(Almost) Everything You Need To Know About Crawling, Indexing, and Rendering in Google

Barry Adams
June 2022
What does Google do?
1. Crawler (Googlebot)
Crawling: Discovery

<a href="...">

<urlset>
  <url>
    <loc>https://friendsofsearch.com/</loc>
    <lastmod>2022-06-09T11:42:50+00:00</lastmod>
  </url>
</urlset>

@badams #FOS22
Crawling: Queue Management

URL Deduplication

@badams #FOS22
Crawling: Queue Management

Prioritisation & Scheduling

1. __________
2. __________
3. __________
4. __________
5. __________

@badams #FOS22
Crawling: Fetch & Parse
Crawl Politeness

@badams #FOS22
Optimise Crawling

- Server Response Time

@badams #FOS22
GSC Crawl Stats

By response

- OK (200): 77%
- Not modified (304): 21%
- Moved permanently (301): 1%
- Not found (404): < 1%
- Moved (other): < 1%

By file type

- JavaScript: 60%
- HTML: 16%
- JSON: 1%
- CSS: < 1%
- Image: < 1%

By purpose

- Refresh: 99%
- Discovery: < 1%

By Googlebot type

- Page resource load: 63%
- Smartphone: 35%
- Desktop: 2%
- Image: < 1%
- AdsBot: < 1%
<table>
<thead>
<tr>
<th>Resource Type</th>
<th>% Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page resource load</td>
<td>63%</td>
</tr>
<tr>
<td>Smartphone</td>
<td>35%</td>
</tr>
<tr>
<td>Desktop</td>
<td>3%</td>
</tr>
<tr>
<td>Image</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>AdsBot</td>
<td>&lt; 1%</td>
</tr>
</tbody>
</table>

Rows per page: 5  1-5 of 7  

@badams #FOS22
## By Googlebot type

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdsBot</td>
<td>66%</td>
</tr>
<tr>
<td>Smartphone</td>
<td>22%</td>
</tr>
<tr>
<td>Desktop</td>
<td>11%</td>
</tr>
<tr>
<td>Page resource load</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>Image</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Rows per page: 5

1-5 of 6
Optimise Crawling

• Serve correct HTTP status codes;

  ➢ 200 OK
  ➢ 301 / 302 Redirects
  ➢ 304 Not Modified
  ➢ 401 / 403 Permission Issues
  ➢ 404 / 410 Not Found/Gone
  ➢ 5xx Error
Optimise Crawling

- ALL resources consume crawl budget;
  - Not just HTML pages
  - Reduce HTTP requests per page
Optimise Crawling

• ALL resources consume crawl budget;
  ➢ Not just HTML pages
  ➢ Reduce HTTP requests per page

• AdsBot can consume crawl budget;
  ➢ Double-check your Google Ads campaigns
Optimise Crawling

• ALL resources consume crawl budget;
  ➢ Not just HTML pages
  ➢ Reduce HTTP requests per page

• AdsBot can consume crawl budget;
  ➢ Double-check your Google Ads campaigns

• Link equity (PageRank) impacts crawl budget;
  ➢ More link equity = more crawl budget
2. Indexer

Crawler → Indexer → Ranker
Two Stages* of Indexing

Crawler → Indexer → Ranker

*At least – indexing is a collection of interconnected processes

@badams #FOS22
Indexing: HTML Lexer & Tokenizer

Text
“The cat sat on the mat.”

↓

Tokens
“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

↓

Vector encoding of the tokens

<table>
<thead>
<tr>
<th>0.0</th>
<th>0.0</th>
<th>0.4</th>
<th>0.0</th>
<th>0.0</th>
<th>1.0</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>1.0</td>
<td>0.2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

the cat sat on the mat.
Indexing: Selection
Lily Ray (Amsive Digital): ‘Any SEO who hasn’t jumped onto the E-A-T wave is not playing close enough attention’. A bold and bold statement, but true if you believe ‘E-A-T should serve as the backbone to a’.

Rodney Ip (Google): ‘Performance Max helps our advertisers stay ahead of ongoing changes in consumer behaviour’.
Indexing: Rendering

- URLs
  - Crawl Queue
  - Crawler
    - URL
  - Processing
    - HTML
  - Index
  - Render Queue
    - Rendered HTML
  - Renderer

@badams #FOS22
Indexing: Index Integrity

Deduplication & Canonicalisation

@badams #FOS22
Rendering
Evergreen Chrome

Tuesday, May 07, 2019

Googlebot is the crawler that visits web pages to include them within Google Search index. The number one question we got from the community at events and social media was if we could make Googlebot evergreen with the latest Chromium. Today, we are happy to announce that Googlebot now runs the latest Chromium rendering engine (74 at the time of this post) when rendering pages for Search. Moving forward, Googlebot will regularly update its rendering engine to ensure support for latest web platform features.

What that means for you

Compared to the previous version, Googlebot now supports 1000+ new features, like:

- ES6 and newer JavaScript features
- IntersectionObserver for lazy-loading
- Web Components v1 APIs

You should check if you’re transpiling or use polyfills specifically for Googlebot and if so, evaluate if this is still necessary. There are still some limitations, so check our troubleshooter for JavaScript-related issues and the video series on JavaScript SEO.

@badams #FOS22
What happens during Rendering in your Browser?

HTML → HTML Parser → DOM Tree → Layout → Render Tree → Painting → Display

CSS → CSS Parser → CSSOM
JavaScript

- HTML
  - HTML Parser
  - DOM Tree
  - JavaScript
  - CSSOM
- CSS
  - CSS Parser

- Layout
- Render Tree
- Painting
- Display

@badams #FOS22
JavaScript…
JavaScript...
Google’s Rendering as part of Indexing

- HTML
  - HTML Parser
  - DOM Tree
  - CSS Parser
  - CSSOM
  - Render Tree
  - JavaScript
  - Layout
  - Painting
  - Display

JavaScript
CSS

@badams #FOS22
Google does not perform actions

```html
<button onclick="myFunction()">Click me</button>
```

```html
<div onscroll="myFunction()"/>
```
Why Rendering?
Raw HTML:
Als professional stelt u hoge eisen aan de bronnen die u raadpleegt. De Collecties van Wolters Kluwer bieden u de meest complete informatie. Collecties zijn onderverdeeld in wet- en regelgeving, jurisprudentie (incl. artikelen), vakliteratuur en meer. Er zijn Collecties voor iedereen die in de juridische wereld terechtkomt.

Navigator is een van de belangrijkste platformen voor juristen en fiscalisten. Met hoogwaardige juridische content en een persoonlijke zoekervaring biedt Navigator de meeste relevante bronnen in één plaats.
Rendering allows Google to...

• ... load all meta data, content, and links on a webpage
• ... understand the page’s layout and content hierarchy
• ... evaluate the usability and quality of the webpage
Rendering Issues
Possible Rendering Issues in GSC

Coverage > Soft 404
- Status: Excluded
- Affected pages: 242K

Coverage > Submitted URL seems to be a Soft 404
- Submitted URL seems to be a Soft 404
- First detected: 20/12/2017
- Status: Error
- Affected pages: 4.61K

Coverage > Page indexed without content
- Status: Warning
- Affected pages: 1
Rendering Issues

- Inaccessible Resources;
  - Make sure all page resources can be crawled
Rendering Issues

- JavaScript inserts invalid HTML in the `<head>`;
  - `<body>` tags in the `<head>` break Google’s processing of meta tags

```html
<!doctype html>
<html lang="en">

<head>
  <meta name="description" content="ACME Anvils are great for all your skull crushing needs. From annoying road runners to opinionated rabbits, our anvils are guaranteed to hit you in the face." />
  <title>ACME Anvils</title>
  <script>
    Some JavaScript
  </script>
  <link href="https://www.acme.com/skull-crushing/anvils/" hreflang="en-us" rel="alternate" />
  <link href="https://www.acme.com/uk/skull-crushing/anvils/" hreflang="en-gb" rel="alternate" />
  <link href="https://www.acme.com/ie/skull-crushing/anvils/" hreflang="en-ie" rel="alternate" />
  <link href="https://www.acme.com/skull-crushing/anvils/" hreflang="x-default" rel="alternate" />
</head>

<body>
  ...
</body>
</html>
```
Rendering Issues

• JavaScript inserts invalid HTML in the <head>;
  ➢ <body> tags in the <head> break Google’s processing of meta tags

```html
<!doctype html>
<html lang="en">
  <head>
    <meta name="description" content="ACME Anvils are great for all your skull crushing needs. From annoying road runners to opinionated rabbits, our anvils are guaranteed to hit you in the face.">
    <title>ACME Anvils</title>
    <iframe src="...."></iframe>
    <link href="https://www.acme.com/skull-crushing/anvils/" hreflang="en-us" rel="alternate" />
    <link href="https://www.acme.com/uk/skull-crushing/anvils/" hreflang="en-gb" rel="alternate" />
    <link href="https://www.acme.com/ie/skull-crushing/anvils/" hreflang="en-ie" rel="alternate" />
    <link href="https://www.acme.com/skull-crushing/anvils/" hreflang="x-default" rel="alternate" />
  </head>
  <body>
    ...
  </body>
</html>
```
Don't use invalid elements in the `<head>`

The following elements are invalid when used in the `<head>`, and therefore aren't supported by Google Search when placed in the `<head>`:

- `<iframe>`
- `<img>`
- Any other HTML element

We strongly recommend that you don't use these invalid elements in the `<head>`, but if you must, place these invalid elements after the ones you want Google to see. Once Google detects one of these invalid elements, it assumes the end of the `<head>` and stops reading any further elements in the `<head>`.

https://developers.google.com/search/docs/advanced/guidelines/valid-html

@badams #FOS22
Rendering Issues

• HTML vs Render mismatch;
  ➢ Different content in raw HTML vs fully rendered page
https://chrome.google.com/webstore/detail/view-rendered-source/ejgngohbdeoabanmclafpkoogegdpob

@badams #FOS22
SEO Crawlers Can Also Render

@badams #FOS22
Google Tools *ALWAYS* Render

@badams #FOS22
Optimise Rendering

• Don’t rely on Google’s rendering;
  ➢ Use SSR & CDN caching

• Minimise page weight;
  ➢ Fewer page resources = better use of crawl budget
  faster load speed & CWV
  less chance of rendering issues

• Optimise your HTML source;
  ➢ Think about where <script> tags exist and what they
    do when their code is executed

@badams #FOS22
Optimise Indexing

• Optimise your page layouts;
  ➢ Prominent content & links are more valuable for users & Google

• Improve internal linking;
  ➢ More PageRank = higher chance of indexing

• Improve your content;
  ➢ Google has **no obligation** to index all your pages
  ➢ Make it worth Google’s while...
Bypassing Rendering* with Edge SEO

*sort of
Edge SEO

Your Webserver ➔ Cloud CDNs ➔ Users

@badams #FOS22
Edge SEO

Your Webserver

Cloud CDNs

Googlebot

@badams #FOS22
Edge SEO

Your Webserver → Cloud CDNs ← Googlebot

Change your webpages here
Edge SEO

• CDNs store cached versions of your webpages;
  ➢ Global coverage with edge nodes worldwide
  ➢ Usually also results in faster crawling and better CWV

• You manipulate your CDN cached pages;
  ➢ Cloud Workers enable a range of functionality

• Googlebot crawls & indexes the changed CDN-cached pages;
  ➢ Your ‘original’ website remains unchanged
  ➢ Google only sees the changed CDN webpages
Why Edge SEO?

• Faster deployment;
  ➢ Bypass your developers’ lengthy queues
  ➢ ‘Ask forgiveness, not permission’
  ➢ No reliance on client-side JavaScript

• No CMS constraints;
  ➢ Change pages directly regardless of your CMS capabilities

• Testing;
  ➢ Perform narrow tests on specific site sections
  ➢ A/B testing for SEO
SEO A/B Split Testing

SearchPilot

RankScience

Use A/B testing to grow traffic with RankScience

Talk to sales

SplitSignal

Enterprise SEO A/B Split Testing

SplitSignal makes it easy to run meaningful A/B split tests – without taking up valuable developer or data science resources.

Request More Info

@badams FOS22
SEO Split Testing Case Studies

Do 301 Internal Links Harm SEO?

https://www.searchpilot.com/resources/newsletter/

@badams #FOS22
Barry Adams

➢ Doing SEO since 1998
➢ Specialist in Technical SEO & News SEO
➢ Newsletter: SEOforGoogleNews.com

@badams #FOS22
Thank You

barry@polemicdigital.com

@badams