google.
- React was developed by Facebook.
- Vue was developed by [Evan You](#), who used to work on Google's Angular team.



Each JavaScript framework has its own pros and cons, so developers will choose to work with the one that best suits them and the project they're working on.

If you want to learn more about how the different frameworks measure up, [this guide](#) gives a detailed comparison.

JavaScript Rendering

JavaScript rendering involves taking the script and the instructions it contains, processing it all, then running it so that the required output is shown in the browser. There are many different methods you can use to control the way in which JavaScript is rendered.

Requiring JavaScript to be rendered on a page can negatively impact two key areas:

- Site speed
- Search engine crawling and indexing

Depending on which rendering method you use, you can reduce page load speed and make sure content is accessible to search engines for crawling and indexing.

Pre-rendering

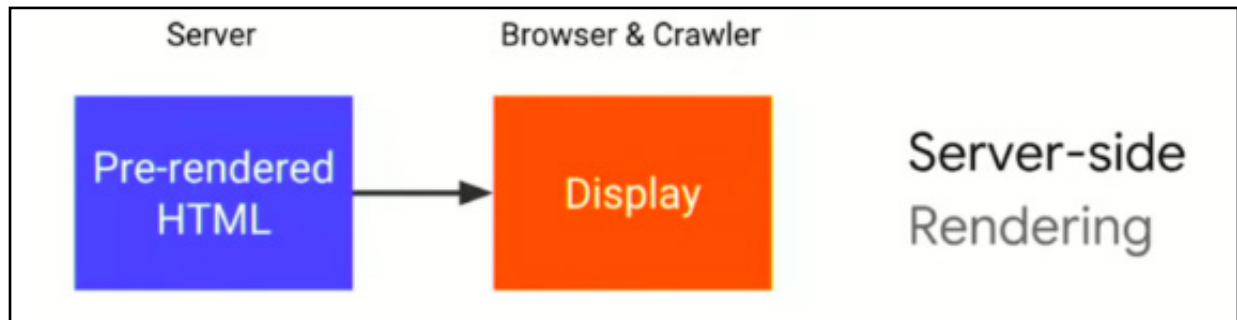
Pre-rendering involves rendering the content on a page before it is requested by the user or search engine, so that they receive a static page with all of the content on there ready to go.

By preloading a page in this way, it means that your content will be accessible rather than a search engine or user's browser having to render the page themselves.

Pre-rendering is usually used for search engine bots rather than humans. This is because a static, pre-rendered page will be less engaging for users as it will lack any dynamic content or interactivity.

Server-side Rendering

The hosting server does the heavy lifting and renders the page so that the JavaScript has already been processed and the content is ready to be handed over to the user's browser or search engine crawler when it is requested.



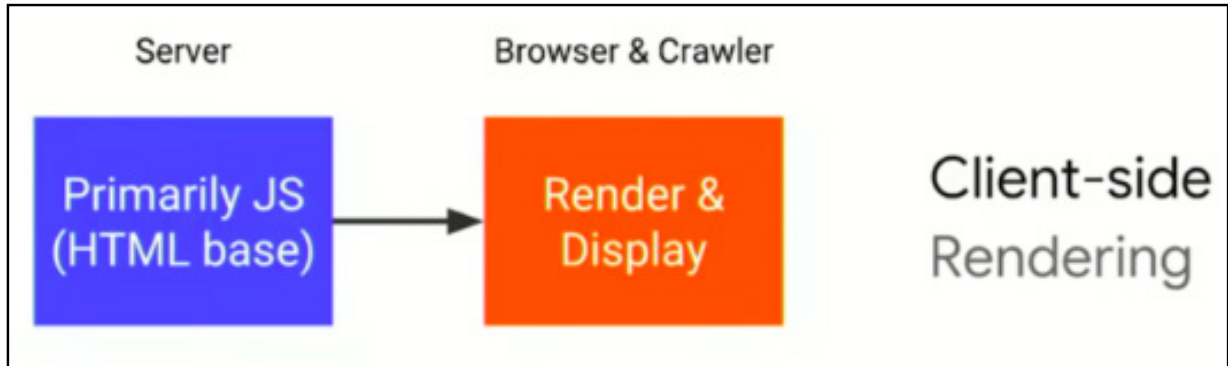
This method helps to reduce any strain on the user's device that would have been caused by processing JavaScript, and this can increase page load speed.

Server-side rendering also ensures the full content can be seen and indexed by search engines.

Client-side Rendering

During client-side rendering, JavaScript is processed by the user's browser or by the search engine that's requesting a page.

The server will handle the initial request, but the rest of the work of processing and rendering a page falls on the user's device or the search engine.



It is often advised against to use client-side rendering as there is a delay between Google crawling pages and then being able to render them.

Google puts pages that need to be rendered into a queue until enough resources become available to process them.

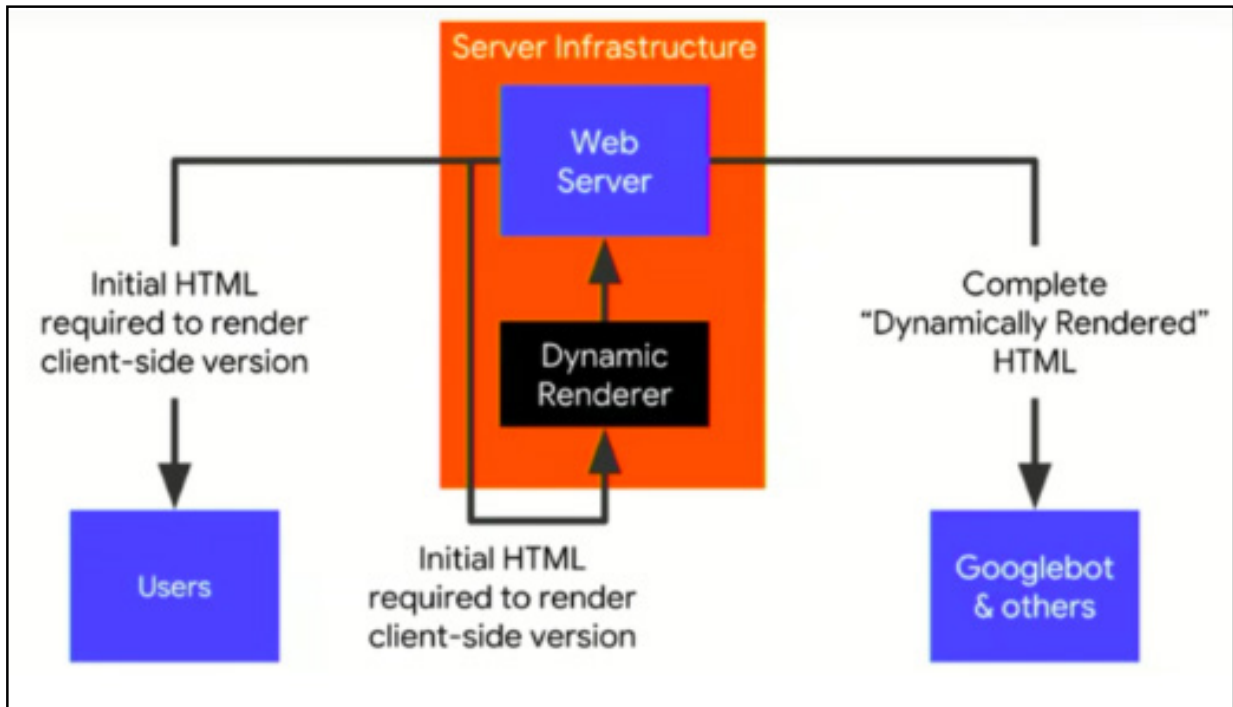
If you're relying on Google to render a page client-side, this can delay indexing by up to a week after it is initially crawled.

Dynamic Rendering

Dynamic rendering involves using different rendering methods depending on whether a user's browser or a search engine bot is requesting a page.



If your site usually renders client-side, when Googlebot is detected the page will be pre-rendered using a mini client-side renderer (for example, [Puppeteer](#) or [Rendertron](#)), so the content can be seen and indexed straight away.

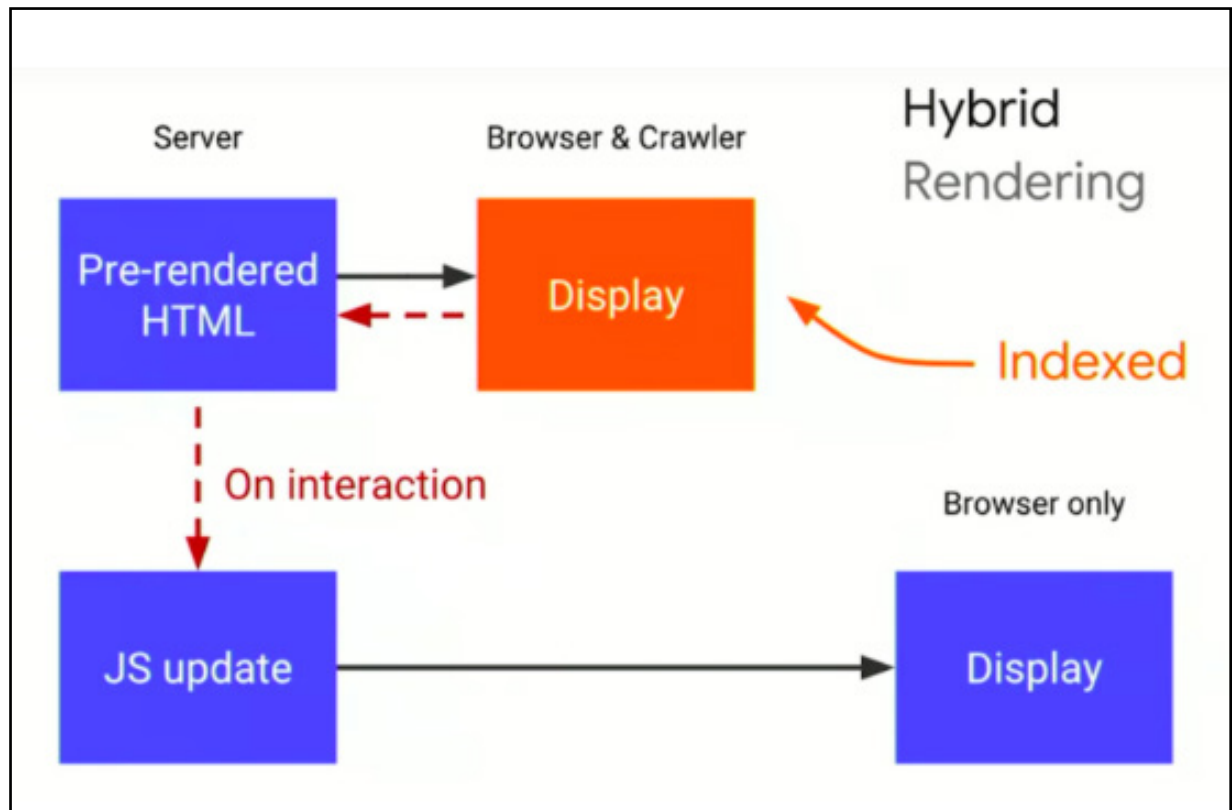


Hybrid Rendering

Hybrid rendering involves a combination of both server-side rendering and client-side rendering.

The core content is pre-rendered server-side and sent to the client, whether that's the user's browser or the search engine crawler that's requesting the content.

After the page is initially loaded, additional JavaScript for any interactivity is then rendered client-side.



Conclusion

Hopefully you found this guide useful, and that it helped you better understand the basics of JavaScript and how it impacts websites.

Now that you've brushed up on the key terms, you should be better equipped to hold your own in conversations with the developers!

Chapter 19

An SEO Guide to URL Parameter Handling

SEJ
EBOOK

Written By
Jes Scholz

International Digital Director, Ringier



While parameters are loved by developers and analytics aficionados, they are often an SEO nightmare.

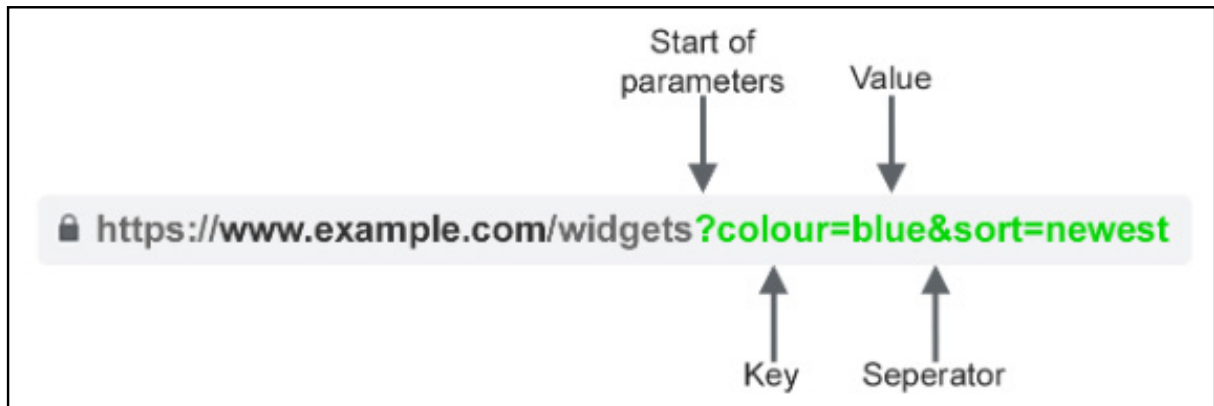
Endless combinations of parameters can create thousands of URL variations out of the same content.

The problem is we can't simply wish parameters away. They play an important role in a website's user experience. So we need to understand how to handle them in an SEO-friendly way.

To do so we explore:

- [The basics of URL parameters](#)
- [SEO issues caused by parameters](#)
- [Assessing the extent of your parameter problem](#)
- [SEO solutions to tame parameter](#)
- [Best practice URL parameter handling](#)

What Are URL Parameters?



Also known by the aliases of query strings or URL variables, parameters are the portion of a URL that follows a question mark. They are comprised of a key and a value pair, separated by an equal sign. Multiple parameters can be added to a single page by using an ampersand.

The most common use cases for parameters are:

- Tracking – For example `?utm_medium=social`, `?sessionid=123` or `?affiliateid=abc`
- Reordering – For example `?sort=lowest-price`, `?order=highest-rated` or `?so=newest`
- Filtering – For example `?type=widget`, `colour=blue` or `?price-range=20-50`
- Identifying – For example `?product=small-blue-widget`, `categoryid=124` or `itemid=24AU`
- Paginating – For example, `?page=2`, `?p=2` or `viewItems=10-30`
- Searching – For example, `?query=users-query`, `?q=users-query` or `?search=drop-down-option`
- Translating – For example, `?lang=fr`, `?language=de` or

SEO Issues with URL Parameters

1. Parameters Create Duplicate Content

Often, URL parameters make no significant change to the content of a page. A re-ordered version of the page is often not so different from the original. A page URL with tracking tags or a session ID is identical to the original.

For example, the following URLs would all return collection of widgets.

- Static URL: <https://www.example.com/widgets>
- Tracking parameter: <https://www.example.com/widgets?sessionID=32764>
- Reordering parameter: <https://www.example.com/widgets?sort=newest>
- Identifying parameter: <https://www.example.com?category=widgets>
- Searching parameter: <https://www.example.com/products?search=widget>

That's quite a few URLs for what is effectively the same content – now imagine this over every category on your site. It can really add up.

The challenge is that search engines treat every parameter based URL is a new page. So they see multiple variations of the same

page. All serving duplicate content and all targeting the same keyword phrase or semantic topic.

While such duplication is unlikely to cause you to be completely filtered out of the search results, it does lead to [keyword cannibalization](#) and could downgrade Google's view of your overall site quality as these additional URLs add no real value.

2. Parameters Waste Crawl Budget

Crawling redundant parameter pages drains crawl budget, reducing your site's ability to index SEO relevant pages and increasing server load.

Google [sums up](#) this point perfectly.

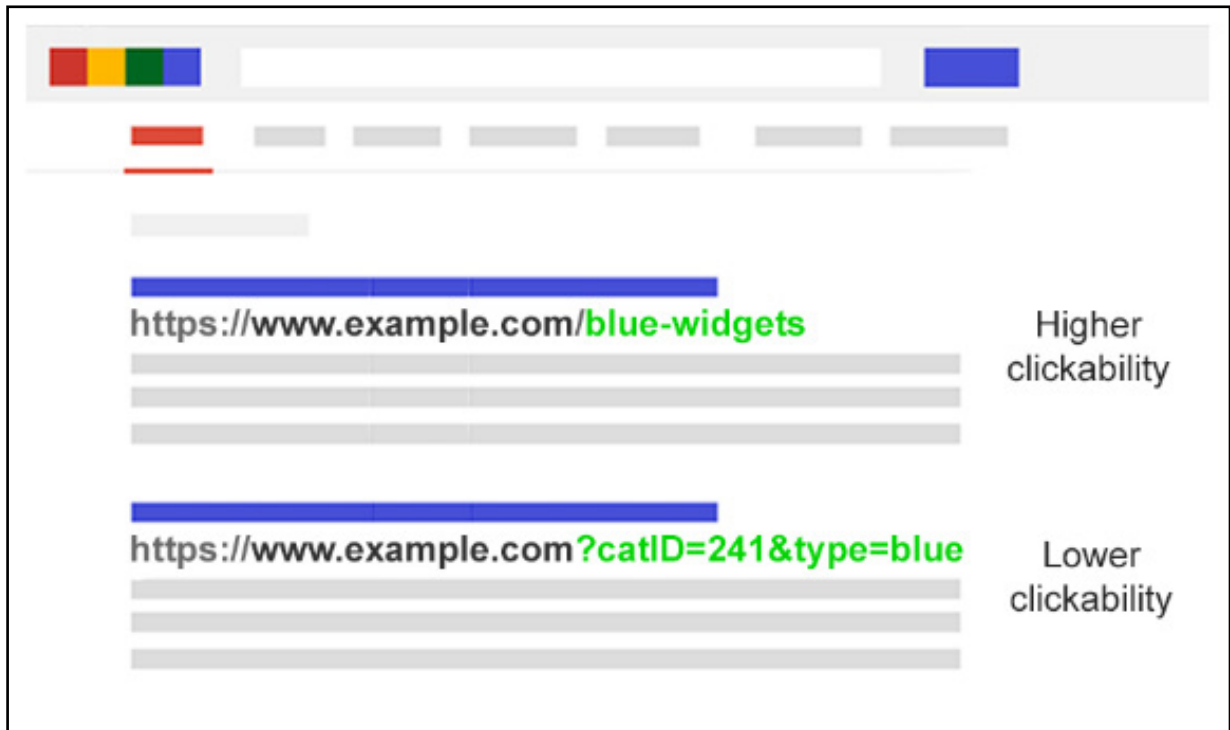
“Overly complex URLs, especially those containing multiple parameters, can cause a problems for crawlers by creating unnecessarily high numbers of URLs that point to identical or similar content on your site. As a result, Googlebot may consume much more bandwidth than necessary, or may be unable to completely index all the content on your site.”

3. Parameters Split Page Ranking Signals

If you have multiple permutations of the same page content, links and social shares may be coming in on various versions.

This dilutes your ranking signals. When you confuse a crawler, it becomes unsure which of the competing pages to index for the search query.

4. Parameters Make URLs Less Clickable



Let's face it. Parameter URLs are unsightly. They're hard to read. They don't seem as trustworthy. As such, they are less likely to be clicked.

This will impact page performance. Not only because CTR can influence rankings, but also because it's less clickable on social media, in emails, when copy pasted into forums or anywhere else the full URL may be displayed.

While this may only have a fractional impact on a single page's amplification, every tweet, like, share, email, link, and mention matters for the domain.

Poor URL readability could contribute to a decrease in brand engagement.

Assess the Extent of Your Parameter Problem

It's important to know every parameter used on your website. But chances are your developers don't keep an up to date list.

So how do you find all the parameter that need handling? Or understand [how search engines crawl and index](#) such pages? Know the value they bring to users?

Follow these five steps:

- Run a crawler: With a tool like Screaming Frog you can search for "?" in the URL.
- Look in Google Search Console URL Parameters Tool: Google auto-adds the query strings it finds.
- Review your log files: See if Googlebot is crawling parameter based URLs.
- Search with site: inurl: advanced operators: Know how Google is indexing the parameters you found by putting the key in a site:example.com inurl:key combination query.

- Look in Google Analytics All Pages report: Search for “?” to see how each of the parameters you found are used by users. Be sure to check that URL query parameters have not been excluded in the view setting.

Armed with this data, you can now decide how to best handle each of your website's parameters.

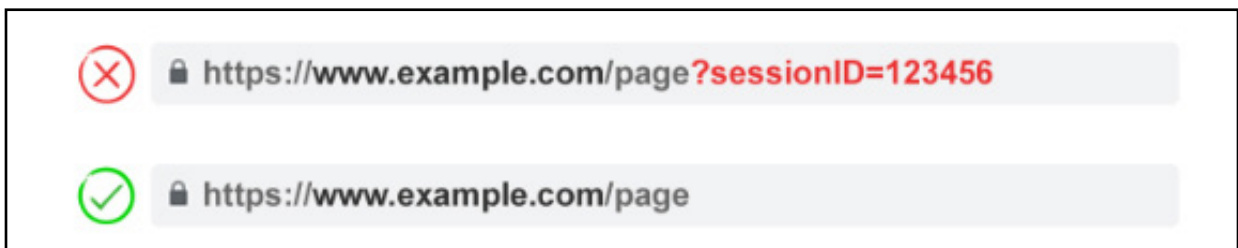
SEO Solutions to Tame URL Parameters

You have six tools in your SEO arsenal to deal with URL parameters on a strategic level.

Limit Parameter-Based URLs

A simple review of how and why parameters are generated can provide an SEO quick win. You will often find ways to reduce the number of parameter URLs and so minimize the negative SEO impact. There are four common issues to begin your review.

1. Eliminate Unnecessary Parameters



Ask your developer for a list of every website parameter and its function. Chances are, you will discover parameters that no longer perform a valuable function.

For example, users can be better identified by [cookies](#) than sessionIDs. Yet the sessionID parameter may still exist on your website as it was used historically.

Or you may discover that a filter in your faceted navigation is rarely applied by your users.

Any parameters caused by technical debt should be immediately eliminated.

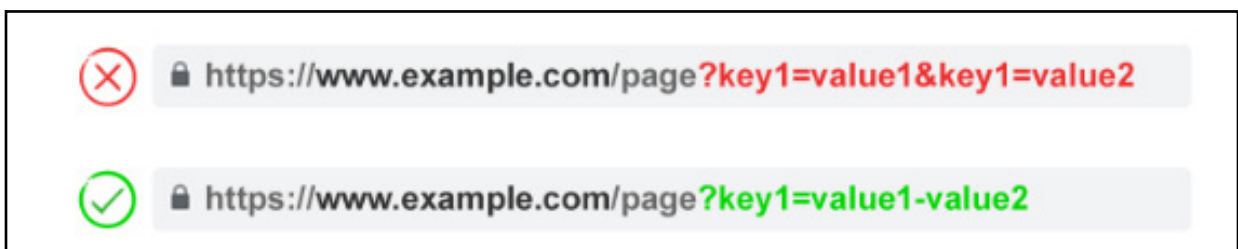
2. Prevent Empty Values



URL parameters should be added to a URL only when they have a function. Don't permit parameter keys to be added if the value is blank.

In the above example, key2 and key3 add no value both literally and figuratively.

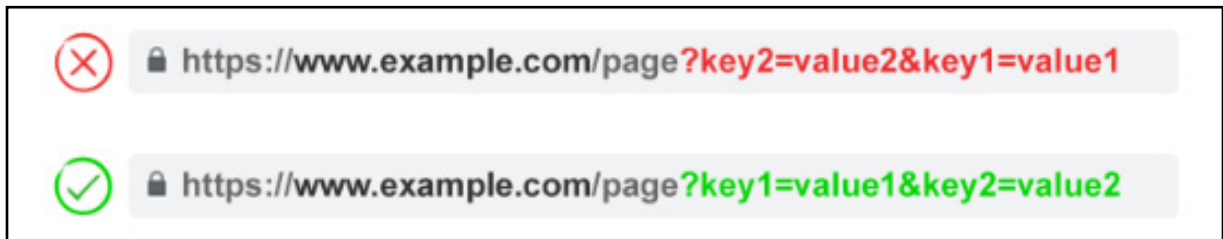
3. Use Keys Only Once



Avoid applying multiple parameters with the same parameter name and a different value.

For multi-select option, it is better to combine the values together after a single key.

4. Order URL Parameters



If the same URL parameter are rearranged, the pages are interpreted by search engines as equal. As such, parameter order doesn't matter from a duplicate content perspective. But each of those combinations burn crawl budget and split ranking signals.

Avoid these issues by asking your developer to write a script to always place parameters in a consistent order, regardless of how the user selected them.

In my opinion, you should start with any translating parameters, followed by identifying, then pagination, then layering on filtering and reordering or search parameters and finally tracking.

Pros:

- Allows more efficient use of crawl budget.
- Reduces duplicate content issues.
- Consolidates ranking signals to fewer pages.
- Suitable for all parameter types.

Cons:

- Moderate technical implementation time

Rel="Canonical" Link Attribute



```
<!DOCTYPE html>
<html>
<head>
<link rel="canonical" href="https://www.example.com/page" />
</head>
<body>
...
</body>
</html>
```

The rel="canonical" link attribute calls out that a page has identical or similar content to another. This encourages search engines to consolidate the ranking signals to the URL specified as canonical.

You can rel=canonical your parameter based URLs to your SEO-friendly URL for tracking, identifying or reordering parameters. But this tactic is not suitable when the parameter page content is not close enough to the canonical, such as pagination, searching, translating or some filtering parameters.

Pros:

- Relatively easy technical implementation.
- Very likely to safeguard against duplicate content issues.
- Consolidates ranking signals to the canonical URL.

Cons:

- Wastes crawl budget on parameter pages.
- Not suitable for all parameter types.
- Interpreted by search engines as a strong hint, not a directive.

Meta Robots Noindex Tag



```
<!DOCTYPE html>
<html>
<head>
<meta name="robots" content="noindex, follow" />
</head>
<body>
...
</body>
</html>
```

Set a noindex directive for any parameter based page that doesn't add SEO value. This tag will prevent search engines from indexing the page.

URLs with a "noindex" tag are also likely to be crawled less frequently and if it's present for a long time will eventually [lead Google to nofollow the page's links.](#)

Pros:

- Relatively easy technical implementation.
- Very likely to safeguard against duplicate content issues.
- Suitable for all parameter types you do not wish to be indexed.
- Removes existing parameter-based URLs from the index.

Cons:

- Won't prevent search engines from crawling URLs, but will encourage them to do so less frequently.
- Doesn't consolidate ranking signals.
- Interpreted by search engines as a strong hint, not a directive.

Robots.txt Disallow



The robots.txt file is what search engines look at first before crawling your site. If they see something is disallowed, they won't even go there.

You can use this file to block crawler access to every parameter based URL (with `Disallow: /*?*`) or only to specific query strings you don't want to be indexed.

Pros:

- Simple technical implementation.
- Allows more efficient use of crawl budget.
- Avoids duplicate content issues.
- Suitable for all parameter types you do not wish to be crawled.

Cons:

- Doesn't consolidate ranking signals.
- Doesn't remove existing URLs from the index.

URL Parameter Tool in Google Search Console

URL Parameters

Help Google crawl your site more efficiently by indicating how we should handle parameters in your URLs. [Learn more.](#)

⚠ Only use this feature if you're sure how parameters work. Incorrectly excluding URLs could result in many pages disappearing from a search.

Parameter	URLs monitored	Configured	Effect	Crawl	
page	677,630	1 Jan 2019	Specifies	No URLs	Edit / Reset
sort	77,440	1 Jan 2019	Sorts	No URLs	Edit / Reset
utm_medium	74,196	1 Jan 2019	None	Representative URL	Edit / Reset
categoryid	52,581	1 Jan 2019	Narrows	Every URL	Edit / Reset
lang	24,924	1 Jan 2019	Translates	Every URL	Edit / Reset

Configure the Google's URL parameter tool to tell crawlers the purpose of your parameters and how you would like them to be handled.

Google Search Console has a warning message that using the tool “could result in many pages disappearing from a search.”

This may sound ominous. But what's more menacing is thousands of duplicate pages hurting your website's ability to rank.

So it's best to learn how to configure URL parameters in Google Search Console, rather than letting Googlebot decide.

The key is to ask yourself how the parameter impacts the page content:

- Tracking parameters don't change page content. Configure them as “representative URLs”.

- Configure parameters that reorder page content as “sorts”. If this is optionally added by the user, set crawl to “No URLs”. If a sort parameter it is applied by default, use “Only URLs with value”, entering the default value.
- Configure parameters that filter page down to a subset of content as “narrows”. If these filters are not SEO relevant, set crawl to “No URLs”. If they are SEO relevant set to “Every URL”.
- Configure parameters that shows a certain piece or group of content as “specifies”. Ideally, this should be static URL. If not possible, you will likely want to set this to “Every URL”.
- Configure parameters that display a translated version of the content as “translates”. Ideally, translation should be achieved via subfolders. If not possible, you will likely want to set this to “Every URL”.
- Configuration parameters that display a component page of a longer sequence as “paginates”. If you have achieved **efficient indexation with XML sitemaps**, you can save crawl budget and set crawl to “No URL”. If not, set to “Every URL” to help crawlers to reach all of the items.

Google will automatically add parameters to the list under the default “Let Googlebot decide”. The challenge is, these can never be removed, even if the parameter no longer exists. So whenever possible, it’s best to proactively add parameters yourself. So that if at any point that parameter no longer exists, you may delete it from GSC.

For any parameter you set in Google Search Console to “No URL”, you should also consider adding it in Bing’s ignore URL parameters tool.

Pros:

- No developer time needed.
- Allows more efficient use of crawl budget.
- Likely to safeguard against duplicate content issues. Suitable for all parameter types.

Cons:

- Doesn’t consolidate ranking signals.
- Interpreted by Google as a helpful hint, not a directive.
- Only works for Google and with lesser control for Bing.

Move From Dynamic to Static URLs

Many people think the optimal way to handle URL parameters is simply avoid them in the first place. After all, subfolders surpass parameters to help Google understand site structure and static, keyword based URLs have always been a cornerstone of on-page SEO.

To achieve this, you can use server-side URL rewrites to convert parameters into subfolder URLs.

For example, the URL:

`www.example.com/view-product?id=482794`

Would become:

`www.example.com/widgets/blue`

This approach works well for descriptive keyword based parameters, such as those which identify categories, products, or filter for search engine relevant attributes. It is also effective for translated content.

But it becomes problematic for non-keyword relevant elements of **faceted navigation**, such as price. Having such a filter as a static, indexable URL offers no SEO value.

It's also an issue for searching parameters, as every user generated query would create a static page that vies for ranking against the canonical – or worse presents to crawlers low quality content pages whenever a user has searched for a item you don't offer.

It's somewhat odd when applied to pagination (although not uncommon due to WordPress), which would give a URL such as **www.example.com/widgets/blue/page2**

Very odd for reordering, which would give a URL such as **www.example.com/widgets/blue/lowest-price**

And is often not a viable option for tracking. Google Analytics will not acknowledge a static version of UTM parameter.

More to the point, by replacing dynamic parameters with static URLs for things like pagination, onsite search box results or sorting does not address duplicate content, crawl budget or internal link equity dilution.

And having all the combinations of filters from your faceted navigation as indexable URLs often results in **thin content** issues. Especially if you offer multi-select filters.

Many SEO pros argue it's possible to provide the same user experience without impacting the URL. For example, by using POST rather than GET requests to modify the page content. Thus, preserving the user experience and avoiding the SEO problems.

But stripping out parameters in this manner would remove the possibility for your audience to bookmark or share a link to that specific page. And if obviously not feasible for tracking parameters and not optimal for pagination.

The crux of the matter is that for many websites, completing avoiding parameters is simply not possible if you want to provide the ideal user experience. Nor would it be best practice SEO.

So we are left with this. For parameters that you don't want to be indexed in search results (paginating, reordering, tracking, etc) implement as query strings. For parameters that you do want to be indexed, use static URL paths.

Pros:

- Shifts crawler focus from parameter based to static URLs which have a higher likelihood to rank.

Cons:

- Significant investment of development time for URL rewrites and 301 redirects.
- Doesn't prevent duplicate content issues.
- Doesn't consolidate ranking signals.
- Not suitable for all parameter types.
- May lead to thin content issues.
- Doesn't always provide a linkable or bookmarkable URL.

Best Practice URL Parameter Handling for SEO

So which of these six SEO tactics should you implement? The answer can't be all of them.

Not only would that create unnecessary complexity. But often the SEO solutions actively conflict with one another.

For example, if you implement robots.txt disallow, Google would not be able to see any meta noindex tag. You also **shouldn't combine a meta noindex tag with a rel=canonical** link attribute.

What becomes clear is there is no one perfect solution.

Even Google's John Mueller can't decide on an approach. In **a Google Webmaster hangout**, he initially recommended against disallowing parameters, but when questioned on this from a faceted navigation perspective, answered "it depends."

There are occasions when crawling efficiency is more important than consolidating authority signals.

Ultimately, what's right for your website will depend on your priorities.

	Easy to implement	Save crawl budget	Manage duplicate content	Consolidate ranking signals	Suitable for all parameter types
Limit parameters	✗	✓	✓	✓	✓
Canonical link attribute	✓	✗	✓	✓	✗
Noindex tag	✓	✗	✓	✗	✓
Robots.txt file	✓	✓	✓	✗	✓
Parameter tool	✓	✓	✓	✗	✓
Static URLs	✗	✗	✗	✗	✗

Personally, I don't use noindex or block access to parameter pages. If Google can't crawl and understand all the URL variables, it can't consolidate the ranking signals to the canonical page.

I take the following plan of attack for SEO-friendly parameter handling:

- Do keyword research to understand what parameters should be search engine friendly, static URLs.
- Implement **correct pagination handling** with rel="next" & rel="prev".
- For all remaining parameter based URLs, implement consistent ordering rules, which use keys only once and prevent empty values to limit the number of URLs.

- Add a rel=canonical link attribute to suitable parameter pages to combine ranking ability.
- Configure URL parameter handling in both Google and Bing as a failsafe to help search engines understand each parameter's function.
- Double check no parameter based URLs are being submitted in the XML sitemap.

No matter what parameter handling strategy you choose to implement, be sure to **document the impact** of your efforts on KPIs.

Chapter 20

How to Perform an In-Depth Technical SEO Audit

SEJ
EBOOK

Written By
Anna Crowe
Assistant Editor, Search Engine Journal



I'm not going to lie: Conducting an in-depth SEO audit is a major deal.

And, as an SEO consultant, there are a few sweeter words than, “Your audit looks great! When can we bring you onboard?”

Even if you haven't been actively looking for a new gig, knowing your SEO audit nailed it is a huge ego boost.

But, are you terrified to start? Is this **your first SEO audit?** Or, you just don't know where to begin? Sending a fantastic SEO audit to a potential client puts you in the best possible place.

It's a rare opportunity for you to organize your processes and rid your potential client of bad habits (cough*unpublishing pages without a 301 redirect*cough) and crust that accumulates like the lint in your dryer.

So take your time. Remember: Your primary goal is to add value to your customer with your site recommendations for both the short-term and the long-term.

Ahead, I've put together the need-to-know steps for [conducting an SEO audit](#) and a little insight to the first phase of my processes when I first get a new client. It's broken down into sections below. If you feel like you have a good grasp on a particular section, feel free to jump to the next.

This is a series, so stay tuned for more SEO audit love.

Jump to:

- [When Should I Perform an SEO Audit?](#)
- [What You Need from a Client Before an SEO Audit](#)
- [Tools for SEO Audit](#)
- [Technical > DeepCrawl](#)
- [Technical > Screaming Frog](#)
- [Technical > Google Search Console & Bing Webmaster Tools](#)
- [Technical > Google Analytics](#)

When Should I Perform an SEO Audit?

After a potential client sends me an email expressing interest in working together and they answer my survey, we set-up an intro call (Skype or Google Hangouts is preferred).

Before the call, I do my own **mini quick SEO audit** (I invest at least one hour to manually researching) based on their survey answers to become familiar with their market landscape. It's like dating someone you've never met.

You're obviously going to stalk them on Facebook, Twitter, Instagram, and all other channels that are public #solcreep.

Here's an example of what my survey looks like:

Here are some key questions you'll want to ask the client during the first meeting:

- 1.** What are your overall business goals? What are your channel goals (PR, social, etc.)?
- 2.** Who is your target audience?
- 3.** Do you have any business partnerships?
- 4.** How often is the website updated? Do you have a web developer or an IT department?
- 5.** Have you ever worked with an SEO consultant before? Or, had any SEO work done previously?

Sujan Patel also has some great recommendations on [questions to ask a new SEO client](#).

After the call, if I feel we're a good match, I'll send over my formal proposal and contract (thank you HelloSign for making this an easy process for me!).

To begin, I always like to offer my clients the first month as a trial period to make sure we vibe.

This gives both the client and I a chance to become friends first before dating. During this month, I'll take my time to conduct an in-depth SEO audit.

These SEO audits can take me anywhere from 40 hours to 60 hours depending on the size of the website.

These audits are bucketed into three separate parts and presented with Google Slides.

- **Technical:** Crawl errors, indexing, hosting, etc.
- **Content:** Keyword research, competitor analysis, content maps, meta data, etc.
- **Links:** Backlink profile analysis, growth tactics, etc.

After that first month, if the client likes my work, we'll begin implementing the recommendations from the SEO audit. And going forward, I'll perform a mini-audit monthly and an in-depth audit quarterly.

To recap, I perform an SEO audit for my clients:

- First month
- Monthly (mini-audit)
- Quarterly (in-depth audit)

What You Need from a Client Before an SEO Audit

When a client and I start working together, I'll share a Google doc with them requesting a list of passwords and vendors.

This includes:

- Google Analytics access and any third-party analytics tools
- Google and Bing ads
- Webmaster tools
- Website backend access
- Social media accounts
- List of vendors
- List of internal team members (including any work they outsource)

Tools for SEO Audit

Before you begin your SEO audit, here's a recap of the tools I use:

- [Screaming Frog](#)
- [Integrity](#) (for Mac users) and [Xenu Sleuth](#) (for PC users)
- [SEO Browser](#)
- [Wayback Machine](#)
- [Moz](#)
- [Buzzsumo](#)
- [DeepCrawl](#)
- [Copyscape](#)
- [Google Tag Manager](#)
- [Google Tag Manager Chrome Extension](#)
- [Annie Cushing's Campaign Tagging Guide](#)
- [Google Analytics](#) (if given access)
- [Google Search Console](#) (if given access)
- [Bing Webmaster Tools](#) (if given access)
- [You Get Signal](#)
- [Pingdom](#)
- [PageSpeed Tool](#)
- [Sublime Text](#)

My 30-Point Technical SEO Checklist

Technical

Tools needed for technical SEO audit:

- Screaming Frog
- DeepCrawl
- Copyscape
- Integrity for Mac (or Xenu Sleuth for PC users)
- Google Analytics (if given access)
- Google Search Console (if given access)
- Bing Webmaster Tools (if given access)

Step 1: Add Site to DeepCrawl and Screaming Frog

Tools:

- DeepCrawl
- Copyscape
- Screaming Frog
- Google Analytics
- Integrity
- Google Tag Manager
- Google Analytics code

What to Look When Using DeepCrawl

The first thing I do is add my client's site to DeepCrawl. Depending on the size of your client's site, the crawl may take a day or two to get the results back.

Once you get your DeepCrawl results back, here are the things I look for:

■ Duplicate Content

Check out the “Duplicate Pages” report to locate duplicate content.

If duplicate content is identified, I’ll make this a top priority in my recommendations to the client to rewrite these pages and in the meantime, I’ll add the `<meta name=“robots” content=“noindex, nofollow”>` tag to the duplicate pages.

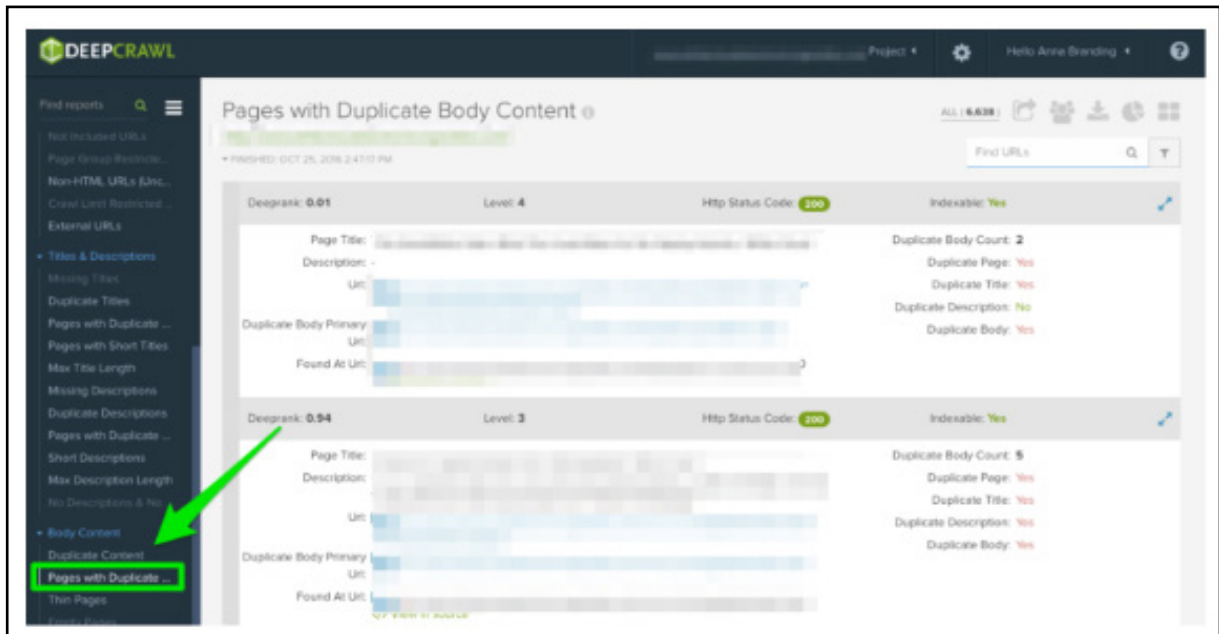
Common duplicate content errors you’ll discover:

- Duplicate meta titles and meta descriptions
- Duplicate body content from tag pages (I’ll use Copyscape to help determine if something is being plagiarized).
- Two domains (ex: yourwebsite.co, yourwebsite.com)
- Subdomains (ex: jobs.yourwebsite.com)
- Similar content on a different domain
- Improperly implemented pagination pages (see below.)

How to fix:

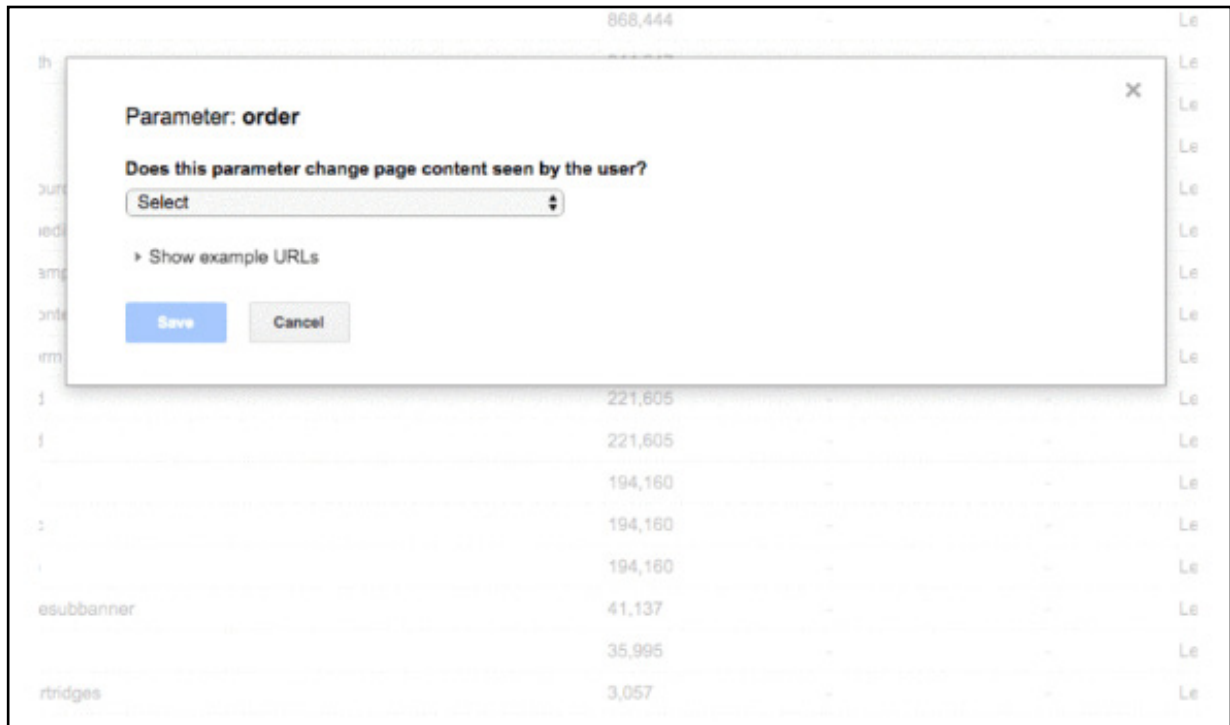
- Add the canonical tag on your pages to let Google know what you want your preferred URL to be.
- Disallow incorrect URLs in the robots.txt.
- Rewrite content (including body copy and meta data).

Here's an example of a duplicate content issue I had with a client of mine. As you can see below, they had URL parameters without the canonical tag.



These are the steps I took to fix the issue:

- I fixed any 301 redirect issues.
- Added a canonical tag to the page, I want Google to crawl.
- Update the Google Search Console parameter settings to exclude any parameters that don't generate unique content.



- Added the disallow function to the robots.txt to the incorrect URLs to improve crawl budget

■ Pagination

There are two reports to check out:

- **First Pages:** To find out what pages are using pagination, review the "First Pages" report. Then, you can manually review the pages using this on the site to discover if pagination is implemented correctly.
- **Unlinked Pagination Pages:** To find out if pagination is working correctly, the "Unlinked Pagination Pages" report will tell you if the rel="next" and rel="prev" are linking to the previous and next pages.

In this example below, I was able to find that a client had reciprocal pagination tags using DeepCrawl:

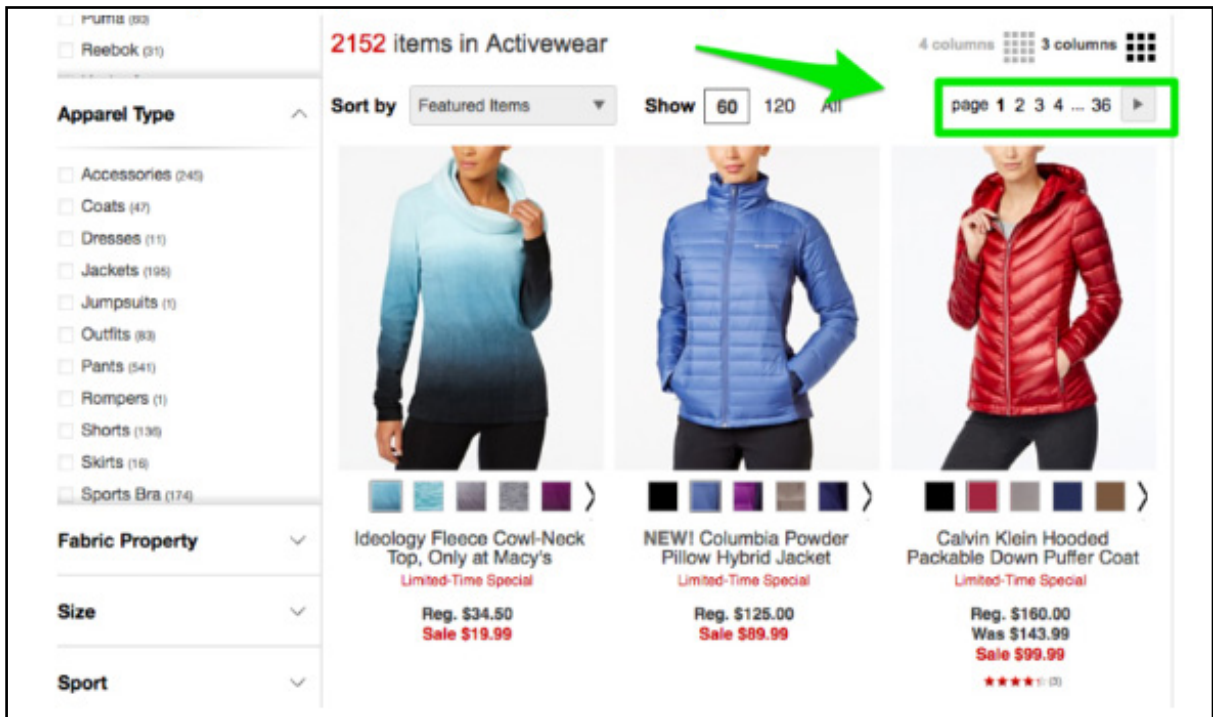
The screenshot shows the DeepCrawl interface with the 'Unlinked Paginated Pages' report. The left sidebar menu has 'Unlinked Paginated P...' highlighted with a green box and a green arrow pointing to it. The main content area shows two entries for Level 4 and Level 5. Each entry includes fields for Page Title, Url, Rel Next Url, Rel Prev Url, Found At Url, Deeprank, Level, Http Status Code, Indexable, Links In Count, and Paginated Page. The 'Paginated Page' field is set to 'Yes' for both entries.

How to fix:

- If you have a “view all” or a “load more” page, add rel=”canonical” tag. Here’s an example from [Crutchfield](#):

The screenshot shows a product page from Crutchfield. The product price is \$1,799.99. There is a green 'Add to Cart' button. Below the product, there is a blue 'Load more products' button. At the bottom, there is a section for 'Related led tv articles from our experts'.

- If you have all your pages on separate pages, then add the standard rel="next" and rel="prev" markup. Here's an example from Macy's:



Max Redirections

Review the "Max Redirections" report to see all the pages that redirect more than 4 times. [John Mueller](#) mentioned in 2015 that Google can stop following redirects if there are more than five.

While some people refer to these crawl errors as eating up the "crawl budget," Gary Illyes refers to this as "host load". It's important to make sure your pages render properly because you want your host load to be used efficiently.

Here's a brief overview of the response codes you might see:

- **301** — These are the majority of the codes you'll see throughout your research. 301 redirects are okay as long as there are only one redirect and no redirect loop.
- **302** — These codes are okay, but if left longer than 3 months or so, I would manually change them to 301s so that they are permanent. This is an error code I'll see often with e-commerce sites when a product is out of stock.
- **400** — Users can't get to the page.
- **403** — Users are unauthorized to access the page.
- **404** — The page is not found (usually meaning the client deleted a page without a 301 redirect).
- **500** — Internal server error that you'll need to connect with the web development team to determine the cause.

How to fix:

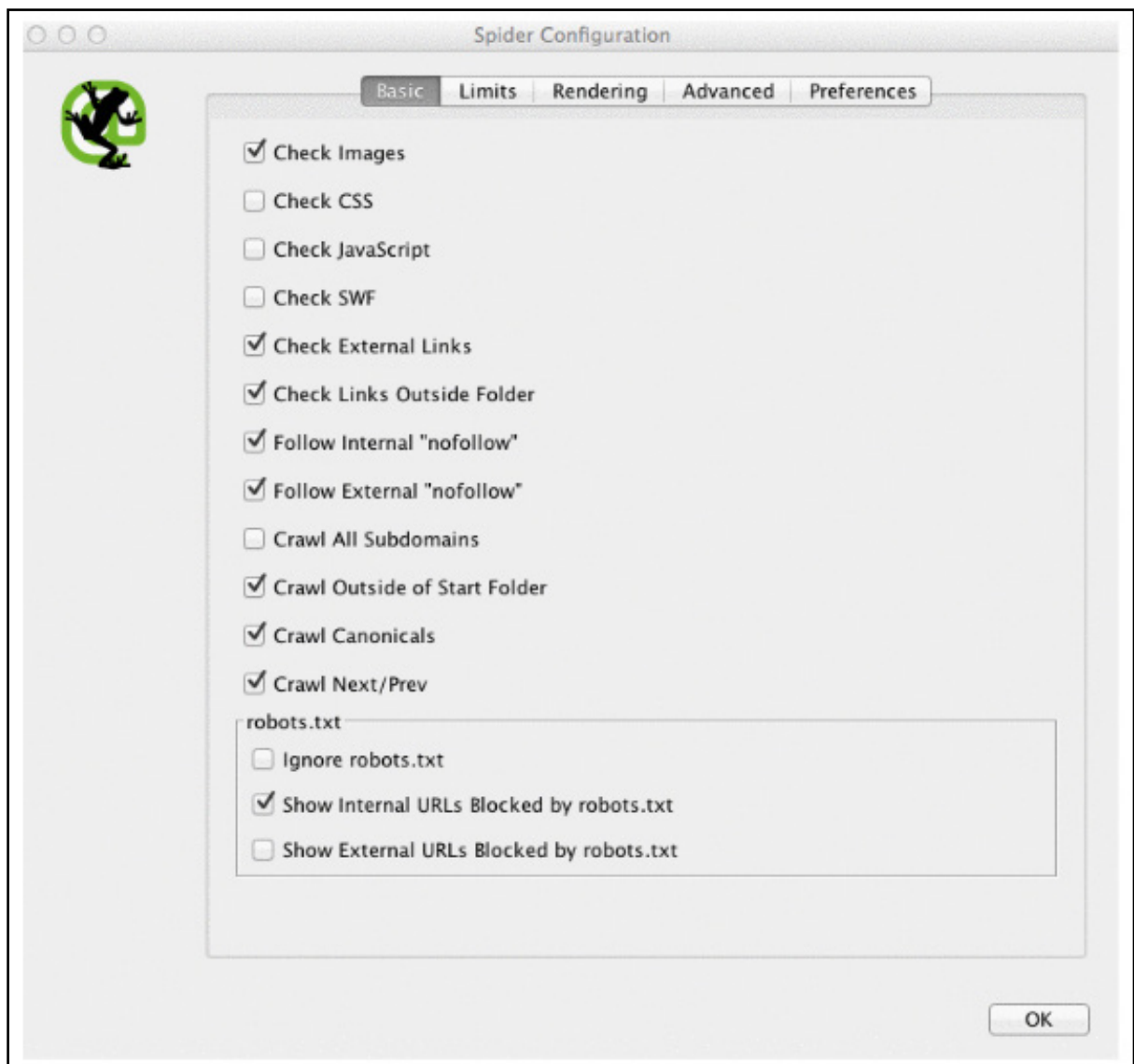
- Remove any internal links pointing to old 404 pages and update them with the redirected page internal link.
- Undo the redirect chains by removing the middle redirects. For example, if redirect A goes to redirect B, C, and D, then you'll want to undo redirects B and C. The final result will be a redirect A to D.
- There is also a way to do this in Screaming Frog and Google Search Console below if you're using that version.

What to Look For When Using Screaming Frog

The second thing I do when I get a new client site is to add their URL to Screaming Frog.

Depending on the size of your client's site, I may configure the settings to crawl specific areas of the site at a time.

Here is what my Screaming Frog spider configurations look like:



You can do this in your spider settings or by excluding areas of the site.

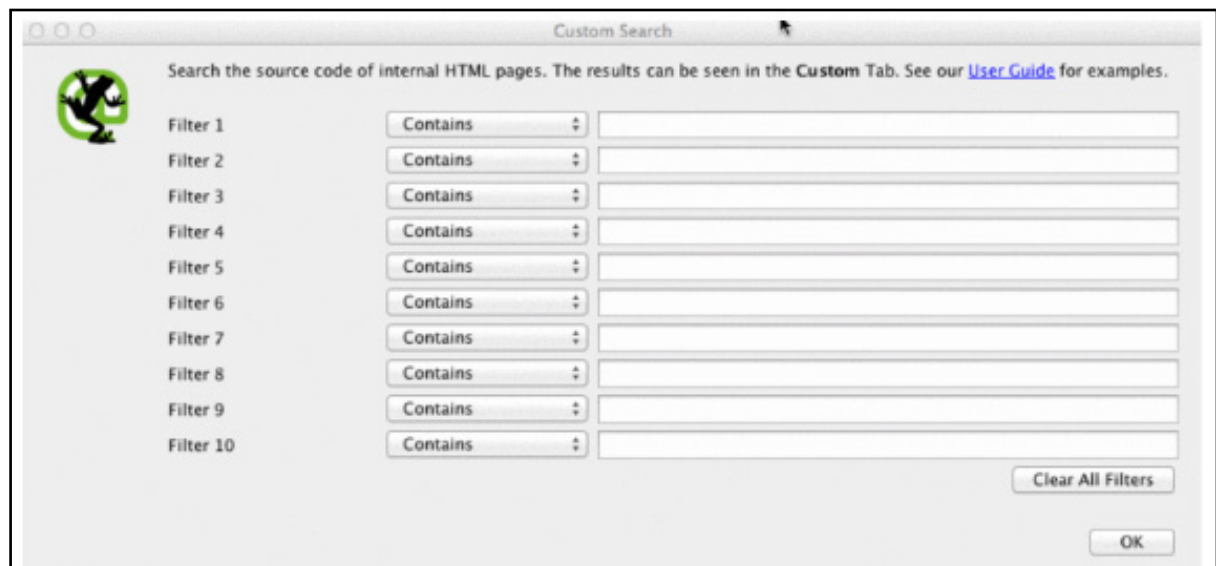
Once you get your Screaming Frog results back, here are the things I look for:

Google Analytics Code

Screaming Frog can help you identify what pages are missing the Google Analytics code (UA-1234568-9).

To find the missing Google Analytics code, follow these steps:

- Go to 'Configuration' in the navigation bar, then Custom.
- Add analytics\.js to Filter 1, then change the drop down to 'Does not contain.'



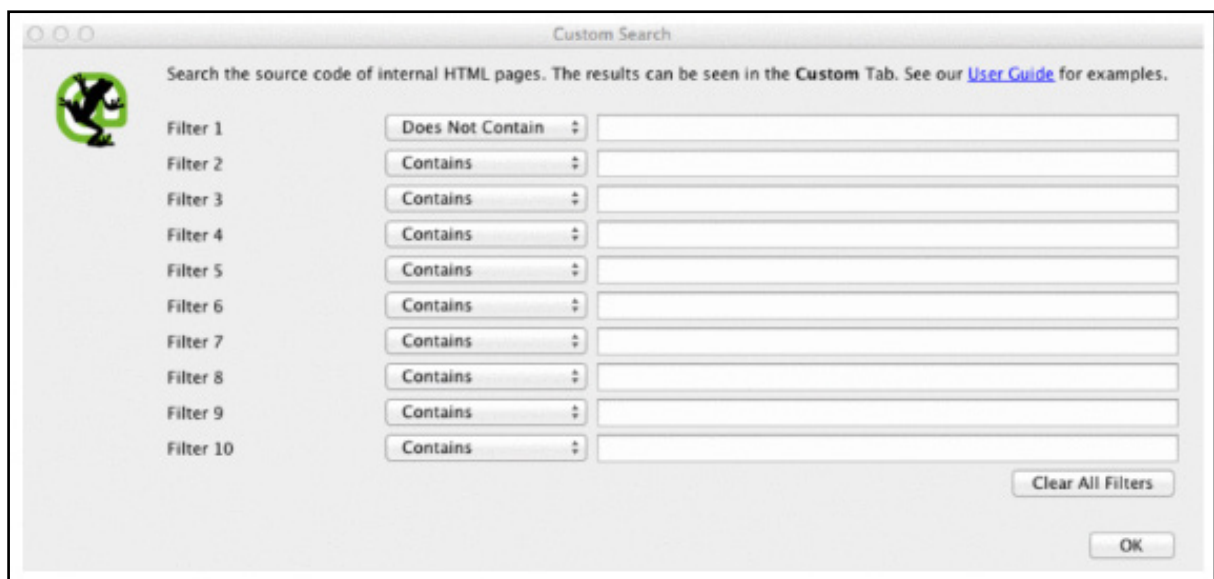
How to fix:

- Contact your client's developers and ask them to add the code to the specific pages that it's missing.
- For more Google Analytics information, skip ahead to that Google Analytics section below.

Google Tag Manager

Screaming Frog can also help you find out what pages are missing the Google Tag Manager snippet with similar steps:

- Go to the 'Configuration' tab in the navigation bar, then Custom. Add `<iframe src="//www.googletagmanager.com/` with 'Does not contain' selected in the Filter.



How to fix:

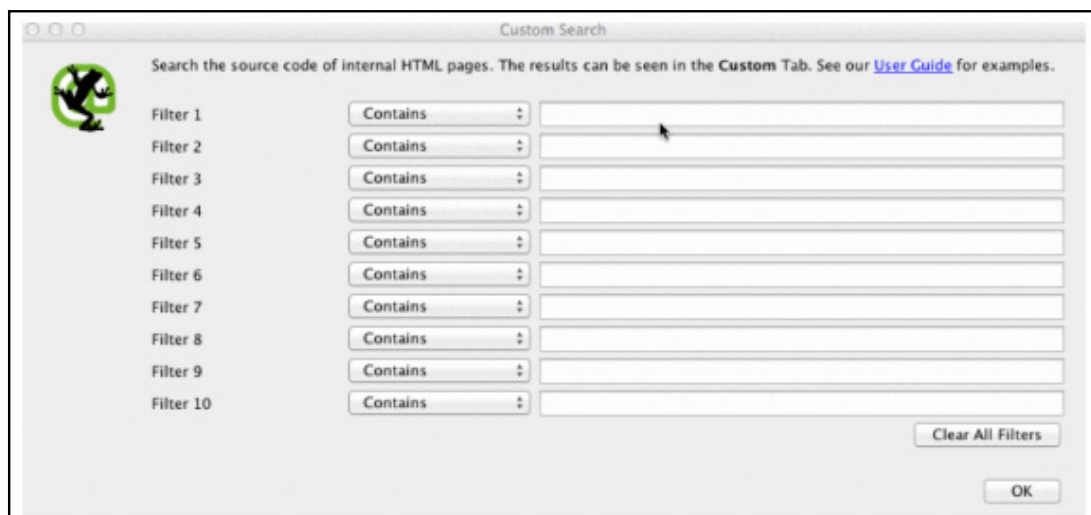
- Head over to Google Tag Manager to see if there are any errors and update where needed.
- Share the code with your client's developer's to see if they can add it back to the site.

Schema

You'll also want to check if your client's site is using schema markup on their site. Schema or structured data helps search engines understand what a page is on the site.

To check for schema markup in Screaming Frog, follow these steps:

- Go to the 'Configuration' tab in the navigation bar, then 'Custom.'
- Add itemtype="http://schema.\.org/" with 'Contain' selected in the Filter.



Indexing

You want to determine how many pages are being indexed for your client, follow this in Screaming Frog:

- After your site is done loading in Screaming Frog, go to Directives > Filter > Index to review if there are any missing pieces of code.



How to fix:

- If the site is new, Google may have no indexed it yet.
- Check the robots.txt file to make sure you're not disallowing anything you want Google to crawl.
- Check to make sure you've submitted your client's sitemap to Google Search Console and Bing Webmaster Tools.
- Conduct manual research (seen below).

Flash

Google announced this year that Chrome will start blocking Flash due to the slow page load times. So, if you're doing an audit, you want to identify if your new client is using Flash or not.

To do this in Screaming Frog, try this:

- Head to the 'Spider Configuration' in the navigation.
- Click 'Check SWF.'
- Filter the 'Internal' tab by 'Flash' after the crawl is done.

Address	Content	Status Code	Status	Title
1 http://annaleacrowe.com/	text/html; charset=UTF-8	200 OK	OK	SEO Freelancer And Designer in Tampa, FL ann
2 http://annaleacrowe.com/wp-includes/js/jquery/jquery.js?ver=1.12.4	application/javascript	200 OK	OK	
3 http://annaleacrowe.com/wp-content/plugins/wp-optimize-by-traffic/public/js/...	application/javascript	200 OK	OK	
4 http://annaleacrowe.com/wp-includes/js/wp-embed.min.js?ver=877636174122...	application/javascript	200 OK	OK	

How to fix:

- Embed videos from YouTube. [Google bought YouTube](#) in 2006, no-brainer here.
- Or, opt for HTML5 standards when adding a video.

Here's an example of HTML5 code for adding a video:

```
<video controls="controls" width="320" height="240">&gt;
  <source class="hiddenSpellError" data-mce-bogus="1" />src="/
  tutorials/media/Anna-Teaches-SEO-To-Small-Businesses.mp4"
  type="video/mp4"&gt;
  <source src="/tutorials/media/Anna-Teaches-SEO-To-Small-
  Businesses.ogg" type="video/ogg" />
```

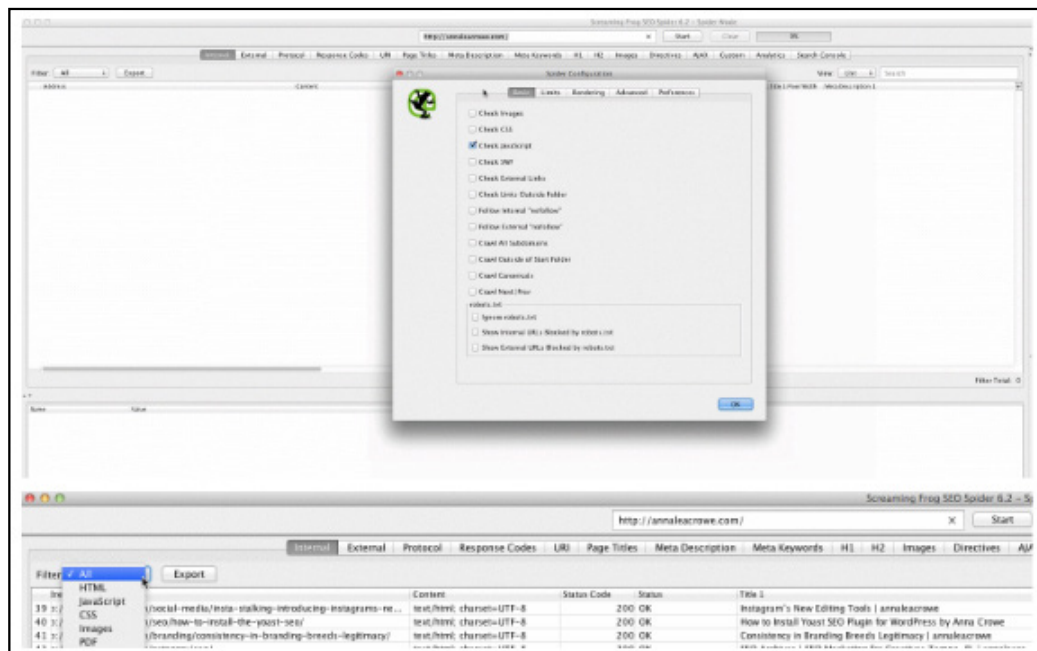
Your browser does not support the video tag.</video>

Javascript

According to [Google's announcement in 2015](#), JavaScript is okay to use for your website as long as you're not blocking anything in your robots.txt (we'll dig into this deeper in a bit!). But, you still want to take a peek at how the Javascript is being delivered to your site.

How to fix:

- Review Javascript to make sure it's not being blocked by robots.txt
- Make sure Javascript is running on the server (this helps produce plain text data vs dynamic).
- If you're running Angular JavaScript, check out this article by Ben Oren on [why it might be killing your SEO efforts](#).
- In Screaming Frog, go to the Spider Configuration in the navigation bar and click 'Check JavaScript.' After the crawl is done, filter your results on the 'Internal' tab by 'JavaScript.'



■ Robots.txt

When you're reviewing a robots.txt for the first time, you want to look to see if anything important is being blocked or disallowed.

For example, if you see this code:

```
User-agent: *
```

```
Disallow: /
```

Your client's website is blocked from all web crawlers.

But, if you have something like Zappos robots.txt file, you should be good to go.

```
# Global robots.txt as of 2012-06-19
```

```
User-agent: *
```

```
Disallow: /bin/
```

```
Disallow: /multiview/
```

```
Disallow: /product/review/add/
```

```
Disallow: /cart
```

```
Disallow: /login
```

```
Disallow: /logout
```

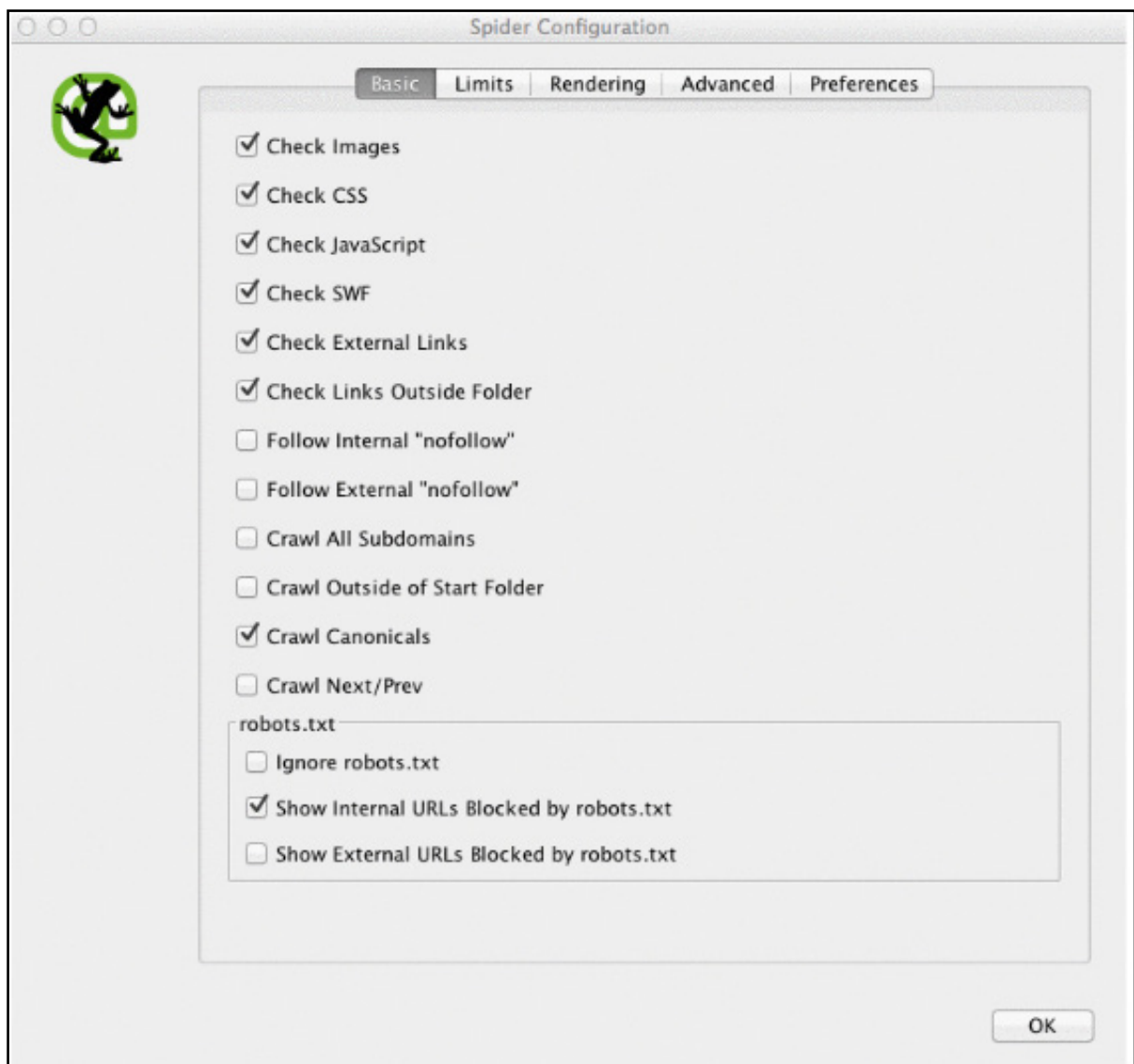
```
Disallow: /register
```

```
Disallow: /account
```

They are only blocking what they do not want web crawlers to locate. This content that is being blocked is not relevant or useful to the web crawler.

How to fix:

- Your robots.txt is case-sensitive so update this to be all lowercase.
- Remove any pages listed as Disallow that you want the search engines to crawl.
- Screaming Frog by default will not be able to load any URLs disallowed by robots.txt. If you choose to switch up the default settings in Screaming Frog, it will ignore all the robots.txt.



- You can also view blocked pages in Screaming Frog under the 'Response Codes' tab, then filtered by 'Blocked by Robots.txt' filter after you've completed your crawl.
- If you have a site with multiple subdomains, you should have a separate robots.txt for each.
- Make sure the sitemap is listed in the robots.txt.

■ Crawl Errors

I use DeepCrawl, Screaming Frog, and Google and Bing webmaster tools to find and cross-check my client's crawl errors.

To find your crawl errors in Screaming Frog, follow these steps:

- After the crawl is complete, go to 'Bulk Reports.'
- Scroll down to 'Response Codes,' then export the server side error report and the client error report.

How to fix:

- The client error reports, you should be able to 301 redirect the majority of the 404 errors in the backend of the site yourself.
- The server error reports, collaborate with the development team to determine the cause. Before fixing these errors on the root directory, be sure to backup the site. You may simply need to create a new .html access file or increase PHP memory limit.
- You'll also want to remove any of these permanent redirects from the sitemap and any internal or external links.
- You can also use '404' in your URL to help track in Google Analytics.

Internal & External Links

When a user clicks on a link to your site and gets a 404 error, it's not a good user experience.

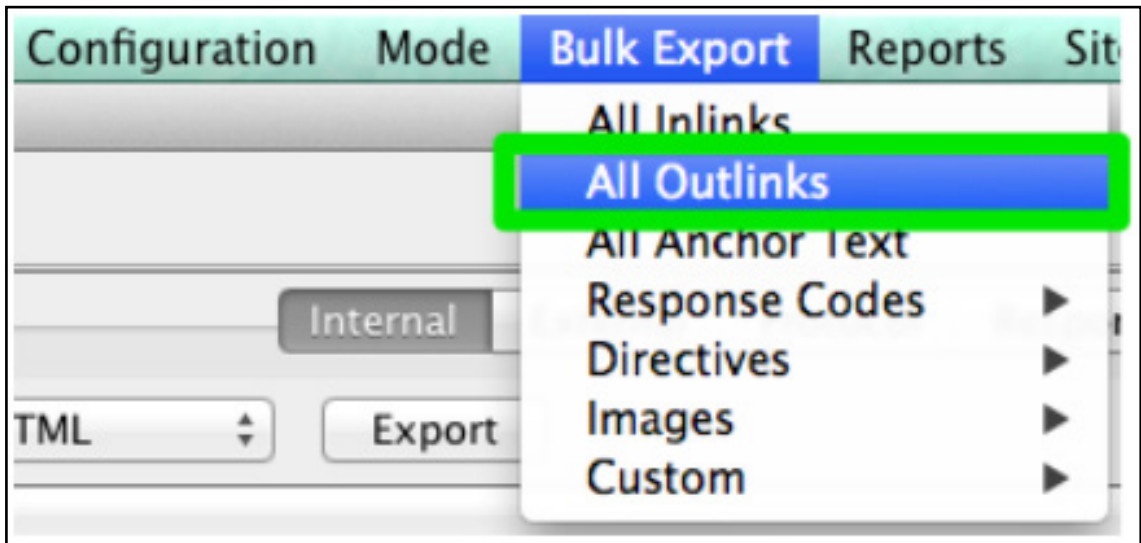
And, it doesn't help your search engines like any better either.

To find my broken internal and external links I use Integrity for Mac. You can also use Xenu Sleuth if you're a PC user.

I'll also show you how to find these internal and external links in Screaming Frog and DeepCrawl if you're using that software.

How to fix:

- If you're using Integrity or Xenu Sleuth, run your client's site URL and you'll get a full list of broken URLs. You can either manually update these yourself or if you're working with a dev team, ask them for help.
- If you're using Screaming Frog, after the crawl is completed, go to 'Bulk Export' in the navigation bar, then 'All Outlinks.' You can sort by URLs and see which pages are sending a 404 signal. Repeat the same step with 'All Inlinks.'



- If you're using DeepCrawl, go to the 'Unique Broken Links' tab under the 'Internal Links' section.

URLs

Every time you take on a new client, you want to review their URL format. What am I looking for in the URLs?

- Parameters – if the URL has weird characters like ?, =, or + it's a dynamic URL which can cause duplicate content if not optimized.
- User-friendly – I like to keep the URLs short and simple while also removing any extra slashes.

How to fix:

- You can search for parameter URLs in Google by doing `site:www.buyaunicorn.com/ inurl: "?"` or whatever you think the parameter might include.
- After you've run the crawl on Screaming Frog, take a look at URLs. If you see parameters listed that are creating duplicates of your content, you need to suggest the following:
- Add a canonical tag to the main URL page. For example, `www.buyaunicorn.com/magical-headbands` is the main page and I see `www.buyaunicorn.com/magical-headbands/?dir=mode123$`, then the canonical tag would need to be added to `www.buyaunicorn.com/magical-headbands`.
- Update your parameters in Google Search Console under 'Crawl' > 'URL Parameters.'

Parameter: **order** ✕

Does this parameter change page content seen by the user?

Select

▶ Show example URLs

	221,605	-	-
	194,160	-	-
	194,160	-	-
	194,160	-	-
yer	41,137	-	-
	35,995	-	-
	3,057	-	-

- Disallow the duplicate URLs in the robots.txt.

Step 2: Review Google Search Console and Bing Webmaster Tools

Tools:

- Google Search Console
- Bing Webmaster Tools
- Sublime Text (or any text editor tool)

Set a Preferred Domain

Since the Panda update, it's beneficial to clarify to the search engines the preferred domain. It also helps make sure all your links are giving one site the extra love instead of being spread across two sites.

How to fix:

- In Google Search Console, click the gear icon in the upper right corner.
- Choose which of the URLs is the preferred domain.
- You don't need to set the preferred domain in Bing Webmaster Tools, just submit your sitemap to help Bing determine your preferred domain.

Backlinks

With the announcement that Penguin is real-time, it's vital that your client's backlinks meet Google's standards. If you notice a large chunk of backlinks coming to your client's site from one page on a website, you'll want to take the necessary steps to clean it up, and FAST!

How to fix:

- In Google Search Console, go to 'Links' > then sort your 'Top linking sites.'

The screenshot shows the Google Search Console interface for the domain <https://annaleacrowe.com/>. The 'Links' report is displayed, showing a total of 309 external links and 952 internal links. The interface includes a sidebar with navigation options like Overview, Performance, URL Inspection, Index, Coverage, Sitemaps, Enhancements, Mobile Usability, Security & Manual Actions, Links, Settings, Submit feedback, and About this report.

External links (Total 309)

Top linked pages	Count
https://annaleacrowe.com/	256
https://annaleacrowe.com/case-studies/hot-dog-collars-case-study-2/	15
https://annaleacrowe.com/meet-anna/	9
https://annaleacrowe.com/services/on-site-seo/	6
https://annaleacrowe.com/services/off-site-seo/	5

Internal links (Total 952)

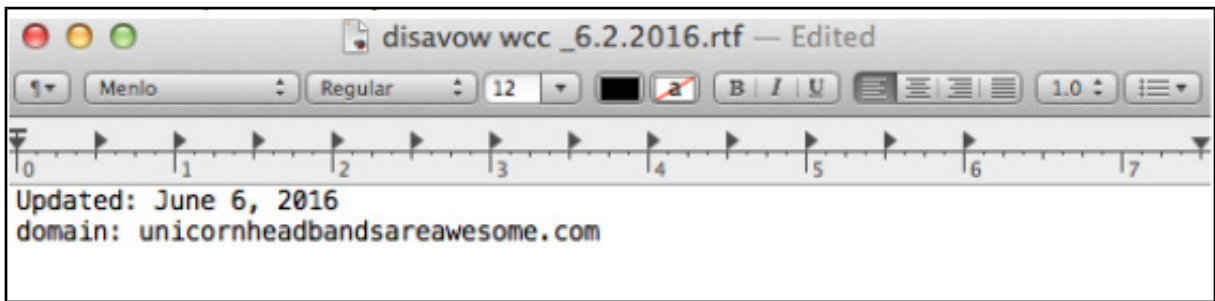
Top linked pages	Count
https://annaleacrowe.com/	62
https://annaleacrowe.com/case-studies/obj-case-study-2/	60
https://annaleacrowe.com/services/	59
https://annaleacrowe.com/blog/	59
https://annaleacrowe.com/faq/	59
https://annaleacrowe.com/case-studies/noodle-case-study-2/	59
https://annaleacrowe.com/case-studies/hot-dog-collars-case-study-2/	59
https://annaleacrowe.com/meet-anna/	59

Top linking sites

Top linking sites	Count
partitions.cf	59

- Contact the companies that are linking to you from one page to have them remove the links.
- Or, add them to your disavow list. When adding companies to your disavow list, be very careful how and why you do this. You don't want to remove valuable links.

Here's an example of what my disavow file looks like:



Keywords

As an SEO consultant, it's my job to start to learn the market landscape of my client. I need to know who their target audience is, what they are searching for, and how they are searching.

To start, I take a look at the keyword search terms they are already getting traffic from.

- In Google Search Console, under 'Search Traffic' > 'Search Analytics' will show you what keywords are already sending your client clicks.

2	music festival guide
3	fashion trends jewelry in california
4	music festival jewelry
5	fashion and music blogs
6	music festival fashion
7	festival jewelry

Sitemap

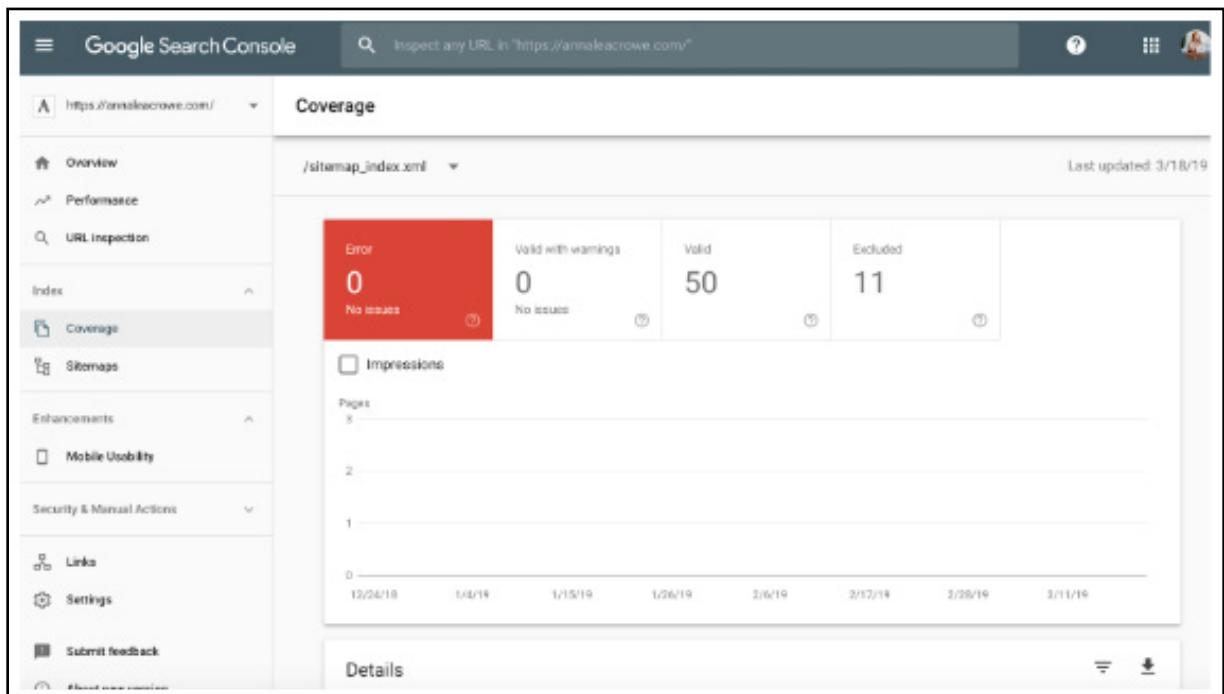
Sitemaps are essential to get search engines to crawl your client's website. It speaks their language.

When creating sitemaps, there are a few things to know:

- Do not include parameter URLs in your sitemap.
- Do not include any non-indexable pages.
- If the site has different subdomains for mobile and desktop, add the rel="alternate" tag to the sitemap.

How to fix:

- Go to 'Google Search Console' > 'Index' > 'Sitemaps' to compare the URLs indexed in the sitemap to the URLs in the web index.



- Then, do a manual search to determine pages are not getting indexed and why.
- If you find old redirected URLs in your client's sitemap, remove them. These old redirects will have an adverse impact on your SEO if you don't remove them.
- If the client is new, submit a new sitemap for them in both Bing and Google webmaster tools.

Sitemaps

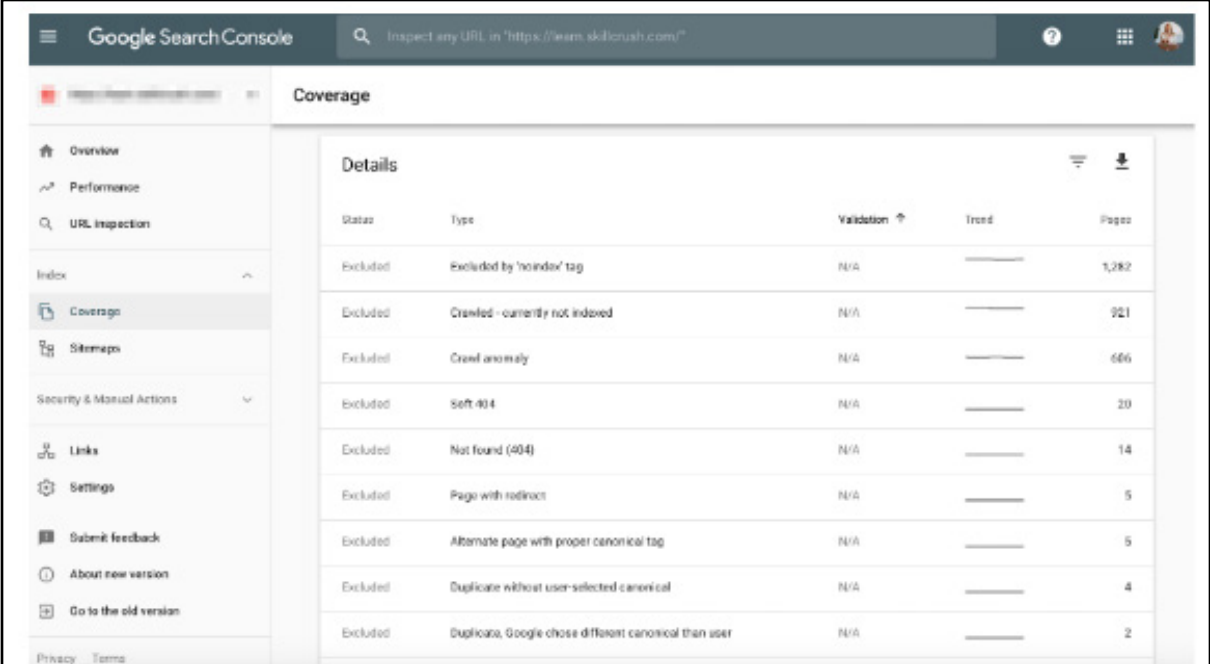
Add a new sitemap

<https://annaleacrowe.com/>

Crawl

Crawl errors are important to check because it's not only bad for the user but it's bad for your website rankings. And, John Mueller stated that low crawl rate may be a sign of a low-quality site.

To check this in Google Search Console, go to 'Coverage' > 'Details.'



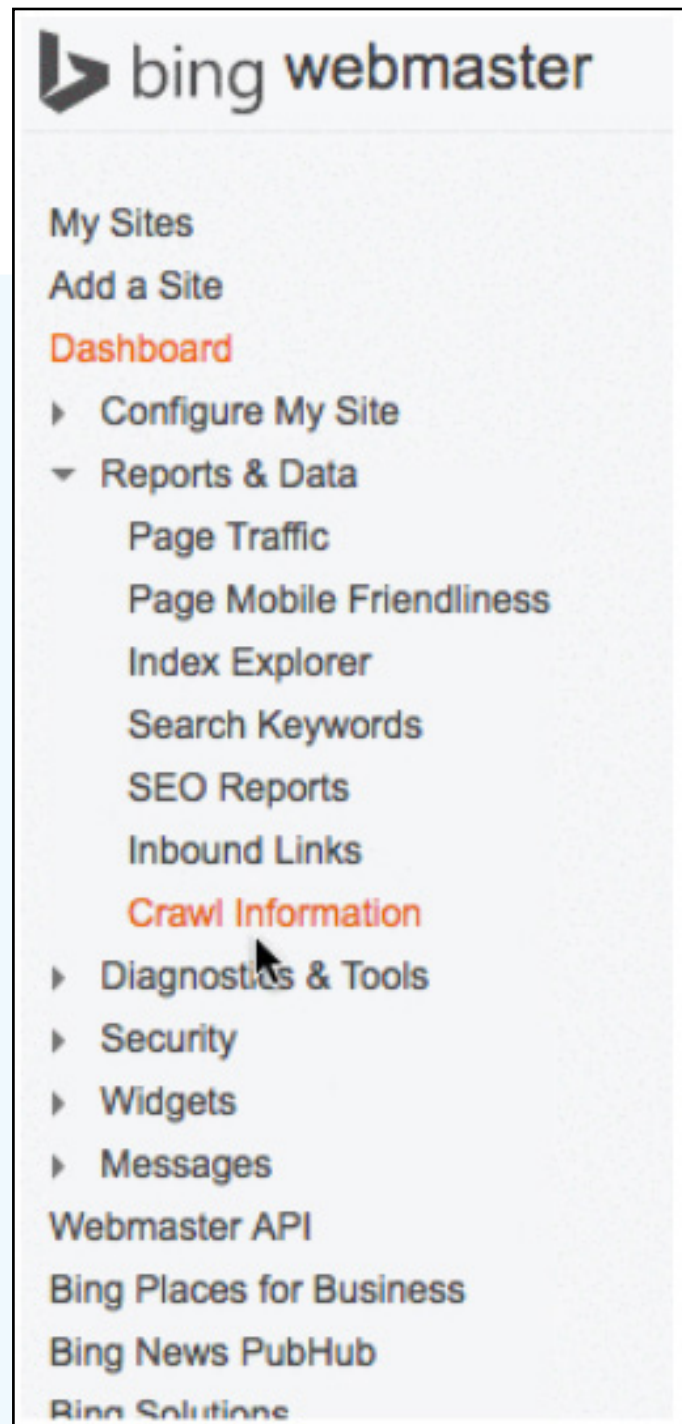
The screenshot shows the Google Search Console interface. The 'Coverage' tab is selected, and the 'Details' view is active. The table below lists various crawl errors with their status, type, validation status, trend, and the number of pages affected.

Status	Type	Validation	Trend	Pages
Excluded	Excluded by 'noindex' tag	N/A	—	1,282
Excluded	Crawled - currently not indexed	N/A	—	921
Excluded	Crawl anomaly	N/A	—	666
Excluded	Soft 404	N/A	—	20
Excluded	Not found (404)	N/A	—	14
Excluded	Page with redirect	N/A	—	5
Excluded	Alternate page with proper canonical tag	N/A	—	5
Excluded	Duplicate without user-selected canonical	N/A	—	4
Excluded	Duplicate, Google chose different canonical than user	N/A	—	2

To check this in Bing Webmaster Tools, go to 'Reports & Data' > 'Crawl Information'.

How to fix:

- Manually check your crawl errors to determine if there are crawl errors coming from old products that don't exist anymore or if you see crawl errors that should be disallowed in the robots.txt file.
- Once you've determined where they are coming from, you can implement 301 redirects to similar pages that link to the dead pages.
- You'll also want to cross-check the crawl stats in Google Search Console with average load time in Google Analytics to see if there is a correlation between time spent downloading and the pages crawled per day.



Structured Data

As mentioned above in the schema section of Screaming Frog, you can review your client's schema markup in Google Search Console.

Use the individual rich results status report in Google Search Console. (Note: The structured data report is no longer available).

This will help you determine what pages have structured data errors that you'll need to fix down the road.

How to fix:

- Google Search Console will tell you what is missing in the schema when you test the live version.
- Based on your error codes, rewrite the schema in a text editor and send to the web development team to update. I use Sublime Text for my text editing. Mac users have one built-in and PC users can use [TextPad](#).

Step 3: Review Google Analytics

Tools:

- Google Analytics
- Google Tag Manager Assistant Chrome Extension
- Annie Cushing Campaign Tagging Guide

Views

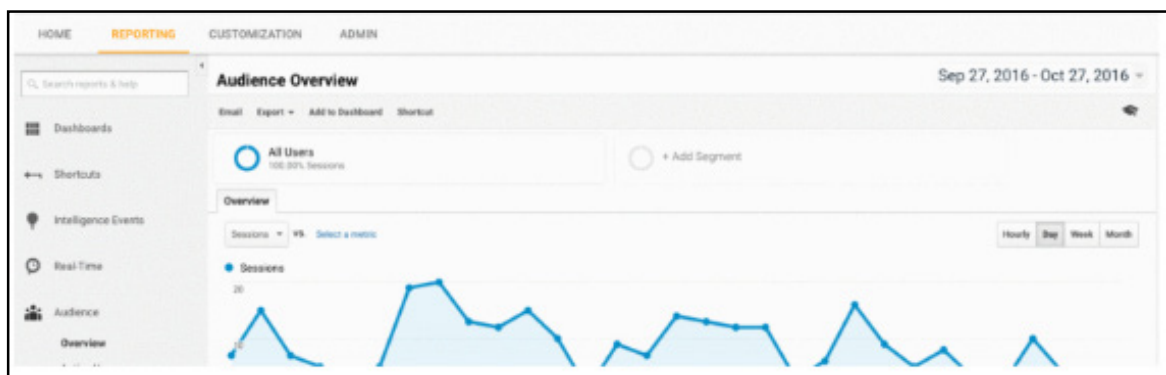
When I first get a new client, I set up 3 different views in Google Analytics.

- Reporting view
- Master view
- Test view

These different views give me the flexibility to make changes without affecting the data.

How to fix:

- In Google Analytics, go to 'Admin' > 'View' > 'View Settings' to create the three different views above.

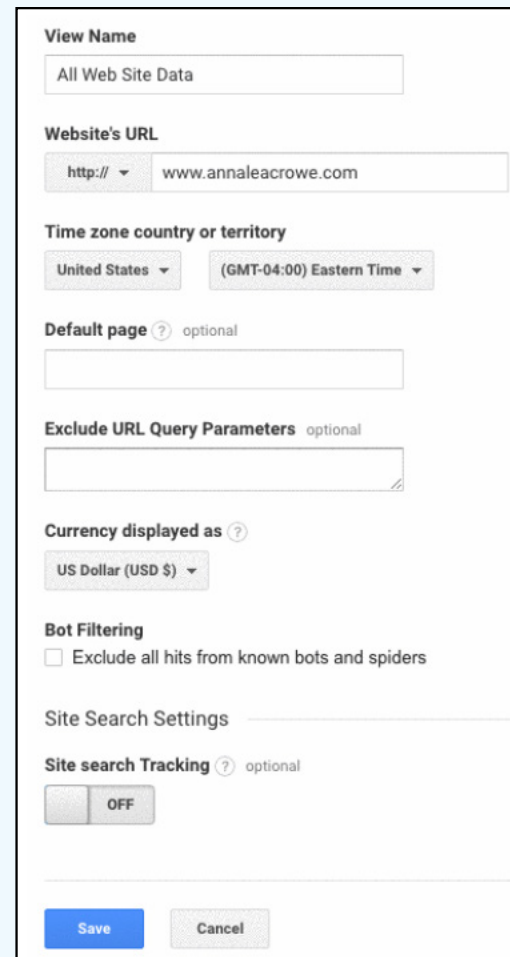


- Make sure to check the 'Bot Filtering' section to exclude all hits from bots and spiders.
- Link AdWords and Google Search Console.
- Lastly, make sure the 'Site search Tracking' is turned on.

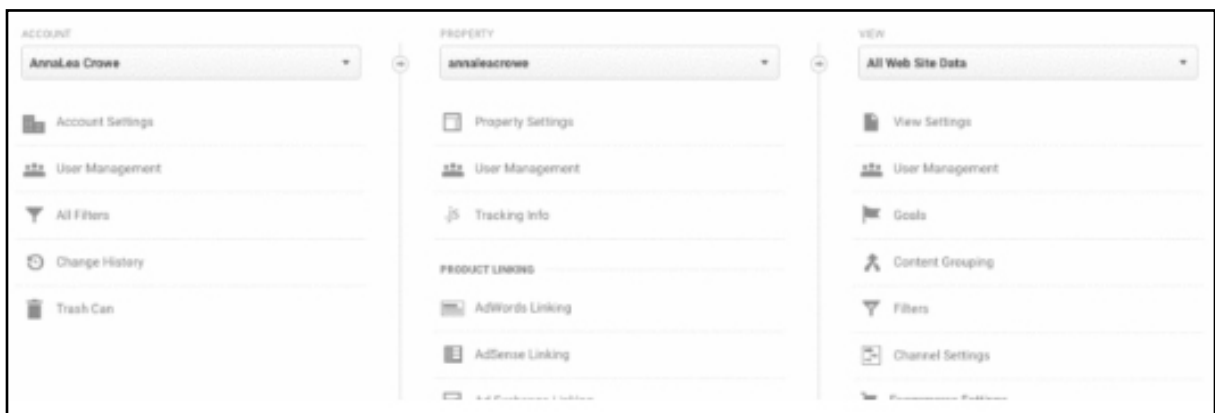
You want to make sure you add your IP address and your client's IP address to the filters in Google Analytics so you don't get any false traffic.

How to fix:

- Go to 'Admin' > 'View' > 'Filters' Then, the settings should be set to 'Exclude' > 'traffic from the IP addresses > 'that are equal to.'



The screenshot shows the 'View Settings' form in Google Analytics. The 'View Name' is 'All Web Site Data'. The 'Website's URL' is 'http:// www.annaleacrowe.com'. The 'Time zone country or territory' is 'United States' and '(GMT-04:00) Eastern Time'. The 'Default page' is optional and empty. The 'Exclude URL Query Parameters' is optional and empty. The 'Currency displayed as' is 'US Dollar (USD \$)'. The 'Bot Filtering' section has an unchecked checkbox for 'Exclude all hits from known bots and spiders'. The 'Site Search Settings' section has 'Site search Tracking' set to 'OFF'.

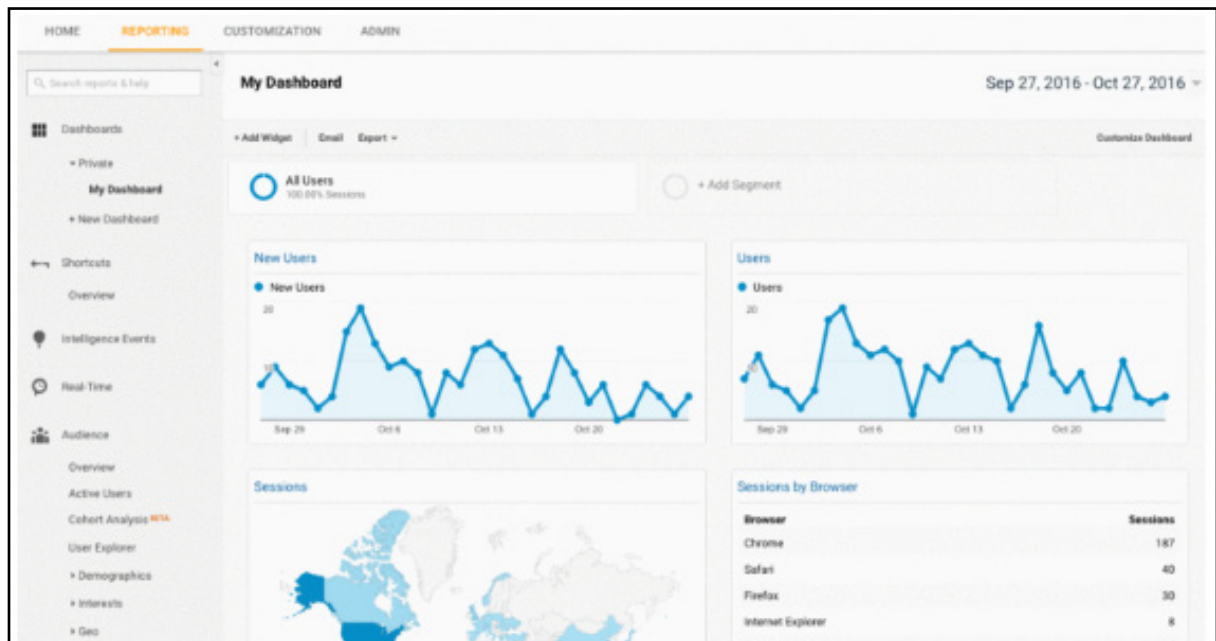


Tracking Code

You can manually check the source code, or you can use my Screaming Frog technique from above.

If the code is there, you'll want to track that it's firing real-time.

- To check this, go to your client's website and click around a bit on the site.
- Then go to Google Analytics > 'Real-Time' > 'Locations,' your location should populate.



If you're using Google Tag Manager, you can also check this with the Google Tag Assistant Chrome extension.

How to fix:

- If the code isn't firing, you'll want to check the code snippet to make sure it's the correct one. If you're managing multiple sites, you may have added a different site's code.
- Before copying the code, use a text editor, not a word processor to copy the snippet onto the website. This can cause extra characters or whitespace.
- The functions are case-sensitive so check to make sure everything is lowercase in code.

Indexing

If you had a chance to play around in Google Search Console, you probably noticed the 'Coverage' section. When I'm auditing a client, I'll review their indexing in Google Search Console compared to Google Analytics.

Here's how:

- In Google Search Console, go to 'Coverage'
- In Google Analytics, go to 'Acquisition' > 'Channels' > 'Organic Search' > 'Landing Page.'



- Once you're here, go to 'Advanced' > 'Site Usage' > 'Sessions' > '9.'

The screenshot shows the Google Analytics interface with the following data:

Landing Page	Acquisition			Behavior			Conversions eCommerce		
	Sessions	% New Sessions	New Users	Source Rate	Pages / Session	Avg. Session Duration	Ecommerce Conversion Rate	Transactions	Revenue
	48,376 <small>% of Total: 58.78% (95,419)</small>	65.75% <small>Avg for View: 61.77% (7.49%)</small>	31,807 <small>% of Total: 54.99% (58,365)</small>	53.24% <small>Avg for View: 59.03% (6.42%)</small>	5.67 <small>Avg for View: 5.98 (-5.17%)</small>	00:03:41 <small>Avg for View: 00:03:55 (-5.97%)</small>	15.22% <small>Avg for View: 16.59% (-7.80%)</small>	7,361 <small>% of Total: 46.75% (15,747)</small>	\$891,564.34 <small>% of Total: 46.11% (\$1,933,363.59)</small>

How to fix:

- Compare the numbers from Google Search Console with the numbers from Google Analytics, if the numbers are widely different, then you know that even though the pages are getting indexed only a fraction are getting organic traffic.

■ Campaign Tagging

The last thing you'll want to check in Google Analytics is if your client is using campaign tagging correctly. You don't want to not get credit for the work you're doing because you forgot about campaign tagging.

How to fix:

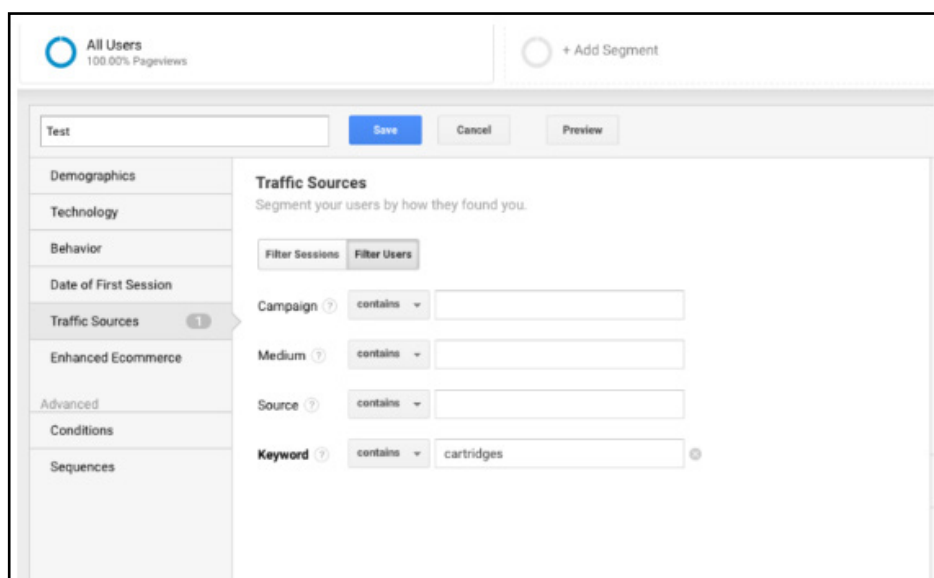
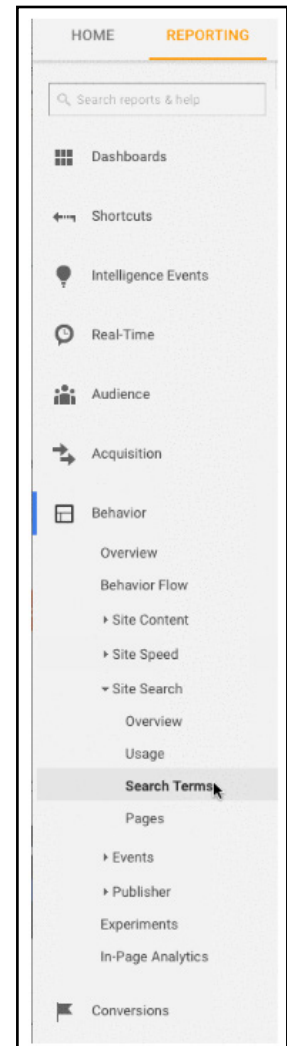
- Set up a campaign tagging strategy for Google Analytics and share it with your client. Annie Cushing put together an [awesome campaign tagging guide](#).
- Set up [Event Tracking](#) if your client is using mobile ads or video.

Keywords

You can use Google Analytics to gain insight into potential keyword gems for your client.

To find keywords in Google Analytics, follow these steps:

- Go to Google Analytics > 'Behavior' > 'Site Search' > 'Search Terms.' This will give you a view of what customers are searching for on the website.
- Next, I'll use those search terms to create a 'New Segment' in Google Analytics to see what pages on the site are already ranking for that particular keyword term.



Step 4: Manual Check

Tools:

- Google Analytics
- Access to client's server and host
- You Get Signal
- Pingdom
- PageSpeed Tools
- Wayback Machine

1 Version of Your Client's Site is Searchable

Check all the different ways you could search for a website.

For example:

- <http://annaisaunicorn.com>
- <https://annaisaunicorn.com>
- <http://www.annaisaunicorn.com>

As Highlander would say, "[there can be only one](#)" website that is searchable.

How to fix:

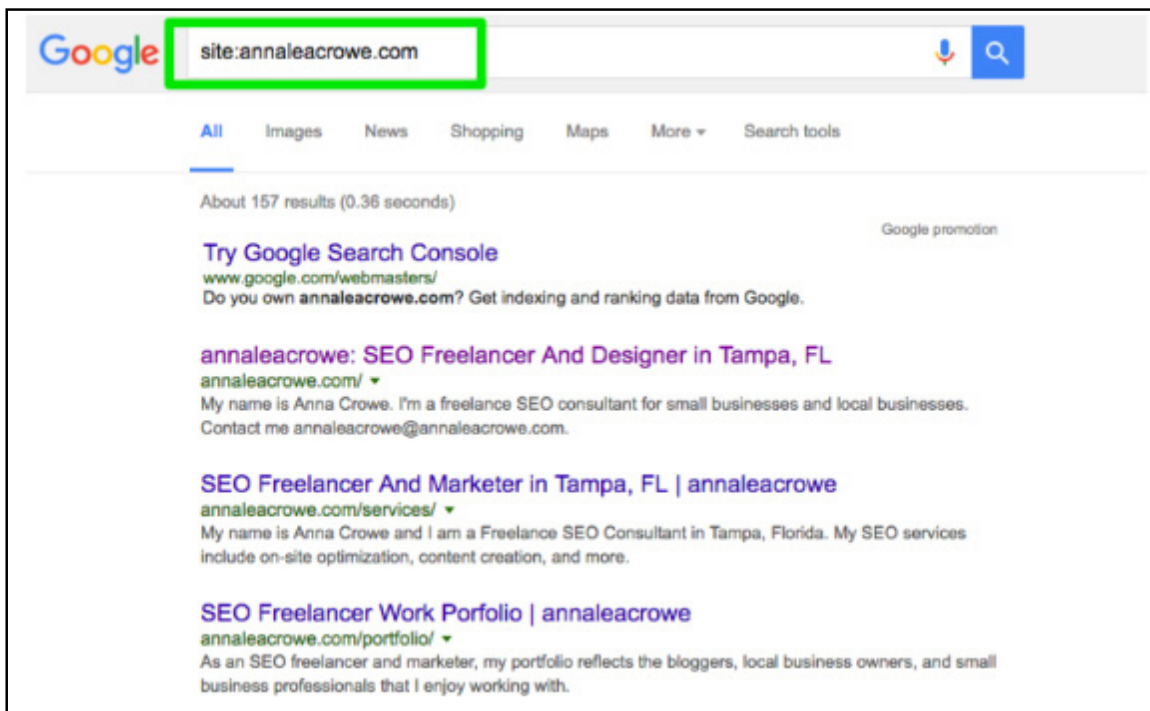
- Use a 301 redirect for all URLs that are not the primary site to the canonical site.

Indexing

Conduct a manual search in Google and Bing to determine how many pages are being indexed by Google. This number isn't always accurate with your Google Analytics and Google Search Console data, but it should give you a rough estimate.

To check, do the following:

- Perform a site search in the search engines.



- When you search, manually scan to make sure only your client's brand is appearing.
- Check to make sure the homepage is on the first page. John Mueller said it isn't necessary for the **homepage to appear as the first result.**

How to fix:

- If another brand is appearing in the search results, you have a bigger issue on your hands. You'll want to dive into the analytics to diagnose the problem.
- If the homepage isn't appearing as the first result, perform a manual check of the website to see what it's missing. This could also mean the site has a penalty or poor site architecture which is a bigger site redesign issue.
- Cross-check the number of organic landing pages in Google Analytics to see if it matches the number of search results you saw in the search engine. This can help you determine what pages the search engines see as valuable.

■ Caching

I'll run a quick check to see if the top pages are being cached by Google. Google uses these cached pages to connect your content with search queries.

To check if Google is caching your client's pages, do this:

<http://webcache.googleusercontent.com/search?q=cache:https://www.searchenginejournal.com/pubcon-day-3-women-in-digital-amazon-analytics/176005/>
Make sure to toggle over to the 'Text-only version.'

You can also check this in Wayback Machine.

How to fix:

- Check the client's server to see if it's down or operating slower than usual. There might be an internal server error or a database connection failure. This can happen if multiple users are attempting to access the server at once.
- Check to see who else is on your server with a reverse IP address check. You can use You Get Signal website for this phase. You may need to upgrade your client's server or start using a CDN if you have sketchy domains sharing the server.
- Check to see if the client is removing specific pages from the site.

■ Hosting

While this may get a little technical for some, it's vital to your SEO success to check the hosting software associated to your client's website. Hosting can harm SEO and all your hard work will be for nothing.

You'll need access to your client's server to manually check any issues. The most common hosting issues I see are having the wrong TLD and slow site speed.

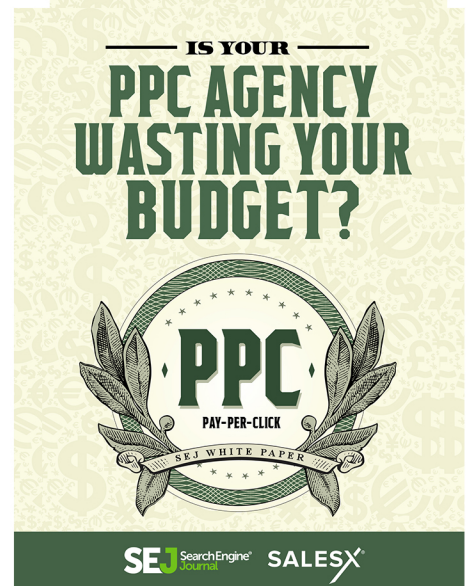
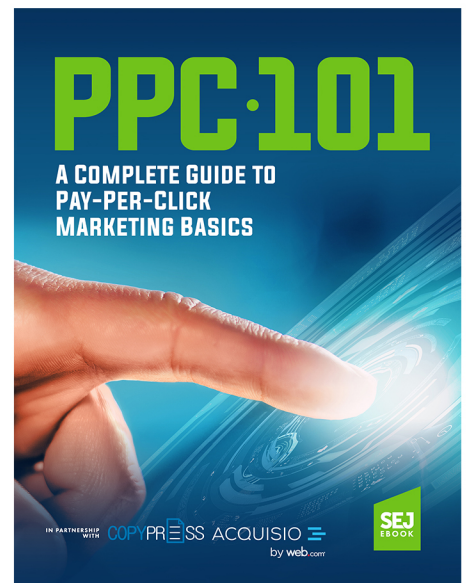
How to fix:

- If your client has the wrong TLD, you need to make sure the country IP address is associated with the country your client is operating in the most. If your client has a .co domain and also a .com domain, then you'll want to redirect the .co to your client's primary domain on the .com.
- If your client has slow site speed, you'll want to address this quickly because **site speed is a ranking factor**. Find out what is making the site slow with tools like PageSpeed Tools and Pingdom. Here's a look at some of the common page speed issues:
 - Host
 - Large images
 - Embedded videos
 - Plugins
 - Ads
 - Theme
 - Widgets
 - Repetitive script or dense code

Over to You!

I'm excited to see you test out DeepCrawl, Screaming Frog, and some of the other tools. And, I'd love to hear about all the creative ways you perform a site audit. What have you experimented with? What tools do you use? Let me know in the comments below.

This is a series of posts which I'll be diving deeper into mobile, site architecture, site speed, content, and off-site. If there's anything particular you want to see, let me know in the comments.



10+ MUST-READ EBOOKS FOR SEO PROFESSIONALS & DIGITAL MARKETERS

Get all of Search Engine Journal's in-depth and free guides and ebooks, covering all things SEO, PPC, content marketing, and social media marketing.

WANT TO ADVERTISE ON OUR EBOOK?

Email jessica@alphabrandmedia.com to learn about our ebook sponsorship options.

