FRIENDS OF SEARCH

A (Shallow) Deep-Dive into Technical SEO in 2021 and Beyond

@badams

POLEMIC DIGITAL

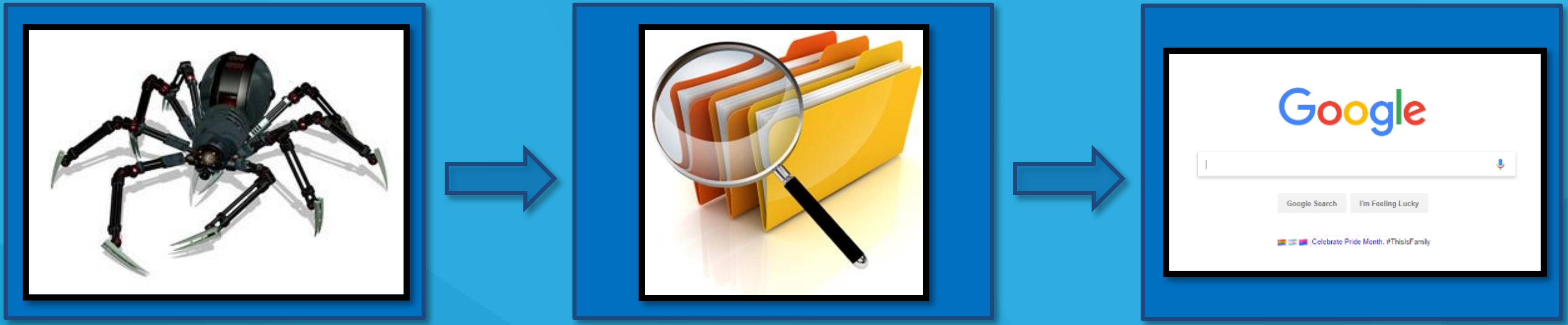
PARENTAL

ADVISORY

EXPLICIT CONTENT

What is Technical SEO?

Google Processes

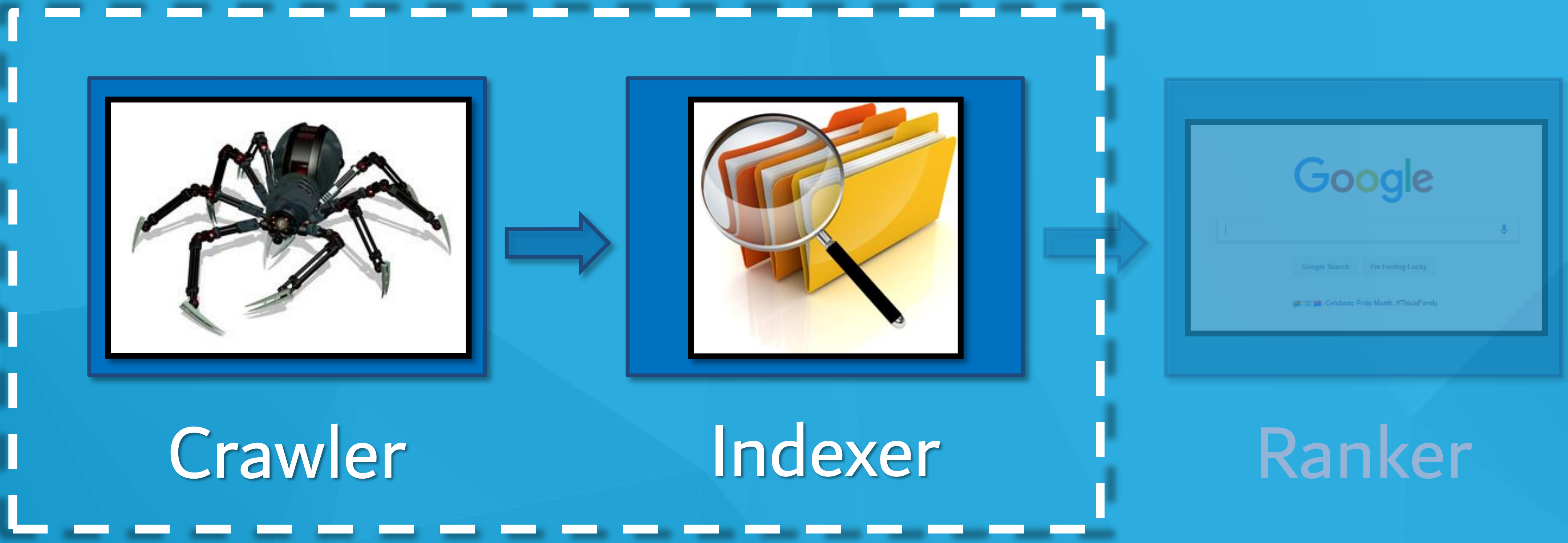


Crawler

Indexer

Ranker

Google Processes

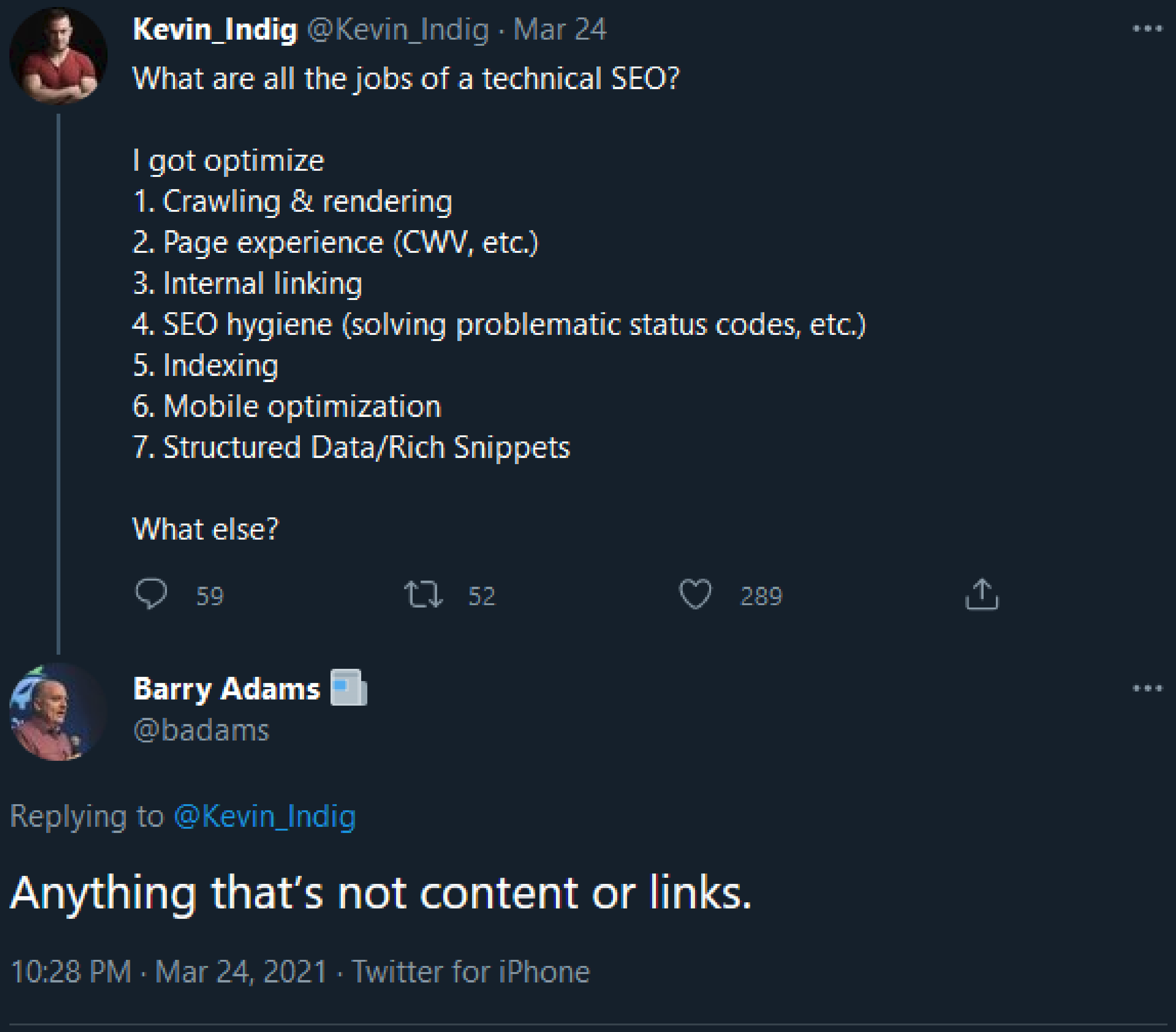


Crawler

Indexer

Ranker

Technical SEO



Kevin_Indig @Kevin_Indig · Mar 24

What are all the jobs of a technical SEO?

I got optimize
1. Crawling & rendering
2. Page experience (CWV, etc.)
3. Internal linking
4. SEO hygiene (solving problematic status codes, etc.)
5. Indexing
6. Mobile optimization
7. Structured Data/Rich Snippets

What else?

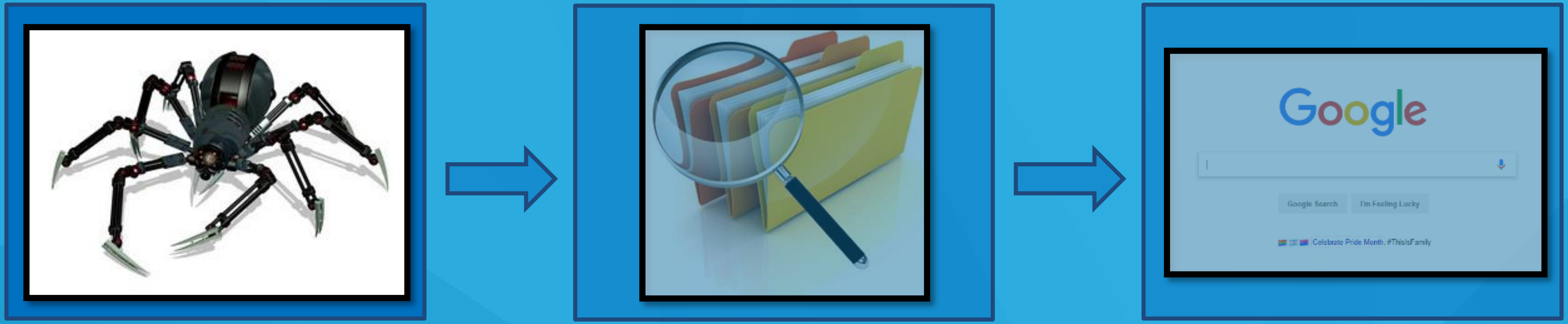
Barry Adams @badams

Replying to @Kevin_Indig

Anything that's not content or links.

10:28 PM · Mar 24, 2021 · Twitter for iPhone

1. Crawler (Googlebot)



Crawler

Indexer

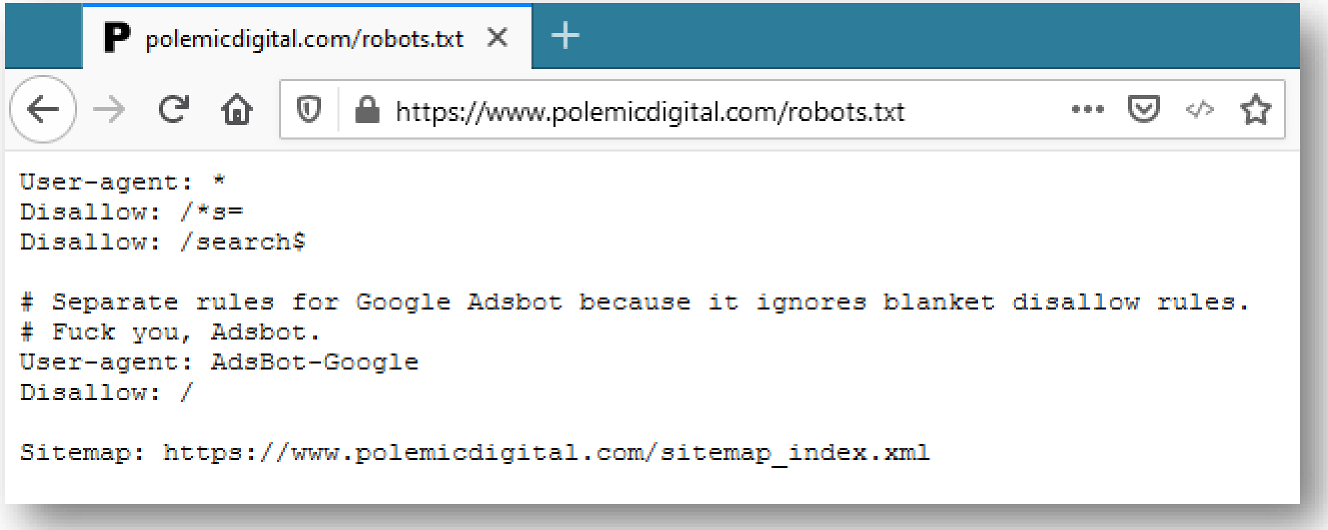
Ranker

Crawling: Googlebot

- URL discovery
  - <a href> tags in HTML
  - XML sitemaps
  - Other sources?
- Crawl queue management
  - De-duplication based on URL patterns
  - Crawl prioritisation & scheduling
- Crawling
  - Fetching raw HTML
  - Crawl ‘politeness’

Crawl Management

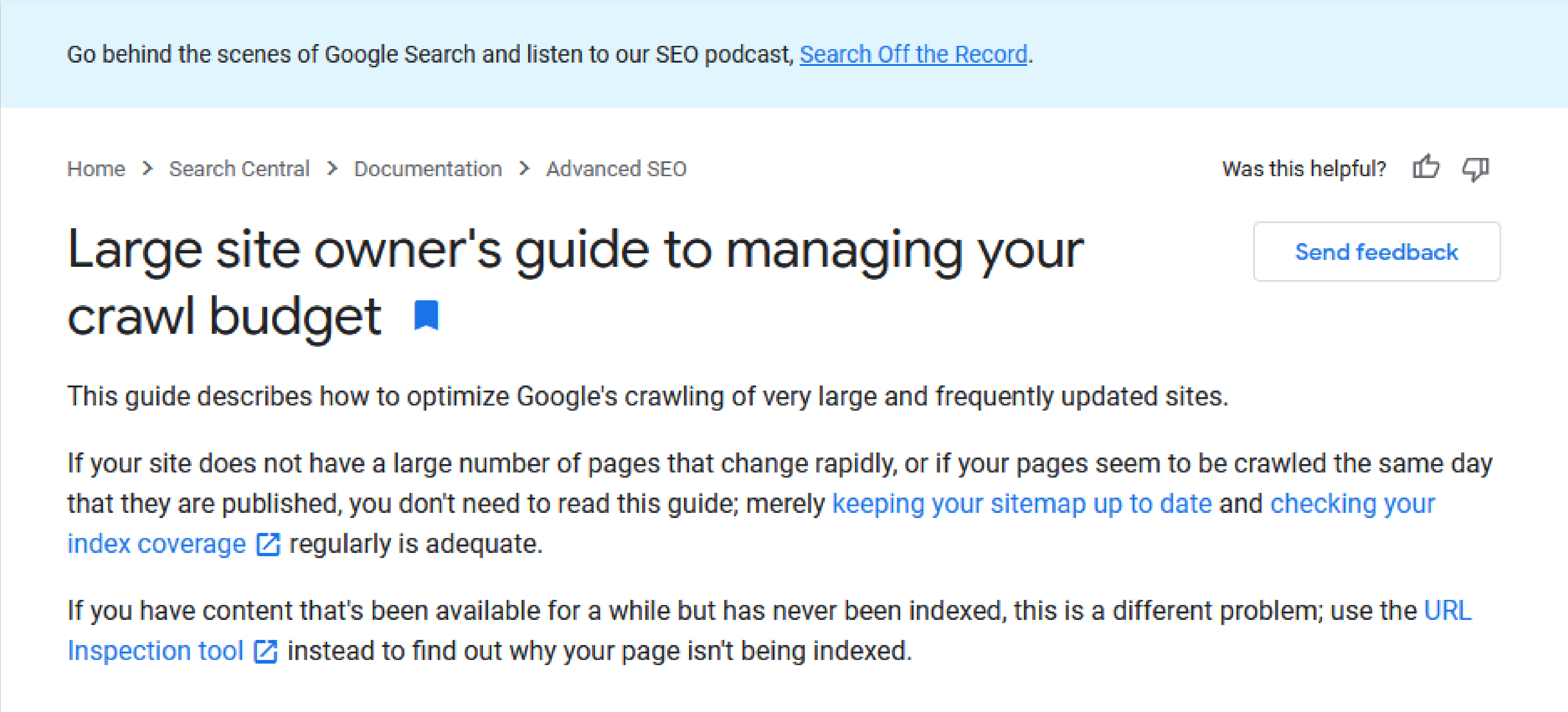
- Robots.txt Disallow
  - ➢ Strongest crawl management signal
  - ➢ Evaporates crawl budget


polemicdigital.com/robots.txt
https://www.polemicdigital.com/robots.txt

```
User-agent: *
Disallow: /*s=
Disallow: /search$

# Separate rules for Google Adsbot because it ignores blanket disallow rules.
# Fuck you, Adsbot.
User-agent: AdsBot-Google
Disallow: /

Sitemap: https://www.polemicdigital.com/sitemap_index.xml
```



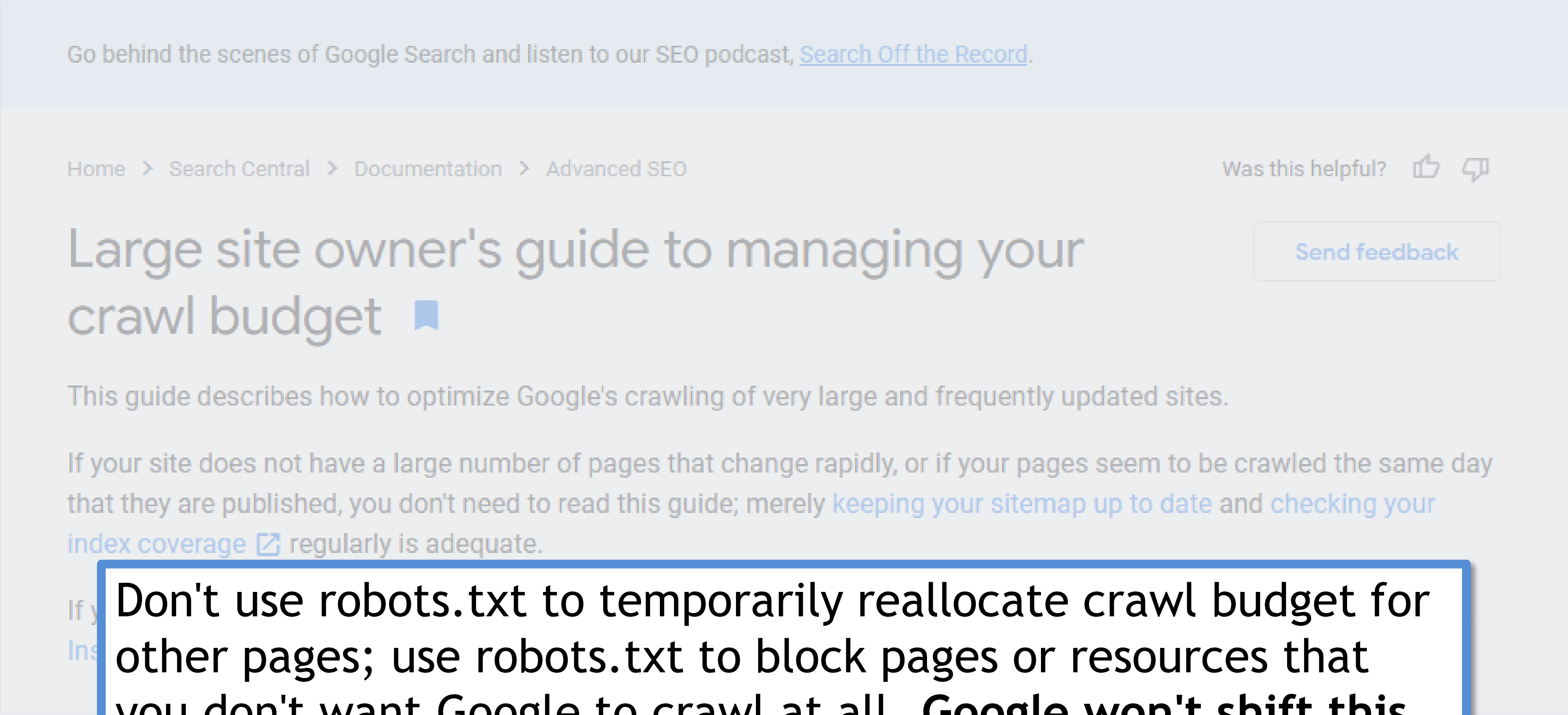
Go behind the scenes of Google Search and listen to our SEO podcast, Search Off the Record.

Large site owner's guide to managing your crawl budget

This guide describes how to optimize Google's crawling of very large and frequently updated sites.

If your site does not have a large number of pages that change rapidly, or if your pages seem to be crawled the same day that they are published, you don't need to read this guide; merely keeping your sitemap up to date and checking your index coverage regularly is adequate.

If you have content that's been available for a while but has never been indexed, this is a different problem; use the URL Inspection tool instead to find out why your page isn't being indexed.



Go behind the scenes of Google Search and listen to our SEO podcast, Search Off the Record.

Large site owner's guide to managing your crawl budget

This guide describes how to optimize Google's crawling of very large and frequently updated sites.

If your site does not have a large number of pages that change rapidly, or if your pages seem to be crawled the same day that they are published, you don't need to read this guide; merely keeping your sitemap up to date and checking your index coverage regularly is adequate.

If y

Ins

Don't use robots.txt to temporarily reallocate crawl budget for other pages; use robots.txt to block pages or resources that you don't want Google to crawl at all. **Google won't shift this newly available crawl budget to other pages** unless Google is already hitting your site's serving limit.

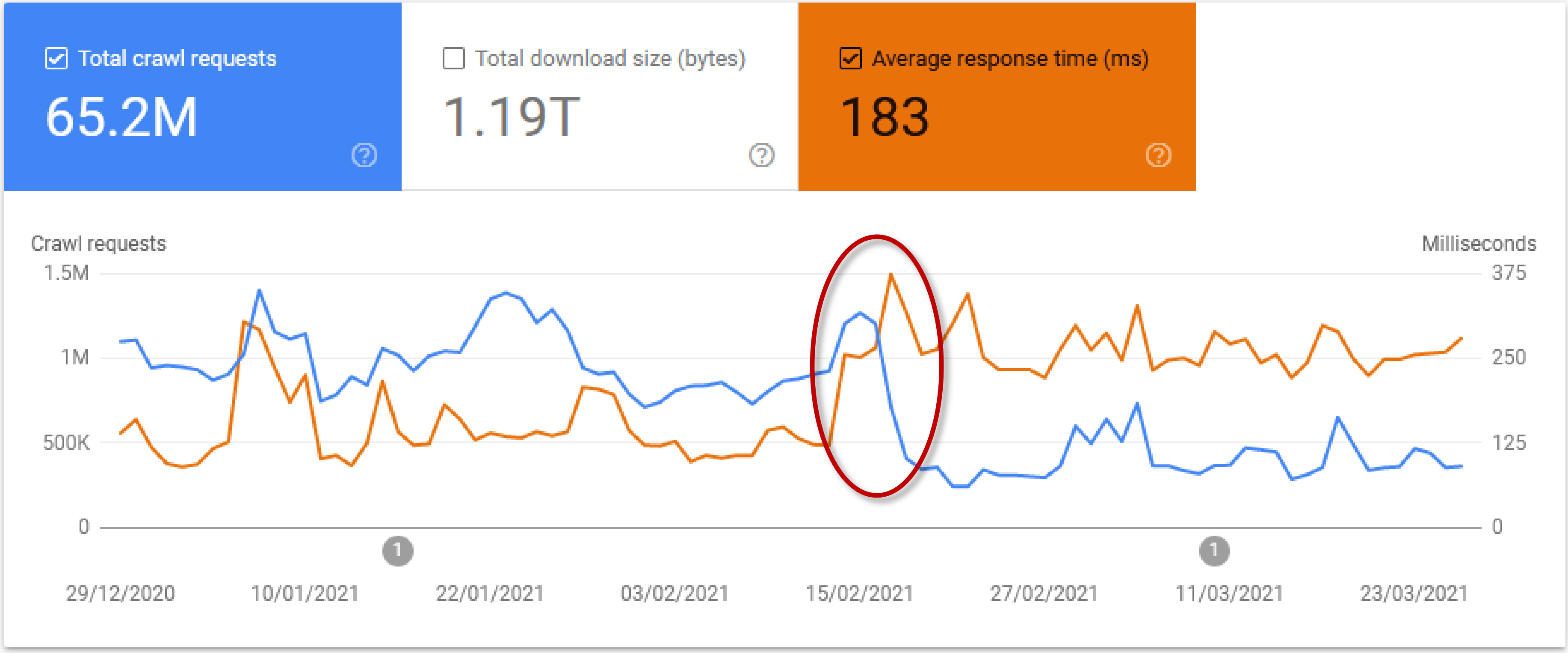
Crawl Management

- Robots.txt Disallow
  - ➢ Strongest crawl management signal
  - ➢ Evaporates crawl budget

- Canonicals & noindex are NOT crawl management
  - ➢ Google needs to **see** meta tags before it can act on them
  - ➢ That means Googlebot still crawls those URLs

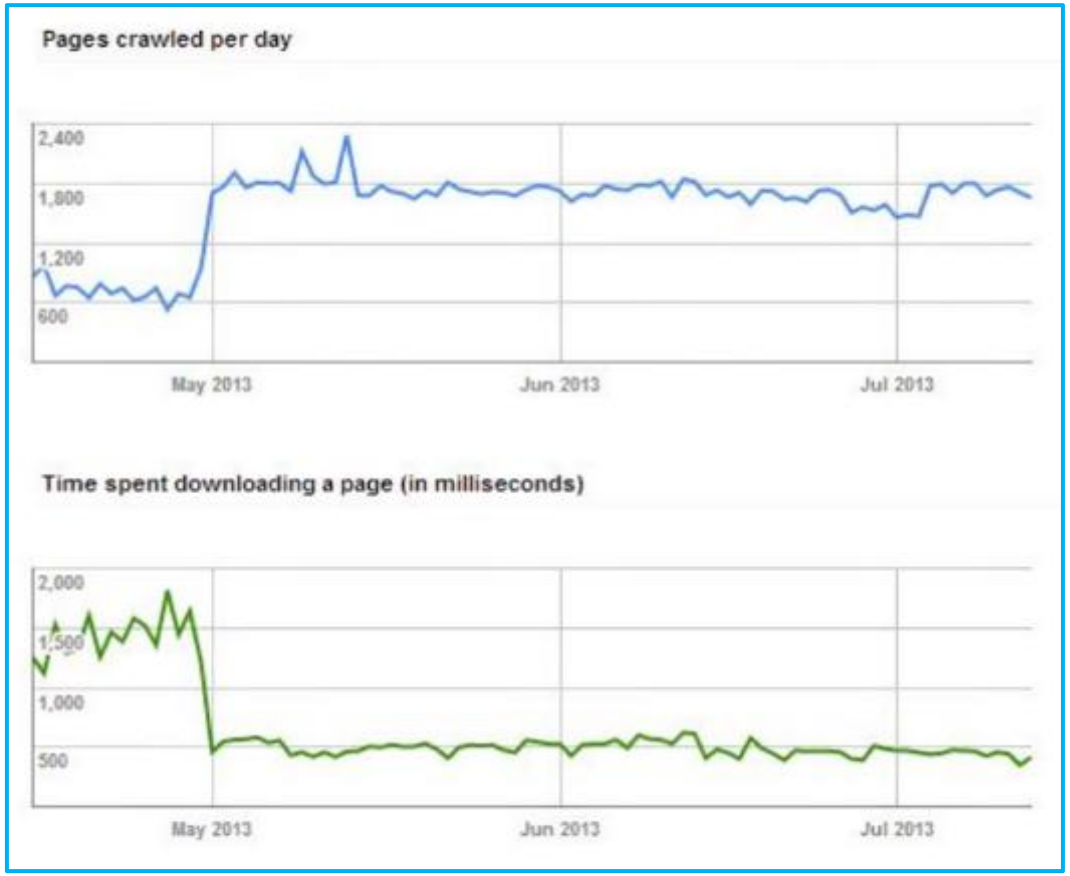
Optimise Crawling

- Server Response Time

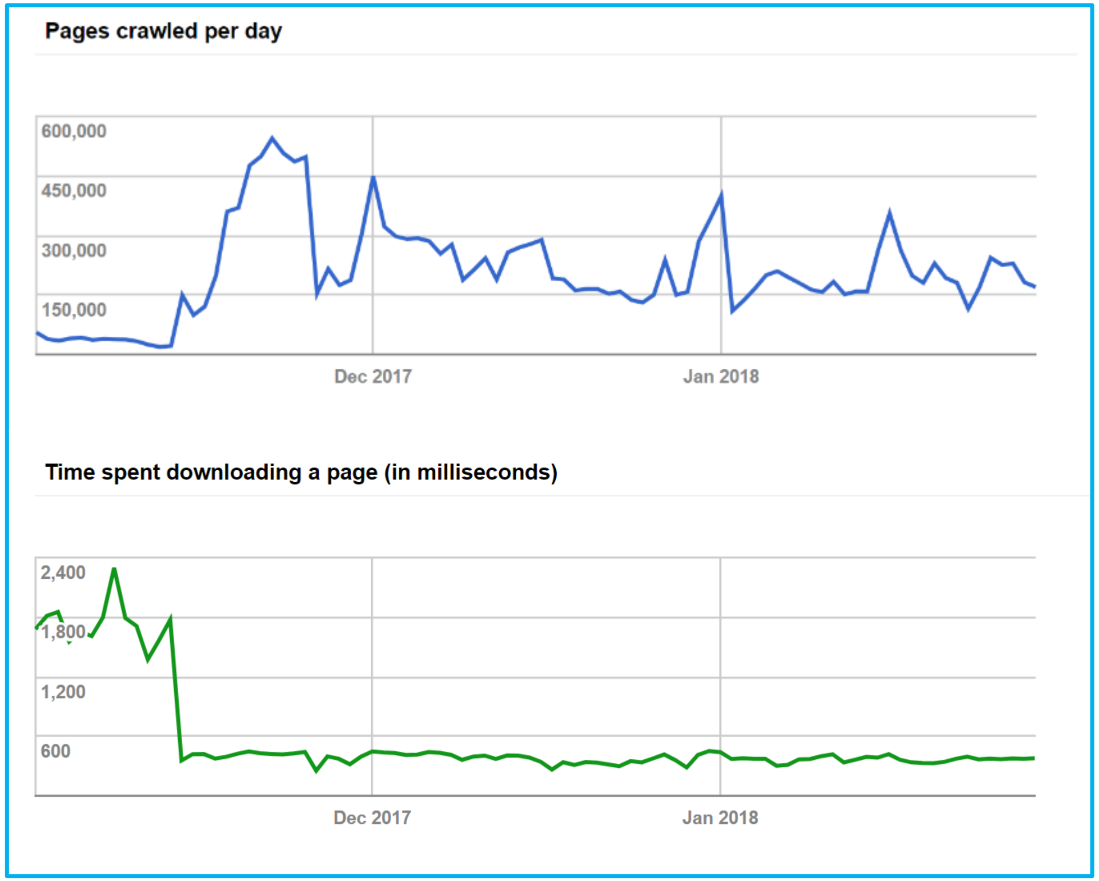


Load Speed

Fast response time = optimal use of Googlebot

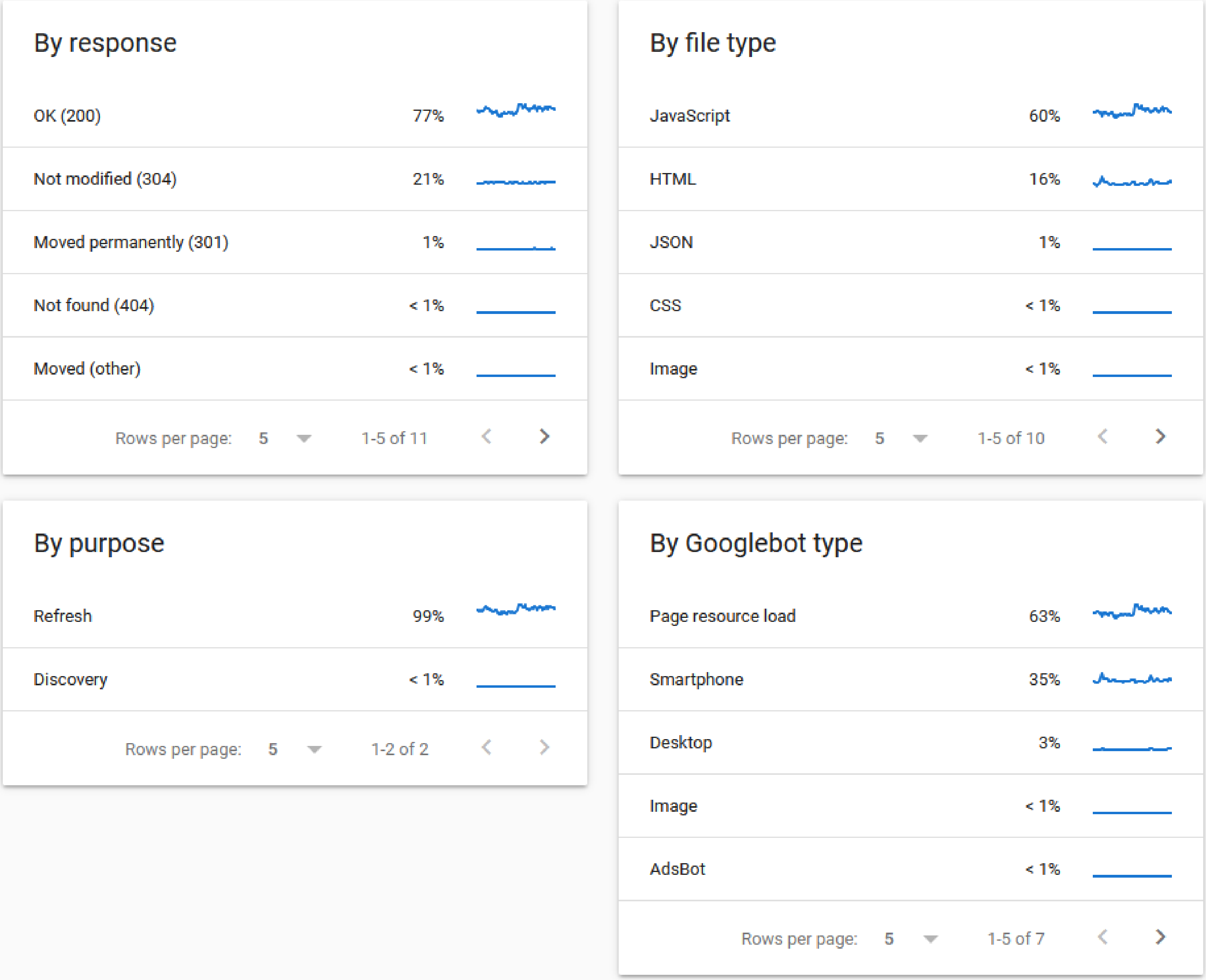

Pages crawled per day

Time spent downloading a page (in milliseconds)


Pages crawled per day

Time spent downloading a page (in milliseconds)

GSC Crawl Stats



By response

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

Rows per page: 5 1-5 of 11

By file type

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

Rows per page: 5 1-5 of 10

By purpose

Refresh	99%
Discovery	< 1%

Rows per page: 5 1-2 of 2

By Googlebot type

Page resource load	63%
Smartphone	35%
Desktop	3%
Image	< 1%
AdsBot	< 1%

Rows per page: 5 1-5 of 7

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

HTTP Status Codes

- All redirects should be **301**
  - ➢ Except geo-targeting redirects, those should be **302**
  - ➢ **304** means the URL hasn’t changed since the last crawl
  - ➢ **307** relates to HSTS preload list

HTTP Status Codes

- Accidental not found: **404**
- Deliberate deletion: **410**

HTTP Status Codes

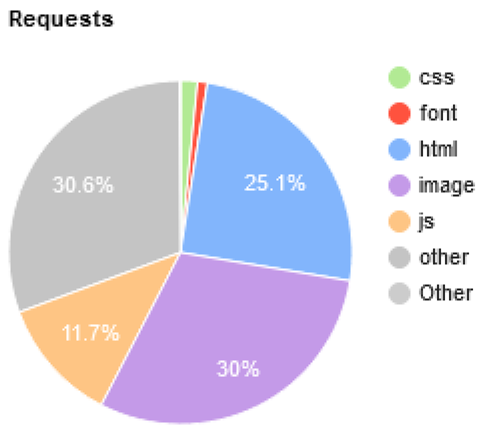
HTTP

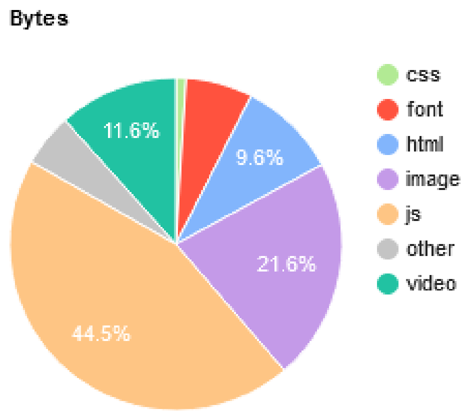
- Accidental server error: **500**
- Deliberate downtime: **503**

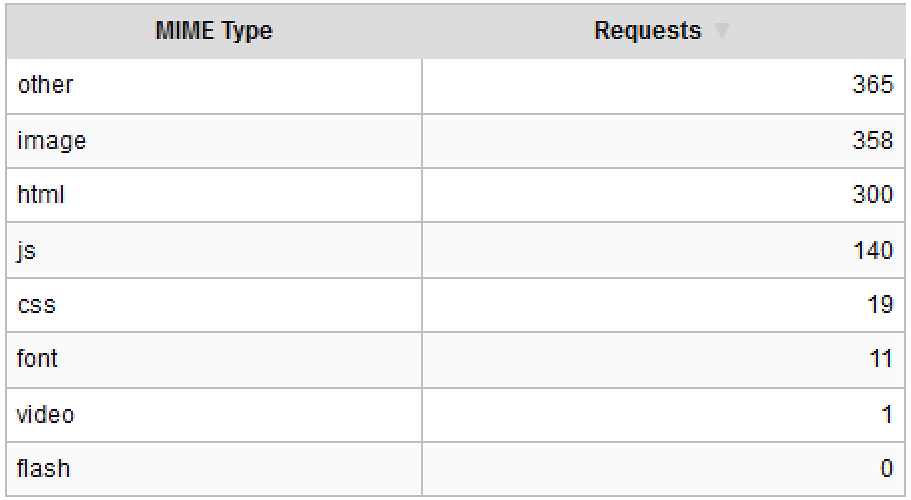
Optimise Crawling

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

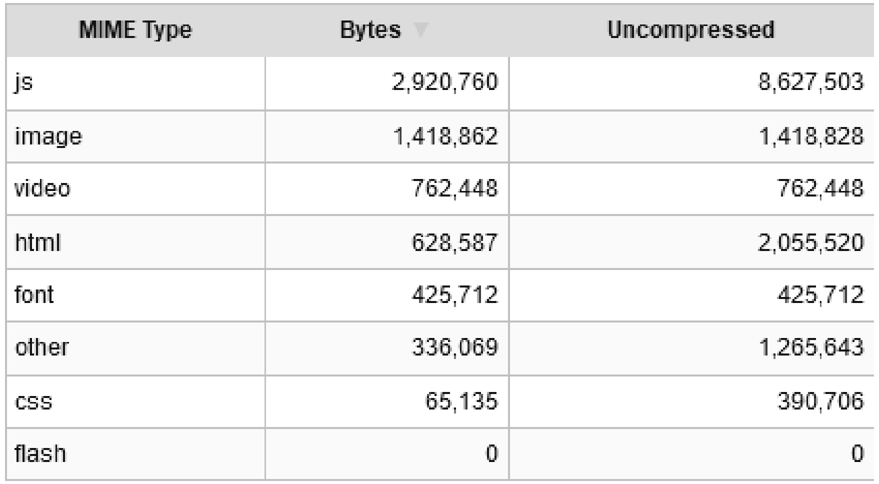
Content breakdown by MIME type (First View)



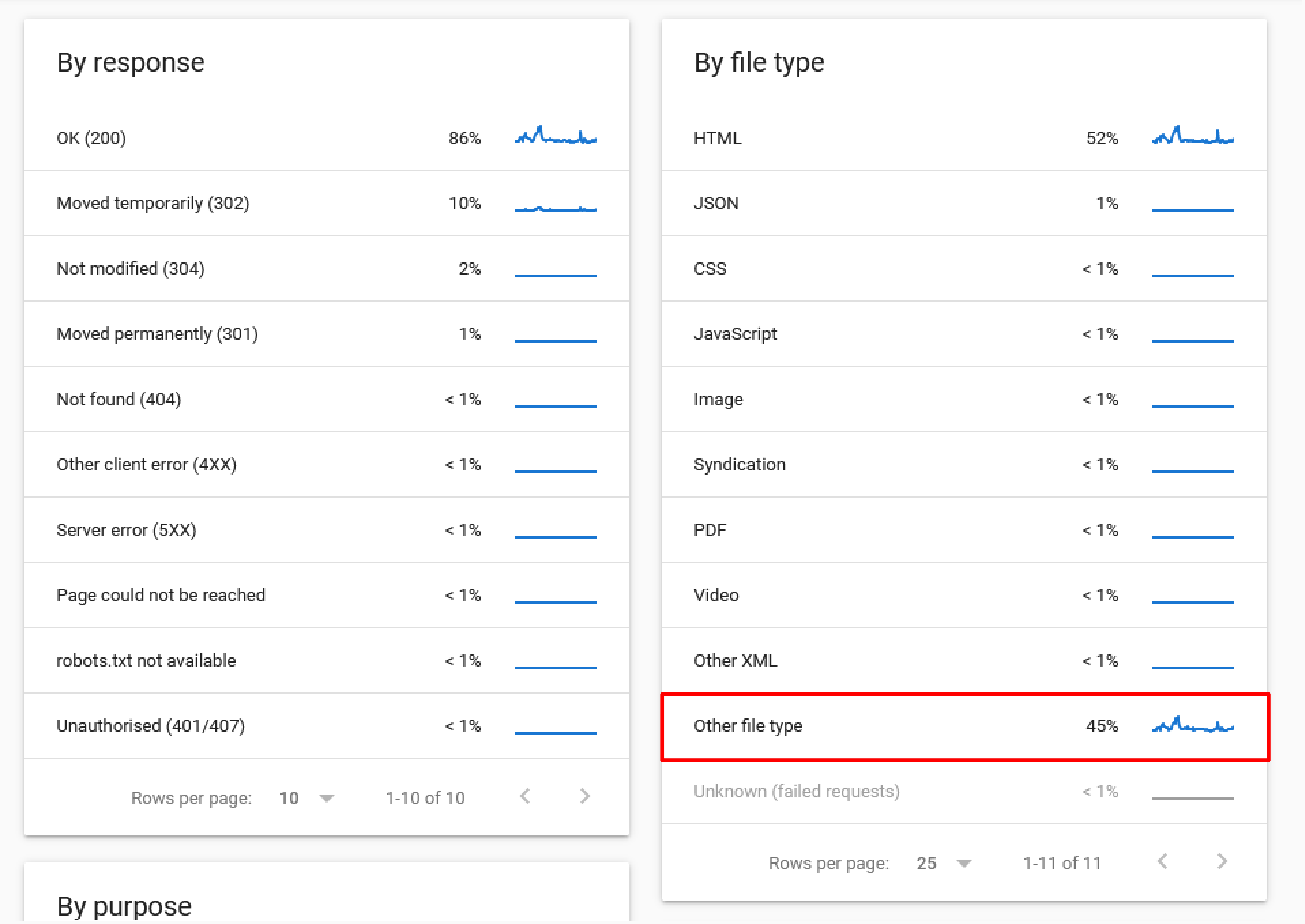




MIME Type	Requests ▼
other	365
image	358
html	300
js	140
css	19
font	11
video	1
flash	0



MIME Type	Bytes ▼	Uncompressed
js	2,920,760	8,627,503
image	1,418,862	1,418,828
video	762,448	762,448
html	628,587	2,055,520
font	425,712	425,712
other	336,069	1,265,643
css	65,135	390,706
flash	0	0



By response

OK (200)	86%
Moved temporarily (302)	10%
Not modified (304)	2%
Moved permanently (301)	1%
Not found (404)	< 1%
Other client error (4XX)	< 1%
Server error (5XX)	< 1%
Page could not be reached	< 1%
robots.txt not available	< 1%
Unauthorised (401/407)	< 1%

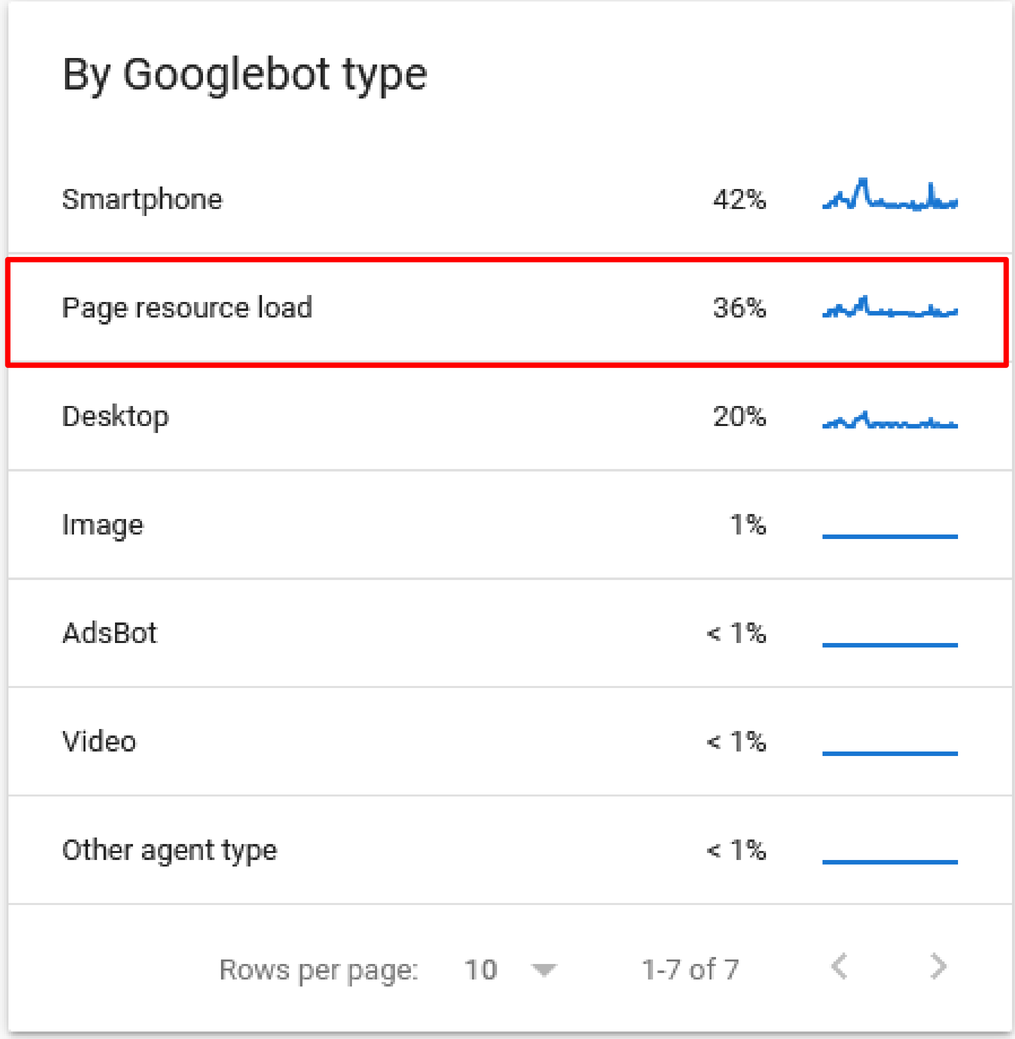
Rows per page: 10 ▾ 1-10 of 10 ‹ ›

By file type

HTML	52%
JSON	1%
CSS	< 1%
JavaScript	< 1%
Image	< 1%
Syndication	< 1%
PDF	< 1%
Video	< 1%
Other XML	< 1%
Other file type	45%
Unknown (failed requests)	< 1%

Rows per page: 25 ▾ 1-11 of 11 ‹ ›

By purpose



By Googlebot type

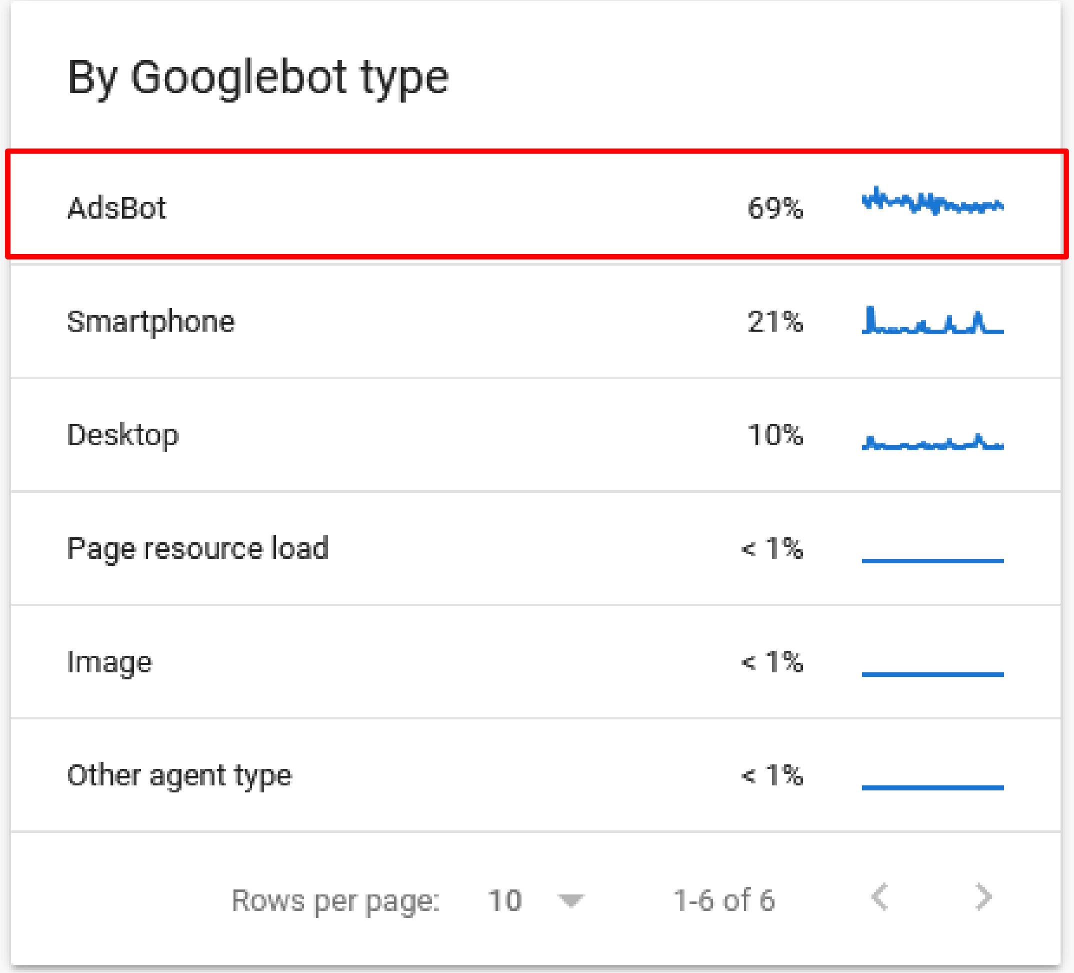
Smartphone	42%
Page resource load	36%
Desktop	20%
Image	1%
AdsBot	< 1%
Video	< 1%
Other agent type	< 1%

Rows per page: 10 ▾ 1-7 of 7 ‹ ›

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



By Googlebot type

AdsBot	69%
Smartphone	21%
Desktop	10%
Page resource load	< 1%
Image	< 1%
Other agent type	< 1%

Rows per page: 10 ▾ 1-6 of 6 ‹ ›

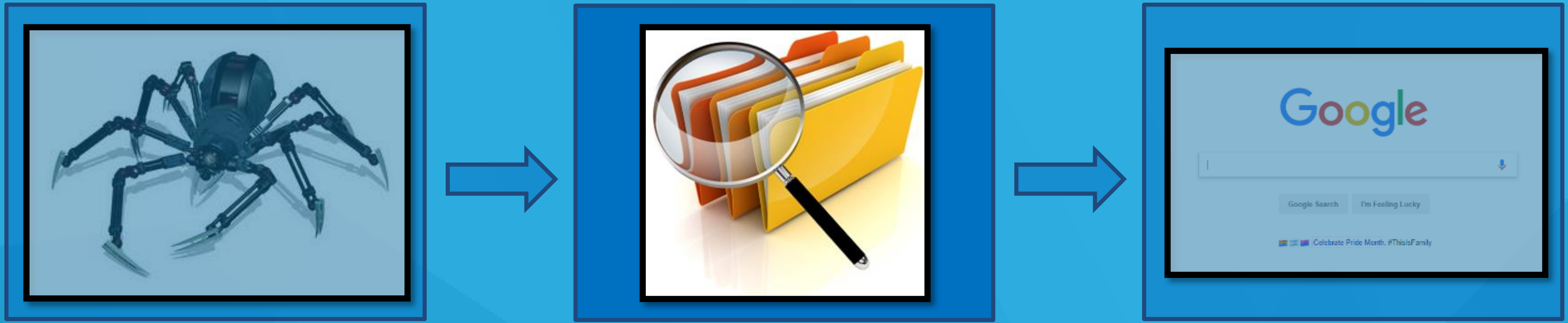
Optimise Crawling

- 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

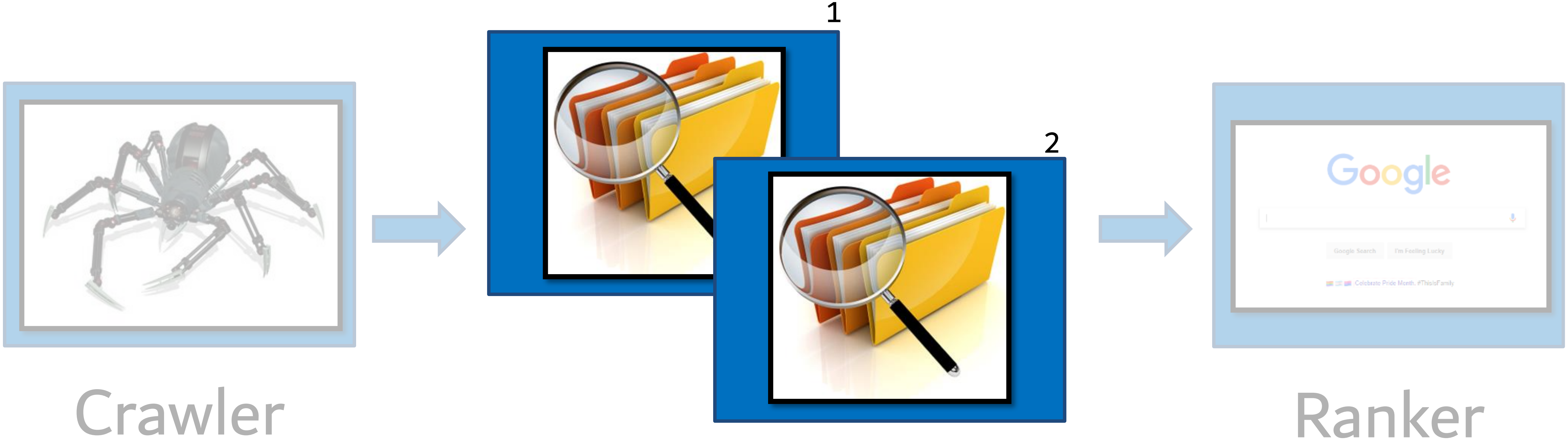


Crawler

Indexer

Ranker

Two Stages* of Indexing



Crawler

Indexer

Ranker

*At least – indexing is a collection of interconnected processes

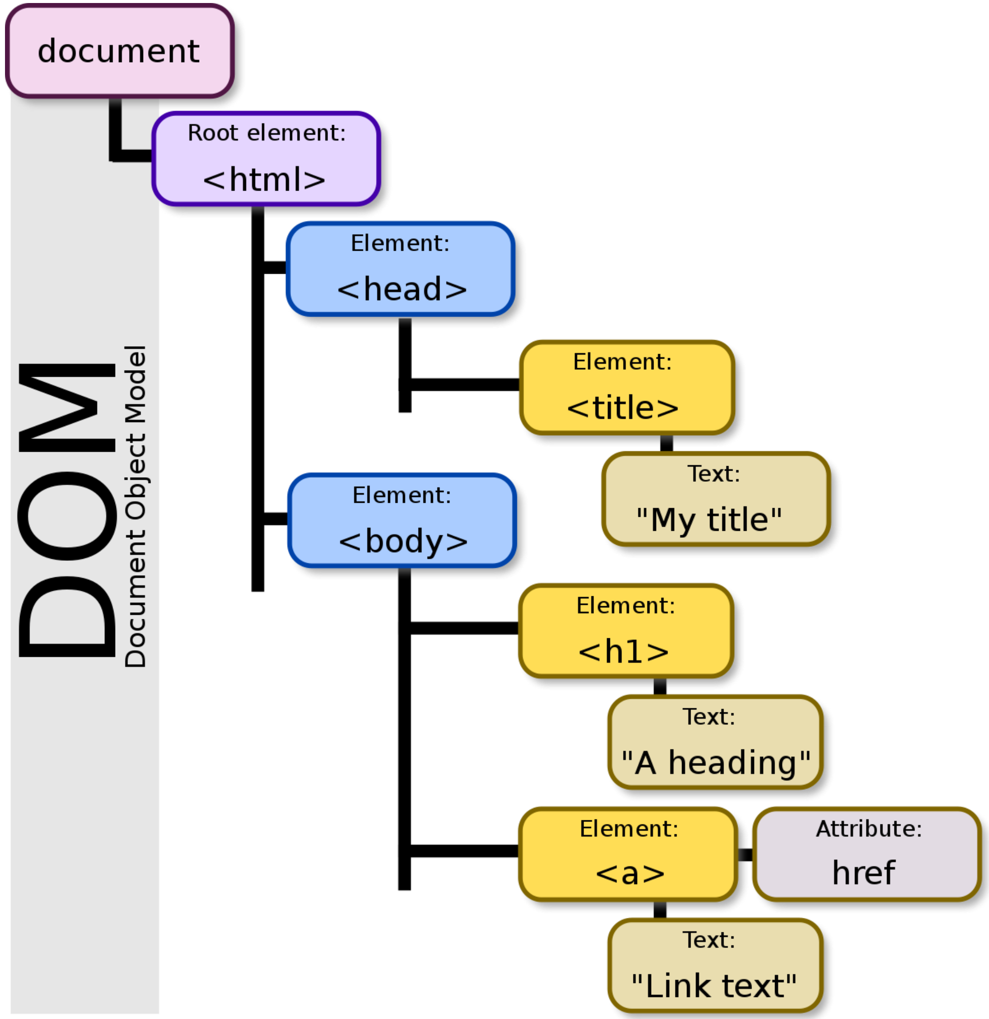
Indexing

- HTML lexer
  - ➢ Cleaning & tokenising the HTML
- Index selection
  - ➢ De-duping prior to indexing
- Indexing
  - ➢ First-pass based on HTML
  - ➢ Potential rendering (not guaranteed)
- Index integrity
  - ➢ Canonicalisation & de-duplication

Indexing
=
Extraction + Semantics

Extraction

Can Google easily extract a page’s content from its DOM?



Optimise Extraction (1)

- Clean HTML;
  - ➢ Yes, really!
  - ➢ There is a max HTML size Google will parse
    - – Speculation: ~1 MB
  - ➢ Less clutter = easier parsing

@badams

POLEMIC DIGITAL

Optimise Extraction (2)

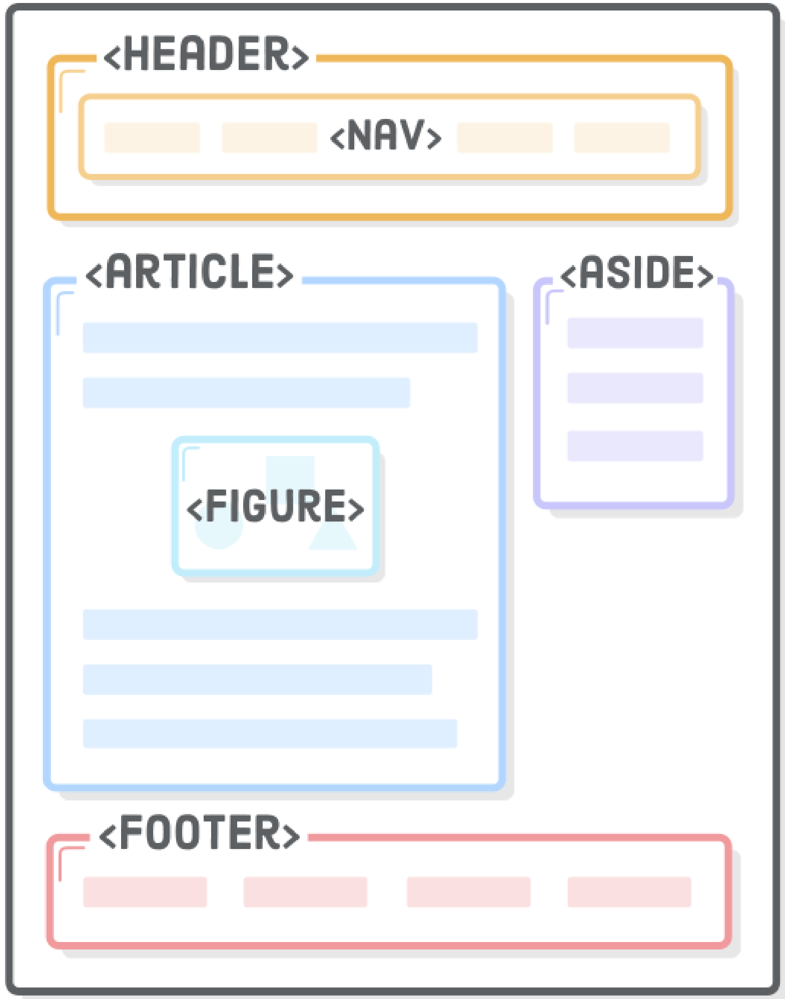
- Clean <head>;
  - ➢ Critical meta tags high in the <head>
    - – Title & description
    - – Open Graph
    - – Canonical, hreflang & mobile alternate
    - – Structured Data
  - ➢ Internal CSS & JS lower in the <head>

Optimise Extraction (3)

- Minimise DOM-manipulation;
  - ➢ Client-side JavaScript that manipulates the DOM can impact extraction
  - ➢ Can also impact Core Web Vitals

Semantics

Can Google understand what the page is about?



Optimise Semantics

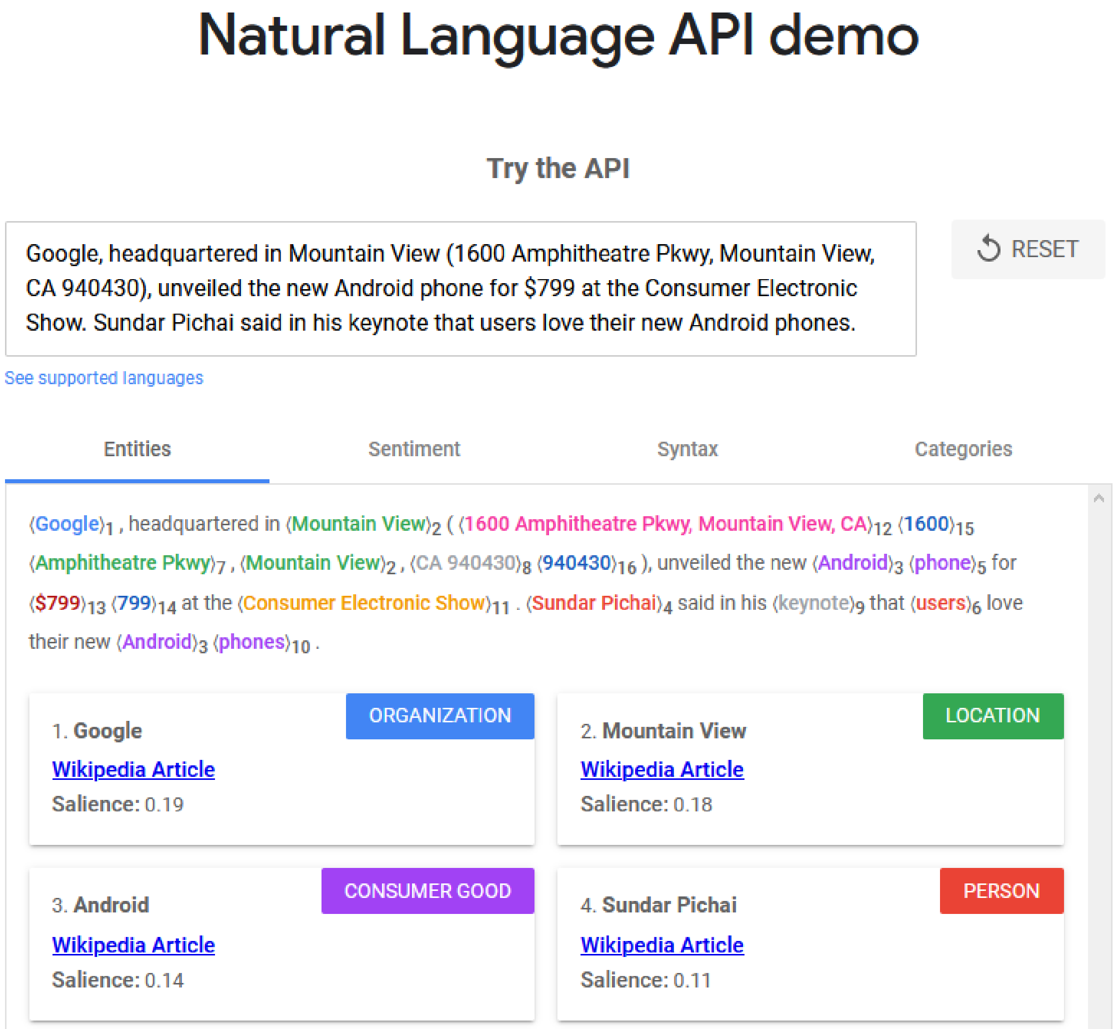
- Good content;
  - Easily identifiable entities and relationships

- Semantic HTML;
  - Enables Google to separate style & boilerplate from content

- Structured Data;
  - Makes page contents explicitly clear

Test Entities in Content

Google NLP API: https://cloud.google.com/natural-language



Natural Language API demo

Try the API

Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.

⟨Google⟩$_1$, headquartered in ⟨Mountain View⟩$_2$ (⟨1600 Amphitheatre Pkwy, Mountain View, CA⟩$_{12}$ ⟨1600⟩$_{15}$ ⟨Amphitheatre Pkwy⟩$_7$, ⟨Mountain View⟩$_2$, ⟨CA 940430⟩$_8$ ⟨940430⟩$_{16}$), unveiled the new ⟨Android⟩$_3$ ⟨phone⟩$_5$ for ⟨$799⟩$_{13}$ ⟨799⟩$_{14}$ at the ⟨Consumer Electronic Show⟩$_{11}$. ⟨Sundar Pichai⟩$_4$ said in his ⟨keynote⟩$_9$ that ⟨users⟩$_6$ love their new ⟨Android⟩$_3$ ⟨phones⟩$_{10}$.

1. Google — ORGANIZATION
Wikipedia Article
Salience: 0.19

2. Mountain View — LOCATION
Wikipedia Article
Salience: 0.18

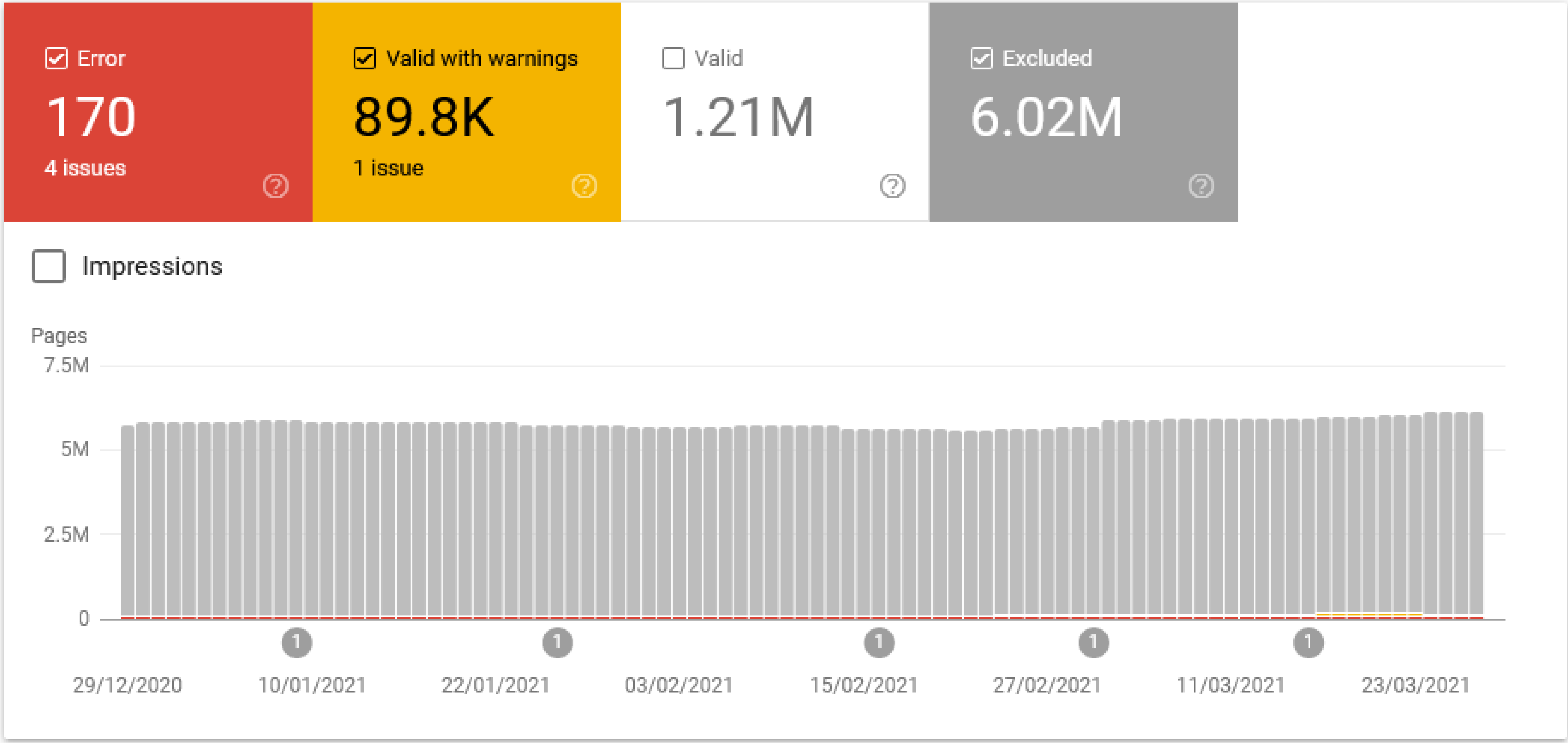
3. Android — CONSUMER GOOD
Wikipedia Article
Salience: 0.14

4. Sundar Pichai — PERSON
Wikipedia Article
Salience: 0.11

@badams

POLEMIC DIGITAL

GSC: Mix of Crawling & Indexing Issues



☑ Error
170
4 issues

☑ Valid with warnings
89.8K
1 issue

☐ Valid
1.21M

☑ Excluded
6.02M

☐ Impressions

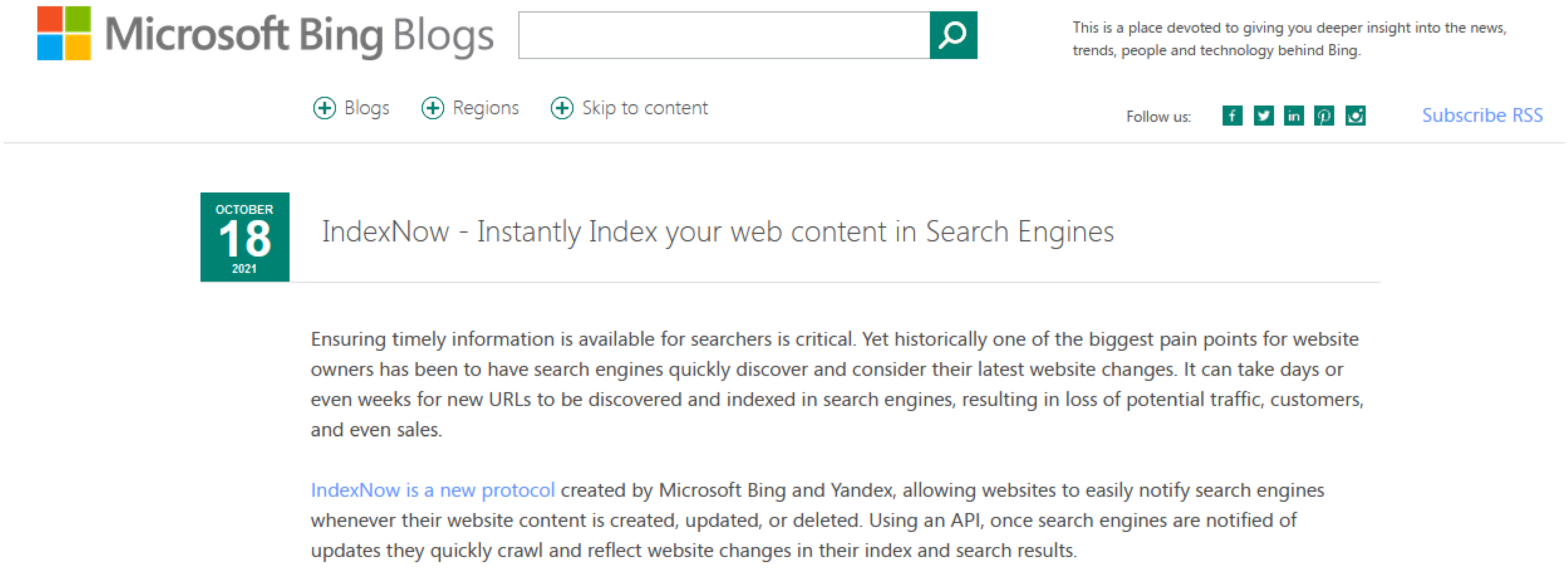
POLEMIC DIGITAL

Technical SEO in 2022

@badams

POLEMIC DIGITAL

IndexNow


Microsoft Bing Blogs
This is a place devoted to giving you deeper insight into the news, trends, people and technology behind Bing.
Blogs
Regions
Skip to content
Follow us:
Subscribe RSS

OCTOBER 18 2021

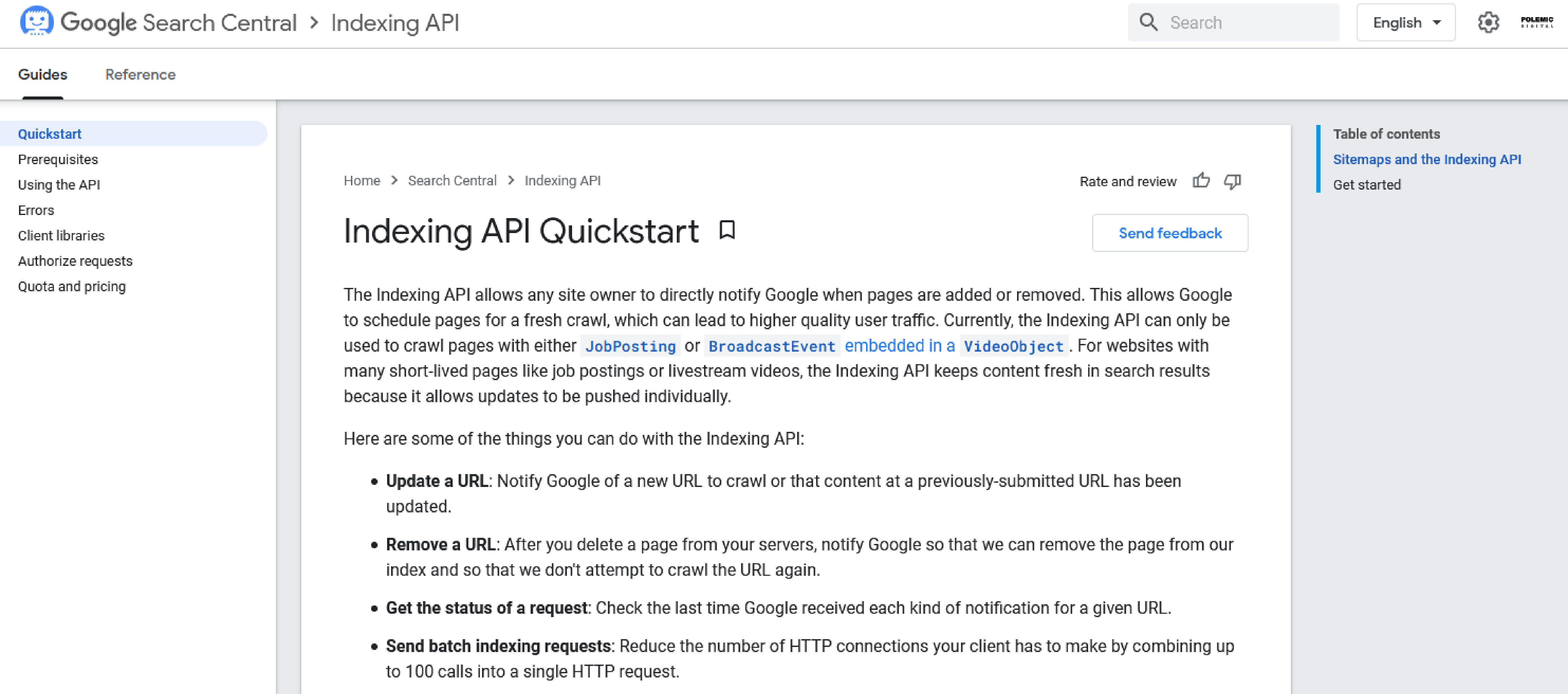
IndexNow - Instantly Index your web content in Search Engines

Ensuring timely information is available for searchers is critical. Yet historically one of the biggest pain points for website owners has been to have search engines quickly discover and consider their latest website changes. It can take days or even weeks for new URLs to be discovered and indexed in search engines, resulting in loss of potential traffic, customers, and even sales.

IndexNow is a new protocol created by Microsoft Bing and Yandex, allowing websites to easily notify search engines whenever their website content is created, updated, or deleted. Using an API, once search engines are notified of updates they quickly crawl and reflect website changes in their index and search results.

https://blogs.bing.com/webmaster/october-2021/IndexNow-Instantly-Index-your-web-content-in-Search-Engines

Live Indexing API (?)



Indexing API Quickstart

The Indexing API allows any site owner to directly notify Google when pages are added or removed. This allows Google to schedule pages for a fresh crawl, which can lead to higher quality user traffic. Currently, the Indexing API can only be used to crawl pages with either `JobPosting` or `BroadcastEvent` embedded in a `VideoObject`. For websites with many short-lived pages like job postings or livestream videos, the Indexing API keeps content fresh in search results because it allows updates to be pushed individually.

Here are some of the things you can do with the Indexing API:

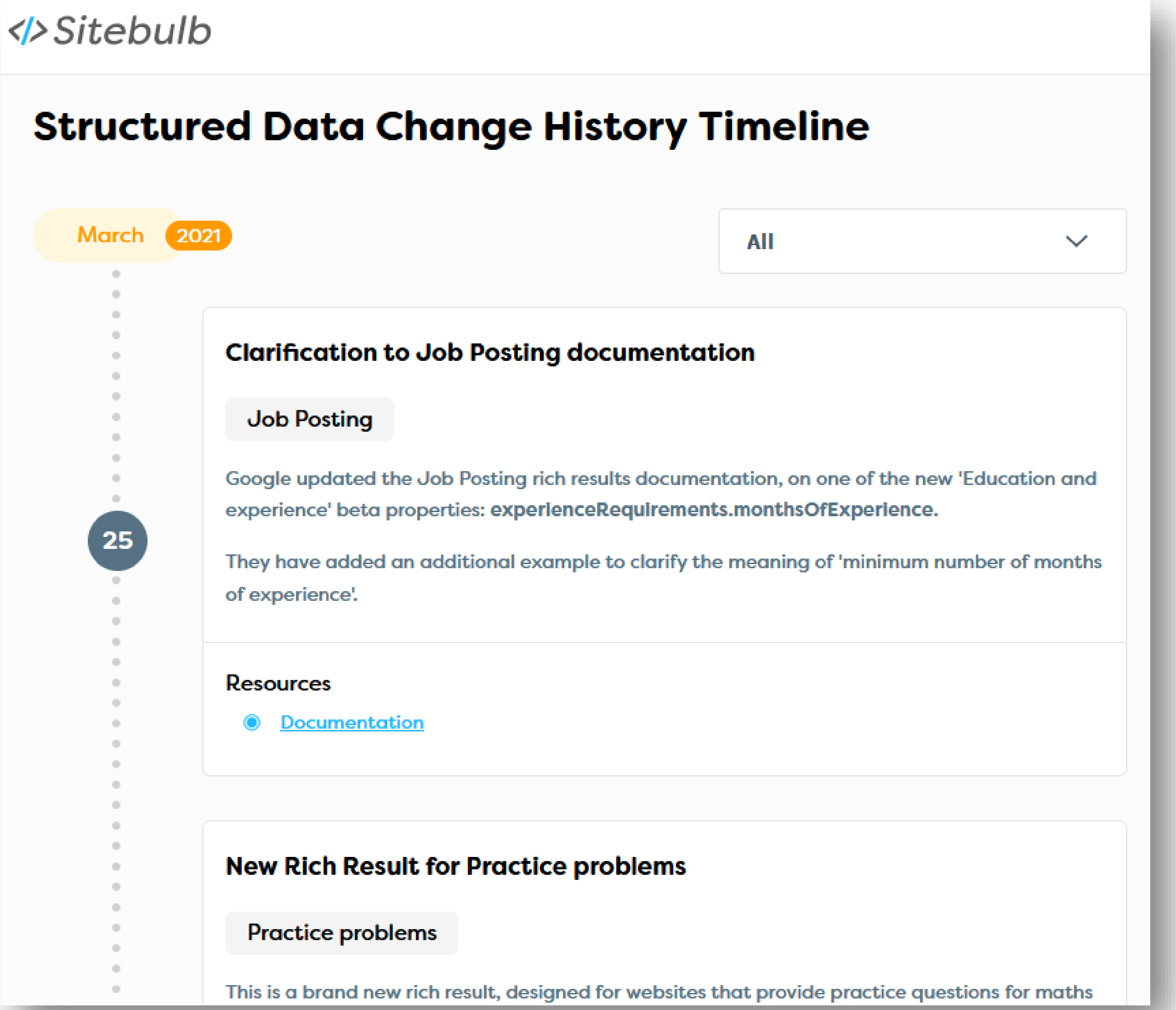
- **Update a URL**: Notify Google of a new URL to crawl or that content at a previously-submitted URL has been updated.
- **Remove a URL**: After you delete a page from your servers, notify Google so that we can remove the page from our index and so that we don't attempt to crawl the URL again.
- **Get the status of a request**: Check the last time Google received each kind of notification for a given URL.
- **Send batch indexing requests**: Reduce the number of HTTP connections your client has to make by combining up to 100 calls into a single HTTP request.

https://developers.google.com/search/apis/indexing-api/v3/quickstart

Structured Data

Constantly evolving schemas

New rich snippets in SERPs



Sitebulb

Structured Data Change History Timeline

March 2021

All

25

Clarification to Job Posting documentation

Job Posting

Google updated the Job Posting rich results documentation, on one of the new 'Education and experience' beta properties: **experienceRequirements.monthsOfExperience.**

They have added an additional example to clarify the meaning of 'minimum number of months of experience'.

Resources

- Documentation

New Rich Result for Practice problems

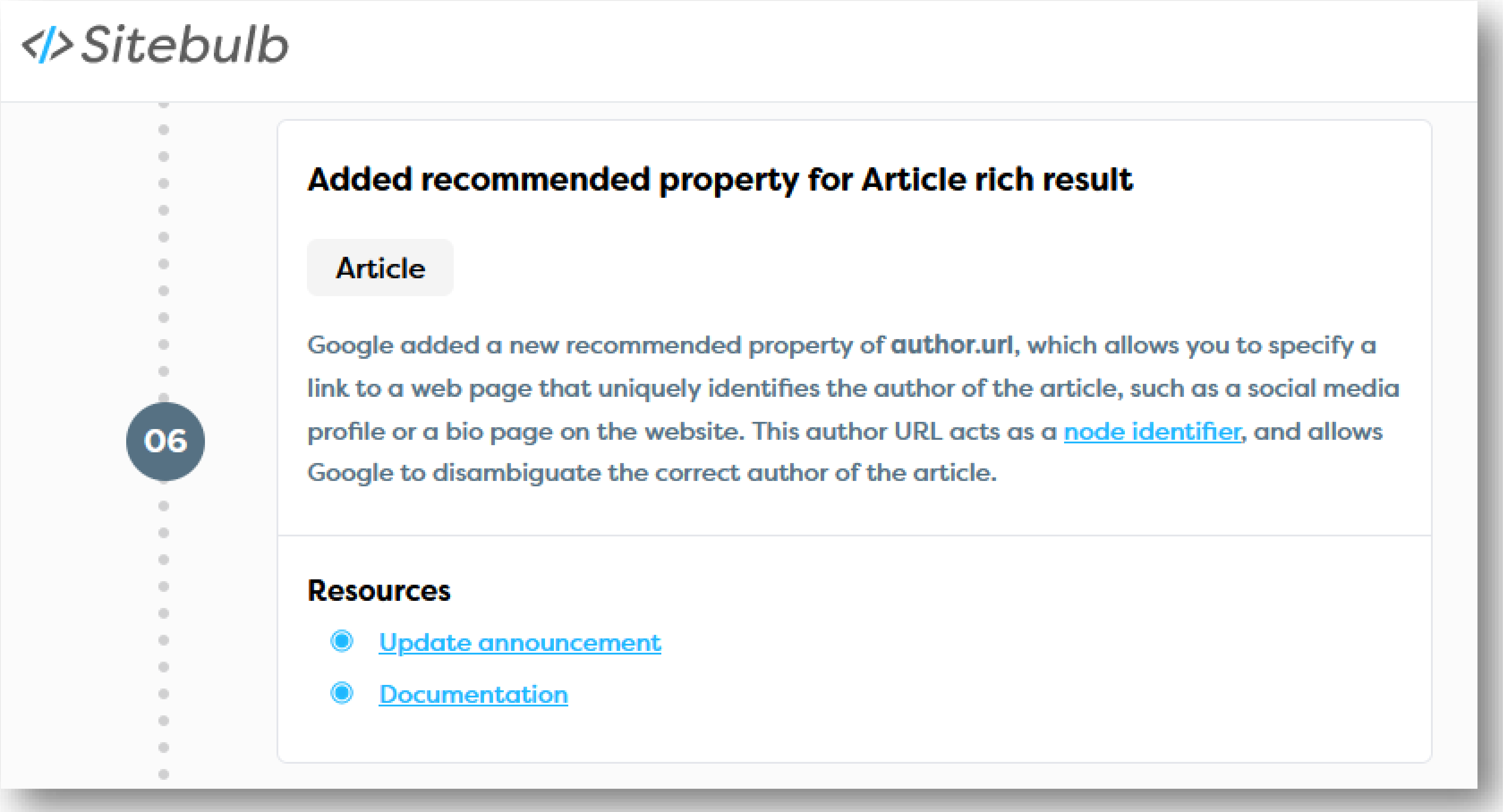
Practice problems

This is a brand new rich result, designed for websites that provide practice questions for maths

https://sitebulb.com/structured-data-history/

Structured Data

- ‘author.url’ now recommended in Article SD



Added recommended property for Article rich result

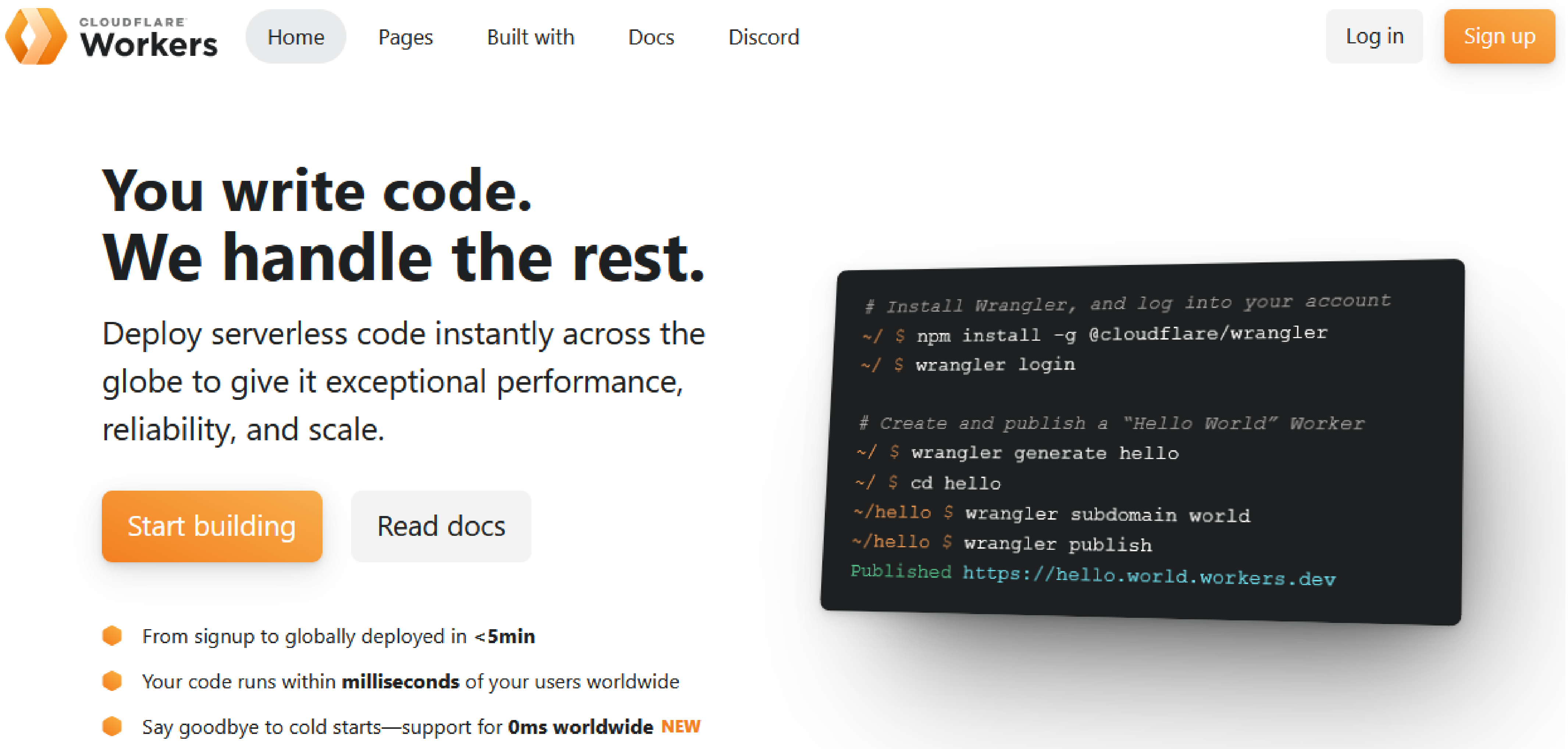
Article

Google added a new recommended property of **author.url**, which allows you to specify a link to a web page that uniquely identifies the author of the article, such as a social media profile or a bio page on the website. This author URL acts as a node identifier, and allows Google to disambiguate the correct author of the article.

Resources

- Update announcement
- Documentation

Edge SEO



You write code.
We handle the rest.

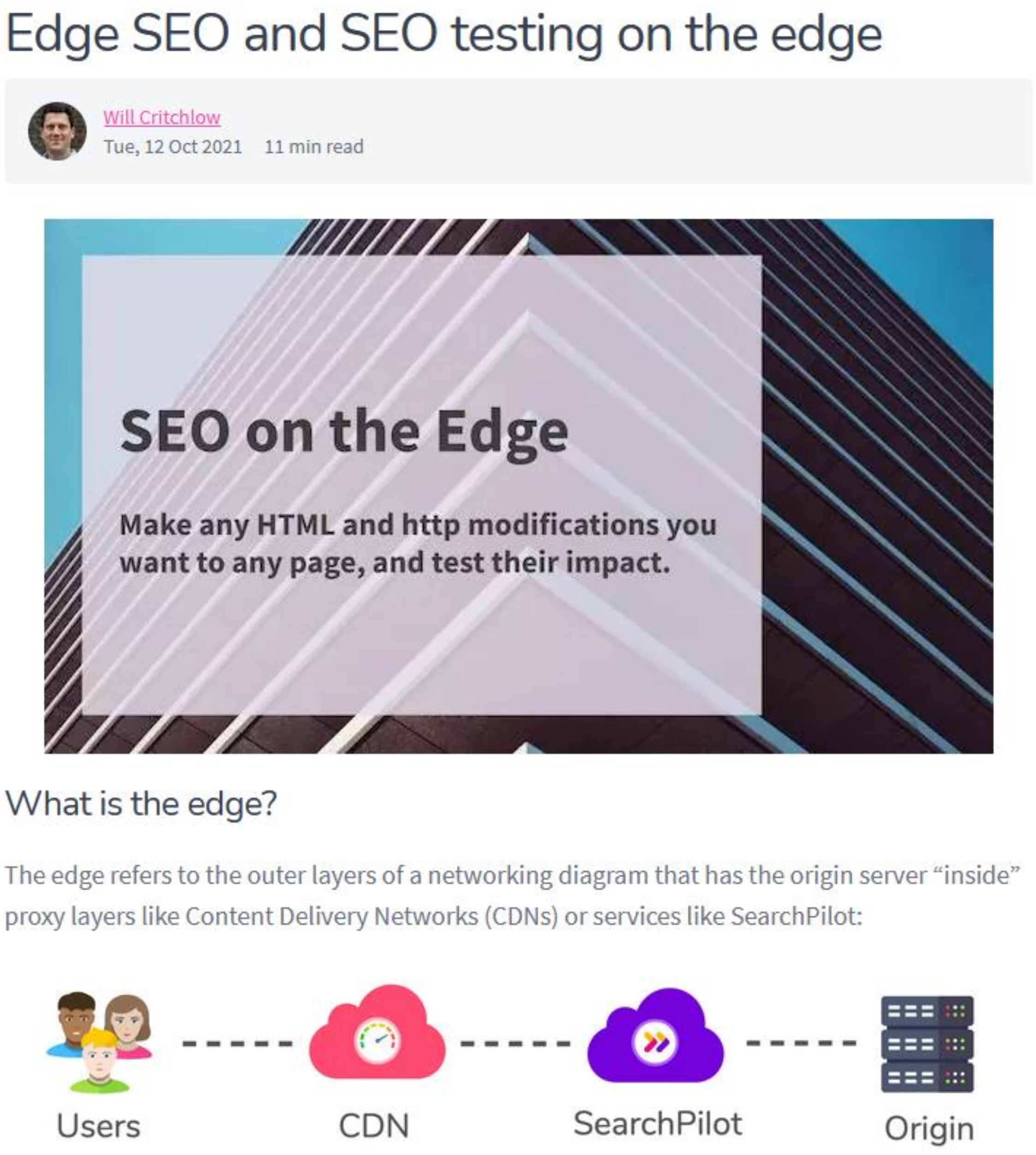
Deploy serverless code instantly across the globe to give it exceptional performance, reliability, and scale.

- From signup to globally deployed in <**5min**
- Your code runs within **milliseconds** of your users worldwide
- Say goodbye to cold starts—support for **0ms worldwide** NEW

```
# Install Wrangler, and log into your account
~/ $ npm install -g @cloudflare/wrangler
~/ $ wrangler login

# Create and publish a "Hello World" Worker
~/ $ wrangler generate hello
~/ $ cd hello
~/hello $ wrangler subdomain world
~/hello $ wrangler publish
Published https://hello.world.workers.dev
```

https://dantaylor.online/edge-seo/



Edge SEO and SEO testing on the edge

Will Critchlow

Tue, 12 Oct 2021 11 min read

SEO on the Edge
Make any HTML and http modifications you want to any page, and test their impact.

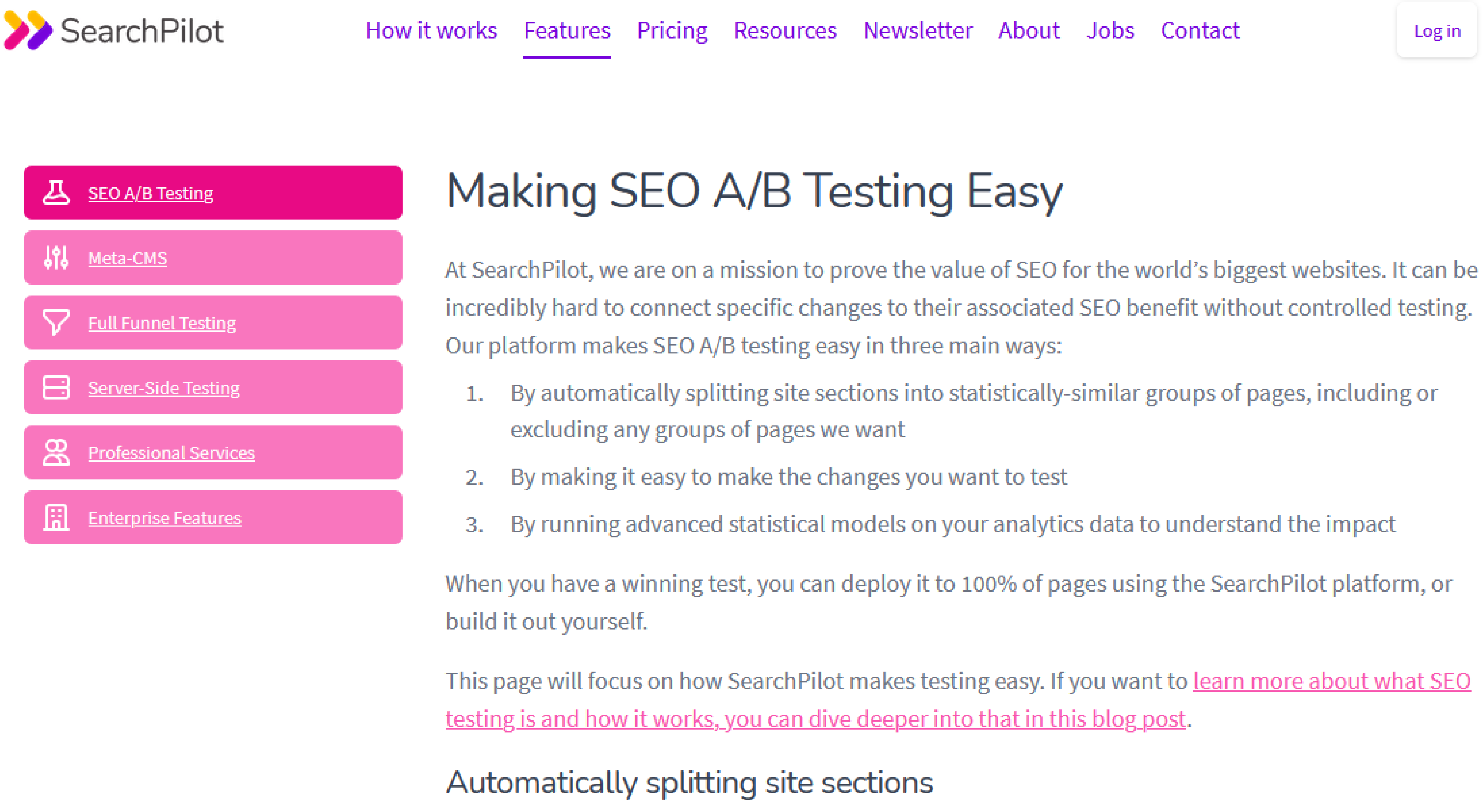
What is the edge?

The edge refers to the outer layers of a networking diagram that has the origin server "inside" proxy layers like Content Delivery Networks (CDNs) or services like SearchPilot:

Users
CDN
SearchPilot
Origin

https://www.searchpilot.com/resources/blog/edge-seo/

A/B Testing



Making SEO A/B Testing Easy

At SearchPilot, we are on a mission to prove the value of SEO for the world's biggest websites. It can be incredibly hard to connect specific changes to their associated SEO benefit without controlled testing. Our platform makes SEO A/B testing easy in three main ways:

1. By automatically splitting site sections into statistically-similar groups of pages, including or excluding any groups of pages we want
2. By making it easy to make the changes you want to test
3. By running advanced statistical models on your analytics data to understand the impact

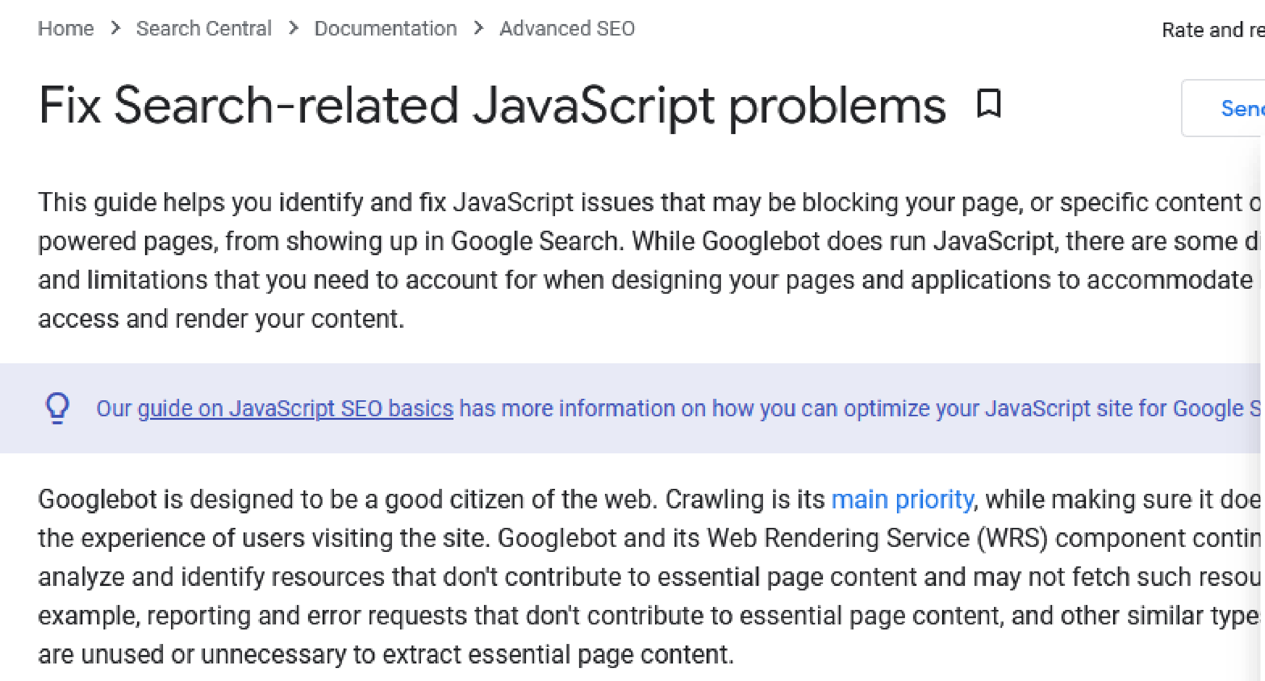
When you have a winning test, you can deploy it to 100% of pages using the SearchPilot platform, or build it out yourself.

This page will focus on how SearchPilot makes testing easy. If you want to learn more about what SEO testing is and how it works, you can dive deeper into that in this blog post.

Automatically splitting site sections

https://www.searchpilot.com/

Less hassle with JavaScript



Home > Search Central > Documentation > Advanced SEO

Rate and review

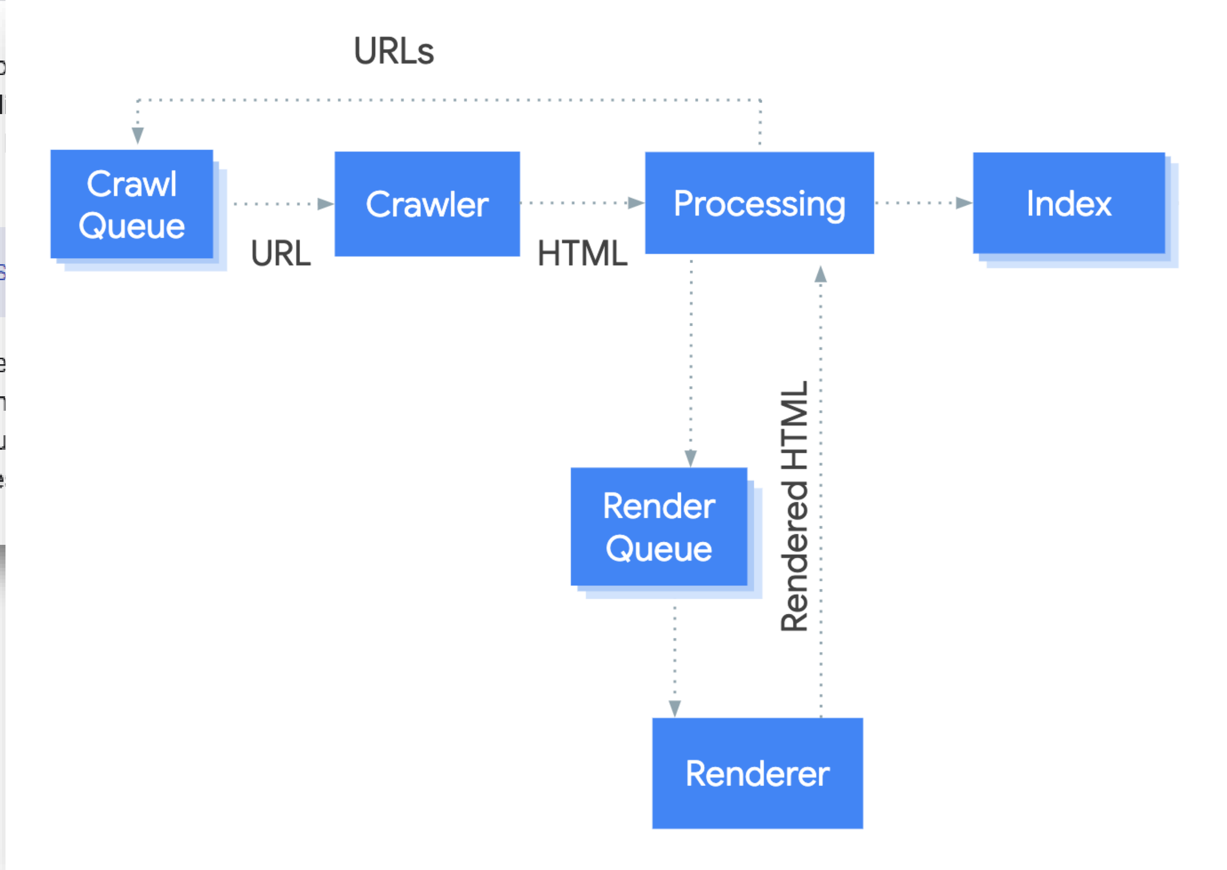
Fix Search-related JavaScript problems

Send feedback

This guide helps you identify and fix JavaScript issues that may be blocking your page, or specific content o
powered pages, from showing up in Google Search. While Googlebot does run JavaScript, there are some d
and limitations that you need to account for when designing your pages and applications to accommodate
access and render your content.

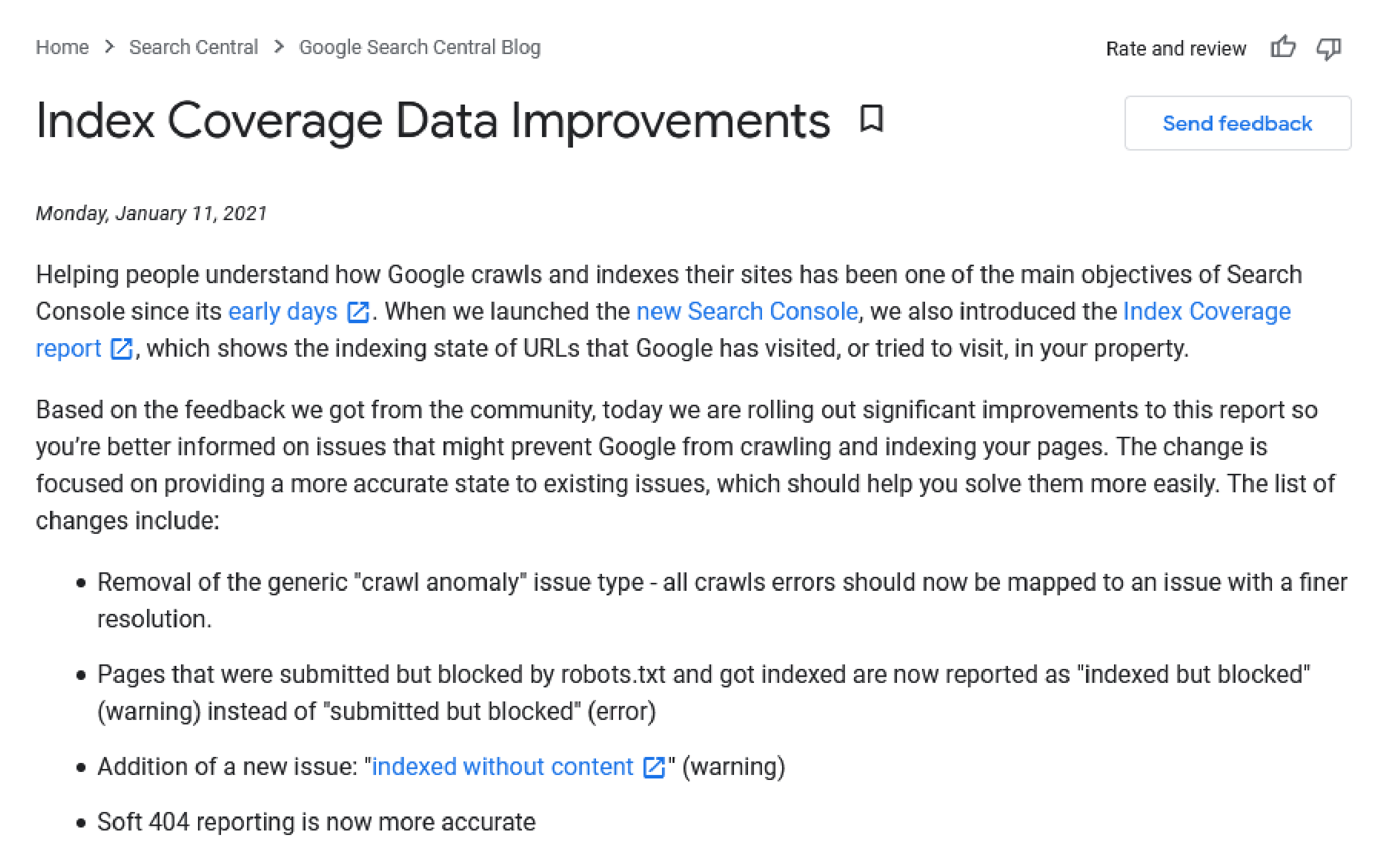
Our guide on JavaScript SEO basics has more information on how you can optimize your JavaScript site for Google S

Googlebot is designed to be a good citizen of the web. Crawling is its main priority, while making sure it doe
the experience of users visiting the site. Googlebot and its Web Rendering Service (WRS) component contin
analyze and identify resources that don't contribute to essential page content and may not fetch such resou
example, reporting and error requests that don't contribute to essential page content, and other similar type
are unused or unnecessary to extract essential page content.



Better GSC Reports

More useful info to empower SEOs & Devs



Index Coverage Data Improvements

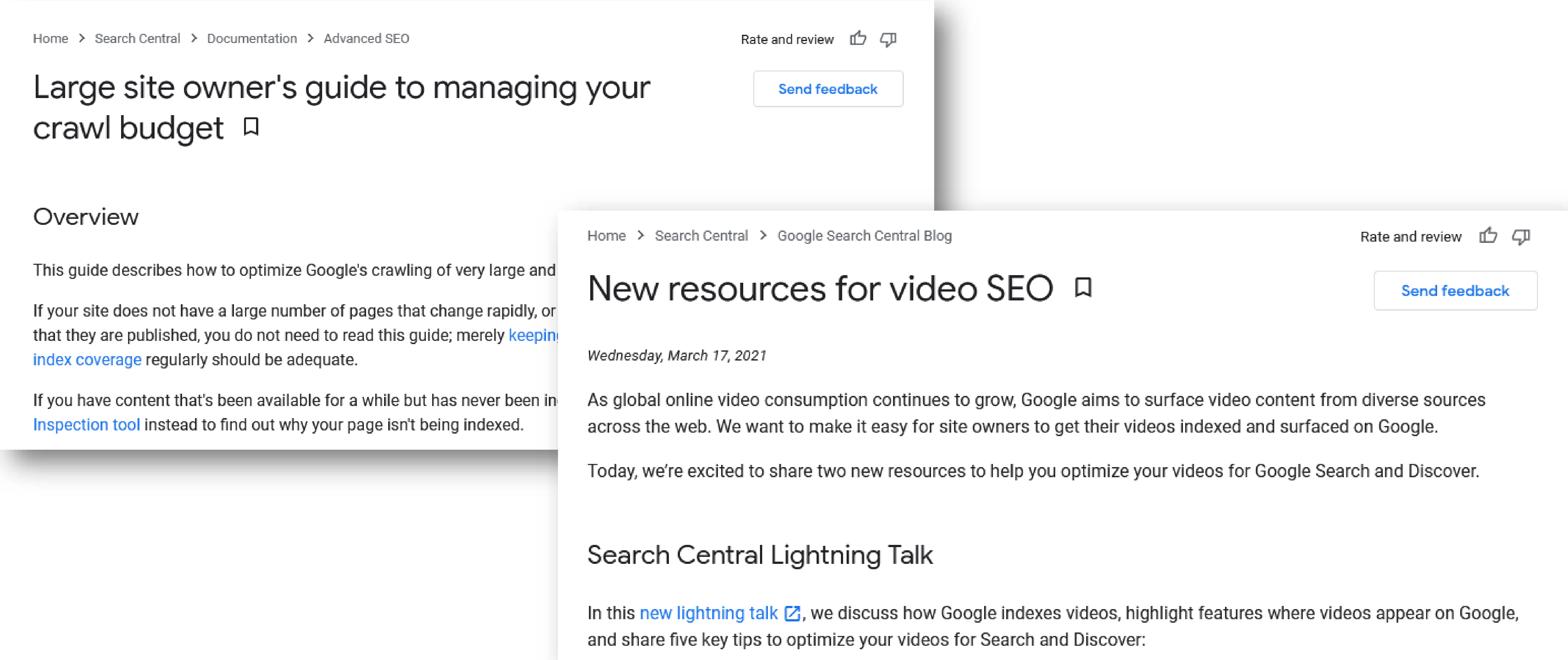
Monday, January 11, 2021

Helping people understand how Google crawls and indexes their sites has been one of the main objectives of Search Console since its early days. When we launched the new Search Console, we also introduced the Index Coverage report, which shows the indexing state of URLs that Google has visited, or tried to visit, in your property.

Based on the feedback we got from the community, today we are rolling out significant improvements to this report so you're better informed on issues that might prevent Google from crawling and indexing your pages. The change is focused on providing a more accurate state to existing issues, which should help you solve them more easily. The list of changes include:

- Removal of the generic "crawl anomaly" issue type - all crawls errors should now be mapped to an issue with a finer resolution.
- Pages that were submitted but blocked by robots.txt and got indexed are now reported as "indexed but blocked" (warning) instead of "submitted but blocked" (error)
- Addition of a new issue: "indexed without content" (warning)
- Soft 404 reporting is now more accurate

Better Google Documentation



Home > Search Central > Documentation > Advanced SEO

Rate and review

Send feedback

Large site owner's guide to managing your crawl budget

Overview

This guide describes how to optimize Google's crawling of very large and

If your site does not have a large number of pages that change rapidly, or that they are published, you do not need to read this guide; merely keepin index coverage regularly should be adequate.

If you have content that's been available for a while but has never been in Inspection tool instead to find out why your page isn't being indexed.

Home > Search Central > Google Search Central Blog

Rate and review

Send feedback

New resources for video SEO

Wednesday, March 17, 2021

As global online video consumption continues to grow, Google aims to surface video content from diverse sources across the web. We want to make it easy for site owners to get their videos indexed and surfaced on Google.

Today, we're excited to share two new resources to help you optimize your videos for Google Search and Discover.

Search Central Lightning Talk

In this new lightning talk, we discuss how Google indexes videos, highlight features where videos appear on Google, and share five key tips to optimize your videos for Search and Discover:

My Tech SEO Tools

SEO Crawlers

- DeepCrawl
  https://www.deepcrawl.com/
- Sitebulb
  https://sitebulb.com/
- Screaming Frog
  https://www.screamingfrog.co.uk/seo-spider/

SEO Review & Monitoring

- Little Warden
  https://littlewarden.com/

- ContentKing
  https://www.contentkingapp.com/

- SEO Info
  https://weeblr.com/doc/products.seoinfo/current/overview/

- SEOBrowse
  https://seobrowse.com/

Performance Analysis

- PageSpeed Insights
  https://pagespeed.web.dev/

- WebPagetest.org
  https://www.webpagetest.org/

- GTmetrix
  https://gtmetrix.com/

Barry Adams

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

PUBCON brightonSEO® unGAggED LEARN INBOUND NESS NEWS & EDITORIAL SEO SUMMIT

SMX® MÜNCHEN GERMANY FRIENDS OF SEARCH optimisey search LEEDS NOTTINGHAM DIGITAL SUMMIT

SEOCAMP'us L'ÉVÈNEMENT SEARCH MARKETING CHIANG MAI SEO CONFERENCE 2019 Digitalzone DIGITAL ELITE DAY MB MARKETING BUSINESS SUMMIT®

FRIENDS OF SEARCH

Thank You!

barry@polemicdigital.com
twitter.com/badams