Drupal SEO pitfalls

And how to avoid them

- By Brent Gees and Wouter De Bruycker

ABOUT US

3

About us

Brent Gees
- Drupal Developer and Drupal Trainer at Dropsolid
- Speaker at various Drupalcamps, DrupalCons and dev days

About us

Wouter De Bruycker

- SEO Strategist & Evangelist at Dropsolid
- Speaker at Drupal events
- Closing Keynote Speaker at DMB Day Ghent "*Today is Yesterday's Future: SEO in Evolution*"
- Speaker at Digital Content Marketing Congress: "*Technical SEO 101 (& 102)*"

5

What we'll talk about in this session

- Public entities
- All pages should be an entity
- Indexable internal search
- Indexed test environments
- Assets blocked by robots.txt
- Security leaks impacting SEO
- Robots.txt disallow != noindex
- Google Analytics horror
- Visual stability issues (CLS)
- Images impacting performance
- Rapid fire best practices

6

PUBLIC ENTITIES

PUBLIC ENTITIES

By default, entities might be **publicly available on their own unique URL.**

Example: a "team member" content type could generate a public node for each team member, while they are really only used for a "team overview page" and the site doesn't need detail pages for each team member.

https://mysite.com/node/42, https://mysite.com/taxonomy/term/42, ...

PUBLIC ENTITIES

This results in low value and "**thin content**" pages, **indexable** by Google.

You **don't** want these pages on your site because they are a **waste of resources** (bandwidth, crawl budget, database storage, …).

ALL PAGES SHOULD BE AN ENTITY

When a page isn't editable as a node, there's no easy way to...

- Edit the meta title or meta description tags
- Configure the XML sitemap inclusion for this page
- Edit the OpenGraph tags
- ...

Content editors need this functionality!

PUBLIC ENTITIES

Prevent those entities from being accessed by visitors (and search engines).
This can be done by modules such as rabbit_hole

ALL PAGES SHOULD BE AN ENTITY

ALL PAGES SHOULD BE AN ENTITY

Oftentimes, some pages on websites are generated based on other content. These pages are not an "editable node" in the backend.
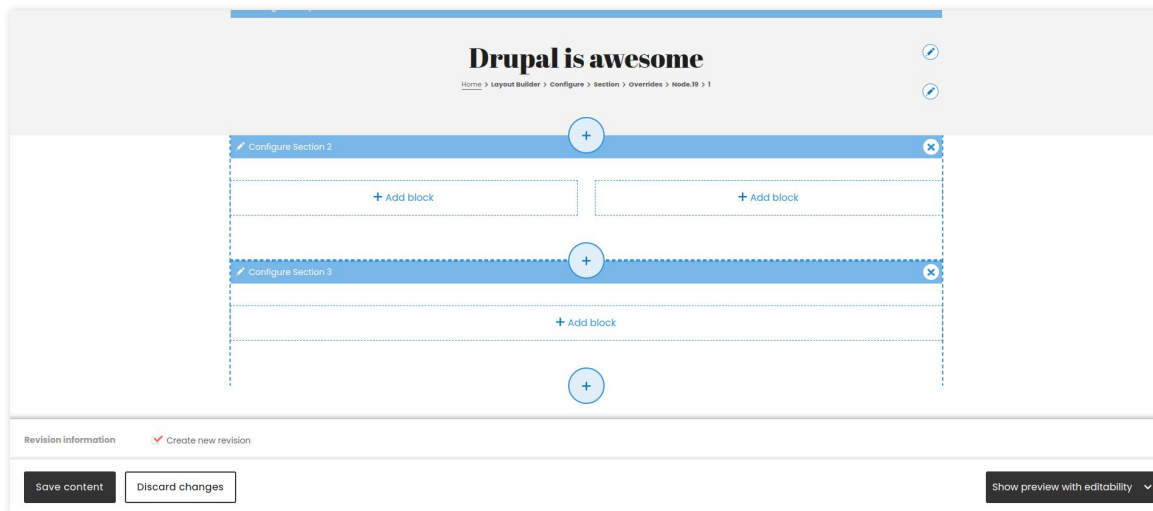
Examples: homepage, overview pages, …

ALL PAGES SHOULD BE AN ENTITY

- Use the core drupal layout builder to construct the homepage

ALL PAGES SHOULD BE AN ENTITY

- Use paragraphs together with modules such as block_field or overview_field to add blocks to the page.

INDEXABLE INTERNAL SEARCH

INDEXABLE INTERNAL SEARCH

By default, **internal** search result pages are often indexable by search engines.

INDEXABLE INTERNAL SEARCH

This results in low value and "**thin content**" pages, **indexable** by Google.

You **don't** want low-value pages in the index.

Google also implicitly mentions this in its quality guidelines to not let Googlebot index internal search results.

INDEXABLE INTERNAL SEARCH

When using a node for your search pages
Install metatag and add the field to your content type

INDEXABLE INTERNAL SEARCH

When using views pages for your search pages
Install metatag and its submodule metatag_views

20

INDEXED TEST ENVIRONMENTS

INDEXED TEST ENVIRONMENTS

Development and staging environments are often crawlable and indexable by search engines because of configuration issues.

Or because of laziness ¯_(ツ)_/¯

Temporary content is also something that often shows up in search results.

INDEXED TEST ENVIRONMENTS

Everybody probably agrees we **don't** want our staging environments or paragraph testing pages to be in the Google index, or even be publicly available, for a wide range of reasons.

INDEXED TEST ENVIRONMENTS

site:acquia-sites.com inurl:prod

Q All | Images | Videos | Maps | News | More | Tools

Testing Post - Matti - STAGING SITE | Kaverina olen cool.
https://staging.maenkiusaa.fi › testing-post-matti ▾
This is a sample text not using the Oxygen Builder.**Lorem** Ipsum is a dummy text of the printing and typesetting industry. **Lorem** Ipsum has been the industry's ...

March 2018 posts - Safety and health at work - EU-OSHA
https://test.osha.europa.eu › blog ▾
Mar 8, 2018 - EU OSHA information about occupational health and safety across EU countries. Find surveys, regulations and standards, risk assessment ...

Main Content Only - Lievegem Frisbee Team
lift-it-up.org › Staging › index.php › features › main-content-only ▾
Lorem Ipsum is simply dummy text of the printing and typesetting industry. **Lorem** Ipsum has been the industry's standard dummy text ever since the 1500s, when ...

---

https://nydmvdev.prod.acquia-sites.com ⋮

https://nydmvdev.prod.acquia-sites.com/
No information is available for this page.
Learn why

http://ricollege.prod.acquia-sites.com ⋮

Rhode Island College
We offer 25 master's degrees in healthcare, counseling, social work, business, management, teaching, community leadership, and arts and sciences – plus doctoral ...

https://lineage.prod.acquia-sites.com › services › prepare ⋮

Food Manufacturing | Lineage Logistics
Our team has expertise and experience in designing a layout to satisfy your emerging supply chain needs while improving supply chain efficiencies and reducing ...

INDEXED TEST ENVIRONMENTS

When you add **test pages on live environments**
Unpublish if possible or prevent from indexing with metatag module

INDEXED TEST ENVIRONMENTS

On test environments:

Use *htpassword* protection for your environments

INDEXED TEST ENVIRONMENTS

Don't assume Google won't index your staging site because you don't have links pointing to it.

Actively prevent Google from indexing it.

INDEXED TEST ENVIRONMENTS

Follow-up by adding a domain property in Google Search Console for your **staging domain**.

You'll be able to see:

- How many staging pages are indexed
- What the URLs are
- **How many people actually visited** the staging pages
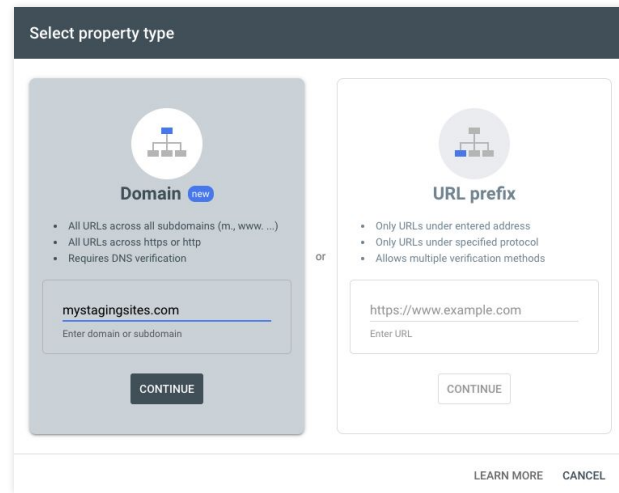
Image: creating a domain property in GSC. In this example for sites that are located on subdomains of this domain.
E.g. "sitename.mystagingsites.com"

28

ASSETS BLOCKED BY ROBOTS.TXT

ASSETS BLOCKED BY ROBOTS.TXT

Sometimes website assets (like favicons or images) are located inside
a folder that's blocked for crawlers via the robots.txt.

```
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
```

ASSETS BLOCKED BY ROBOTS.TXT

We want Google to understand our entire page and "see" it as a regular website visitor would see it.

If some assets are blocked via robots.txt, Google can't view the page as if it was a regular website visitor.

Image: drupalcamp.be/robots.txt
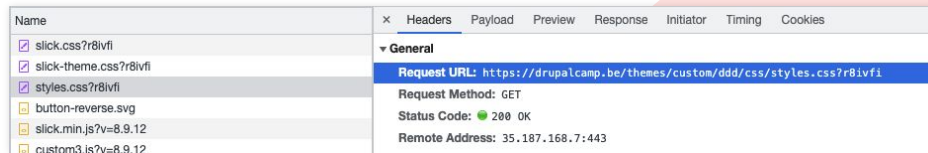
Image: http-requests to files that are not crawlable by Google

ASSETS BLOCKED BY ROBOTS.TXT

We want Google to understand our entire page and "see" it as a regular website visitor would see it.

If some assets are blocked via robots.txt, Google can't view the page as if it was a regular website visitor.

Page partially loaded

Not all page resources could be loaded. This can affect how Google sees and understands your page. Fix availability problems for any resources that can affect how Google understands your page.

Image: notification in Google Search Console due to blocked site assets.

32

ASSETS BLOCKED BY ROBOTS.TXT

Make sure your website assets (images, icons, favicon, …) are in a publicly available folder, not blocked by robots.txt.

Keep an eye on your Google Search Console property for notifications regarding blocked resources.

ASSETS BLOCKED BY ROBOTS.TXT

When **would** you block URLs from Google?

When you have a big website (**millions** of pages) and a big portion of those pages are for example product overviews with filtering parameters, you could block Google from spending time crawling the filtered pages and thus optimizing "crawl budget".

For most websites: crawl budget isn't an issue.

SECURITY LEAKS IMPACTING SEO

SECURITY LEAKS IMPACTING SEO

Allowing public file-uploads could result in lower organic traffic when not setup correctly.

SECURITY LEAKS IMPACTING SEO

If files can be uploaded without some form of authentication or CAPTCHA, this could result in thousands of files being uploaded by spammers, and **indexed by Google**.

SECURITY LEAKS IMPACTING SEO

38
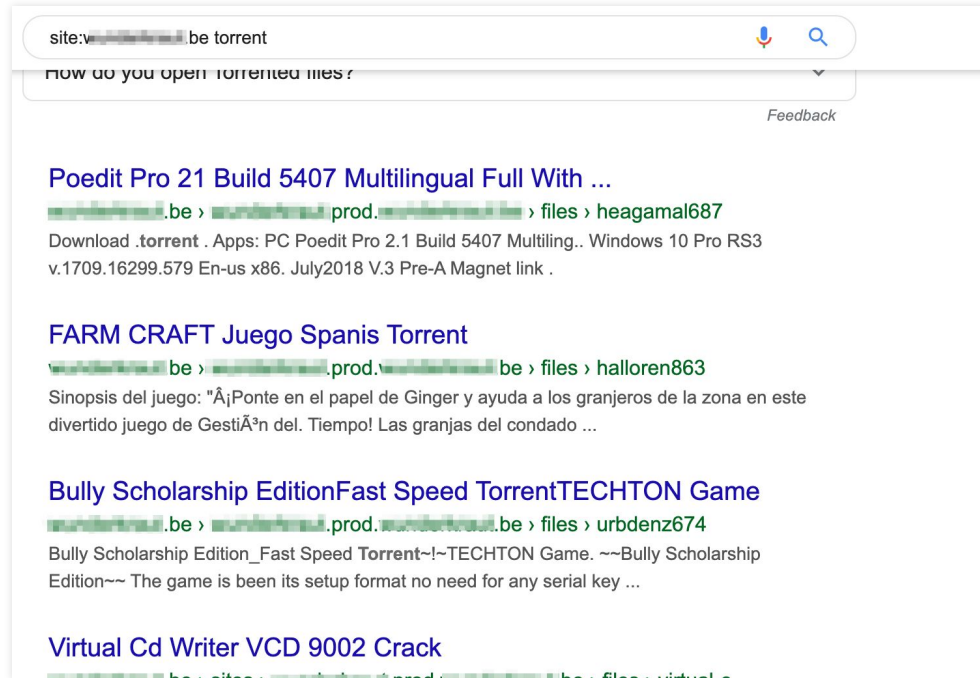
SECURITY LEAKS IMPACTING SEO

When Google notices spam on your website, they could punish you with a "manual action", potentially **lowering your organic search visibility**.

SECURITY LEAKS IMPACTING SEO

*"Google issues a manual action against a site when a **human reviewer** at Google has determined that **pages on the site are not compliant** with Google's webmaster quality guidelines"*

*"If a site has a manual action, some or all of that site **will not be shown in Google** search results."*

Source: https://support.google.com/webmasters/answer/9044175?hl=en

SECURITY LEAKS IMPACTING SEO

Use a module to implement reCAPTCHA, recaptcha, simple_recaptcha, hcaptcha, ...
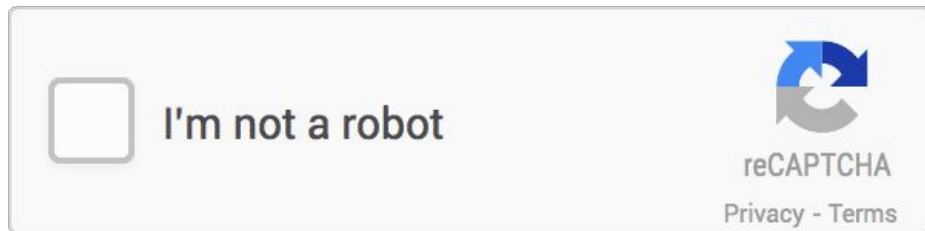
SECURITY LEAKS IMPACTING SEO

Know when to use **private** files and place them **outside** of Drupal's web root.

File system ☆

Home » Administration » Configuration » Media » File system

Public file system path

sites/default/files

A local file system path where public files will be stored. This directory must exist and be writable by Drupal. This directory must be relative to the Drupal installation directory and be accessible over the web. This must be changed in settings.php

Public file base URL

https://example.com/sites/default/files

The base URL that will be used for public file URLs. This can be changed in settings.php

Private file system path

../private/files

An existing local file system path for storing private files. It should be writable by Drupal and not accessible over the web. This must be changed in settings.php

42
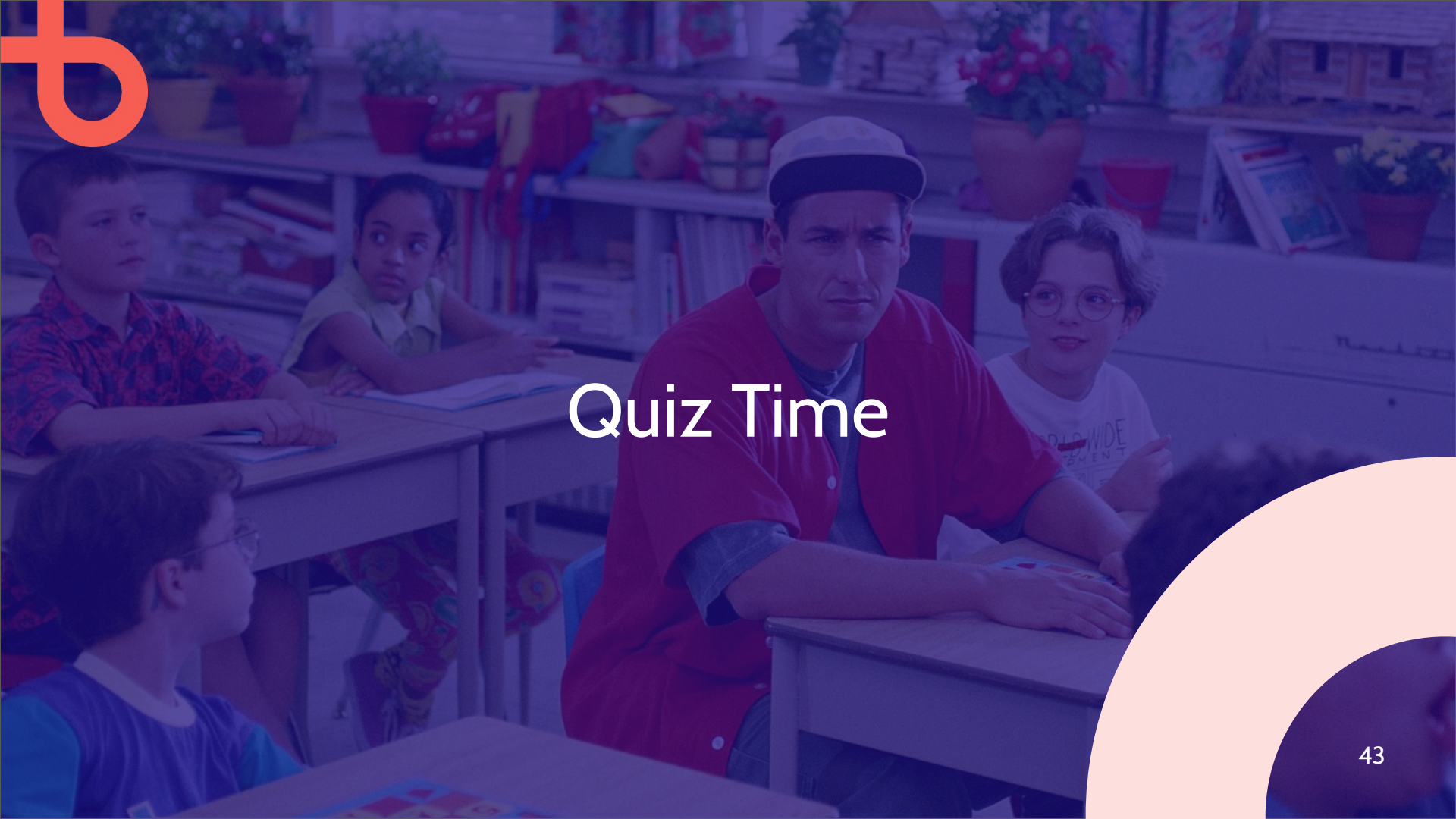
Quiz Time

QUIZ TIME!

Case: a customer noticed a lot of "old" website pages are indexed by Google and would like to remove these from Google (but *still* keep them online for his website visitors just for reference).

What would you advise:

a. Simply block the pages (or folders) via robots.txt. For example by adding one line of instruction *disallow: /old-blog/**

b. Add a "robots" meta-tag to all old pages and set them to *"noindex"*

c. Block using robots.txt **AND** add the robots noindex meta tag

QUIZ TIME!

Case: a customer noticed a lot of "old" website pages are indexed by Google and would like to remove these from Google (but *still* keep them online for his website visitors just for reference).

What would you advise:
a.   ~~Simply block the pages (or folders) via robots.txt. For example by adding one line of instruction *disallow: /old-blog/*~~
b.   Add a "robots" meta-tag to all old pages and set them to *"noindex"*
c.   ~~Block using robots.txt **AND** add the robots noindex meta tag~~

ROBOTS.TXT DISALLOW
!=
NOINDEX

ROBOTS.TXT DISALLOW != NOINDEX

Contrary to popular belief, blocking a page (or folder) via robots.txt or adding a noindex directive to the meta robots tag are **not** the same thing.

Robots.txt instructions impact **crawling**, not indexing.
Noindex directives using a meta tag impact **indexing**, not crawling.

ROBOTS.TXT DISALLOW != NOINDEX

It sounds weird, but Google is able to stumble upon a link on an external website linking to a page blocked by robots.txt, **and still index it.**

The result will most likely be a snippet in the search results without a title or description, since Google **can't read** the title or meta-description.

ROBOTS.TXT DISALLOW != NOINDEX

*"Now wait a second Wouter. I understand it **could** happen in theory a link is indexed even though it's blocked by robots.txt, but do things like this **really happen** in practice?"*

ROBOTS.TXT DISALLOW != NOINDEX

Why yes, yes they do:

ROBOTS.TXT DISALLOW != NOINDEX

So what happened here?

A weird "/home//home" page from drupal.org is indexed by Google.

Let's take a look at the robots.txt: https://www.drupal.org/robots.txt
Bingo:

```
# Googlebot picked up strange homepage URLs somewhere
Disallow: /home/*
```

They noticed "strange homepage URLs" and tried
removing this page from the Google index by blocking
Google from crawling the page.

Oops, my bad!

GOOGLE ANALYTICS HORROR

GOOGLE ANALYTICS HORROR

Correct data is very important when analysing SEO-efforts.

Pay close attention to sudden drops and spikes in Google Analytics data. Big sudden changes are often a sign of a **configuration issue.**
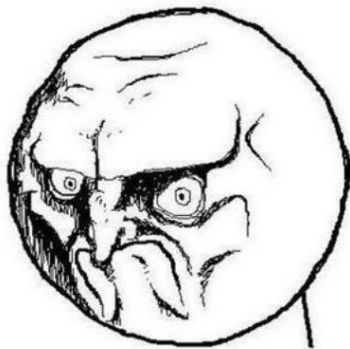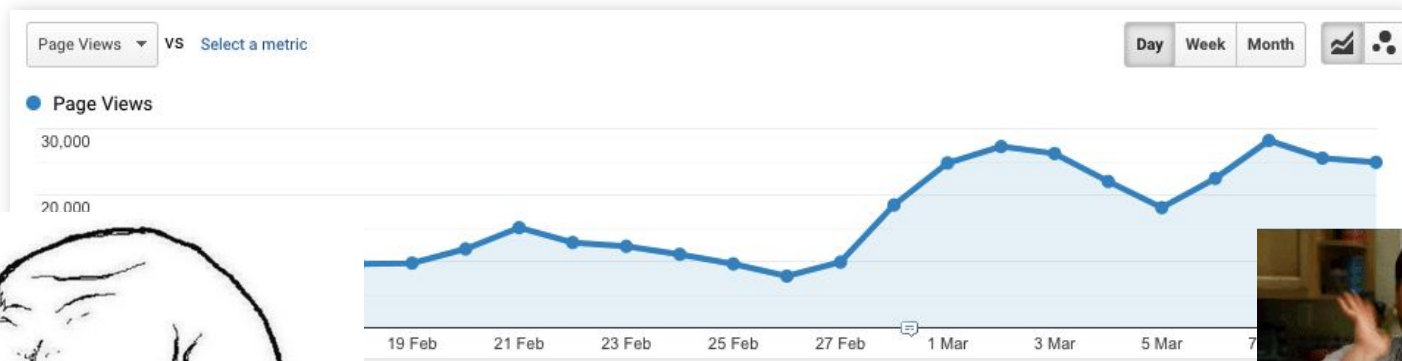
GOOGLE ANALYTICS HORROR

GOOGLE ANALYTICS HORROR

Always review Google Analytics data **after all cookie or analytics module patches and configuration updates** (right after + 24h and 48h later).

Things to check:
- Only execute GA when analytical cookies are accepted (GDPR)
- Fire a GA pageview **as soon as visitors accept** the cookies
- Review (in GA) to check if anything is out of the ordinary
  - Session count
  - Bounce rate
  - % of direct traffic
  - Avg. pages/session

Visual stability issues (CLS)

VISUAL STABILITY ISSUES (CLS)

Cumulative Layout Shift (CLS) measures the **visual stability** of a page, which is an important user experience (and SEO) factor (mobile + desktop)

Elements *jumping around* on a page during pageLoad have a negative impact on the CLS metric.

More info on CLS: https://web.dev/cls/

CLS is part of a bigger group of "Core Web Vitals", which is in turn part of the "page experience" metrics.
More info: https://developers.google.com/search/docs/advanced/experience/page-experience

VISUAL STABILITY ISSUES (CLS)

VISUAL STABILITY ISSUES (CLS)

Top causes of CLS issues:

- Hero images
- Images in the body content
- CSS animations
- Ads loading/popping in during pageLoad or when scrolling
- External plugins (like notification bars at the top of a page)

VISUAL STABILITY ISSUES (CLS)

When using Blazy, Drimage, custom images this is no issue.

When integrating **custom functionalities** that load through ajax or writing some JavaScript, make sure to test this and add a fixed height in CSS whenever possible.

VISUAL STABILITY ISSUES (CLS)

In general:

- **Images are the most frequent CLS culprit**. Make sure to *reserve space* for all images so the layout doesn't shift when they load in

- Review all pages for visual stability
    - Manually (emulate a slow internet connection or use **Lighthouse**)
    - **Or** via crawling software (like ScreamingFrog, free to use for small websites) connected to PageSpeed Insights to automatically measure **all** pages of a website
    - **Or** by following up the notifications in Google Search Console

61

IMAGES IMPACTING PERFORMANCE

IMAGES IMPACTING PERFORMANCE

Faster is always better. For SEO and for UX.

IMAGES IMPACTING PERFORMANCE

Content editors often upload high quality images that (if not handled correctly) can **heavily** impact site performance in a negative way.

IMAGES IMPACTING PERFORMANCE

Make sure the image shown to the end client is as small as possible
- At least use drupal core image styles
- Enhance with blazy and/or drimage
- Conversion to *webp* (will be replaced by avif in the future)
  - *Webp* not supported by IE11
  - *Avif* almost nowhere supported
- Check the compression in Image toolkit (default is ok, 75%)
  - Can be enhanced with Image optimize

IMAGES IMPACTING PERFORMANCE

And also...

- Educate your customers and content-editors!
  - Explain what compression is and how images can impact site speed
  - Explain what image types to use and when to use them
  - Explain how they could manually compress images using free online tools like Squoosh if they want to optimize as much as possible

- Make sure static (fixed) images are optimized!
  (site logo, images in the footer, ...)

RAPID FIRE BEST PRACTICES

RAPID-FIRE BEST PRACTICES   #1

Aggregate and minify CSS and JS files whenever possible.
Less HTTP requests = good.

https://www.drupal.org/project/advagg

RAPID-FIRE BEST PRACTICES #2

Use [Pathauto module](#) to generate nice urls for your pages.
Don't forget to translate them as well.

RAPID-FIRE BEST PRACTICES   #2

70

RAPID-FIRE BEST PRACTICES   #2

RAPID-FIRE BEST PRACTICES   #3

Make sure you use all the available tools possible to increase the speed of your website.

Examples:
- Default drupal caching
- Varnish (More info https://dropsolid.io/knowledge-hub/memcache-drupal-8)
- Memcached (More info https://dropsolid.io/knowledge-hub/varnish-drupal-8)
- Redis (More info https://www.drupal.org/project/redis)

RAPID-FIRE BEST PRACTICES   #4

Make sure each page has a **correct/well-configured** canonical tag.

Depending on the complexity of the website you might want to let a SEO specialist review the canonicals. There can be a lot of small issues depending on your setup.

RAPID-FIRE BEST PRACTICES   #4

Canonical URL best practices:

- Always use **absolute** URLs
- Make sure **all canonicals return a 200**
  - E.g. don't set a canonical to "www" if this URL redirects to the non-www version
- In **most** cases you want to **omit URL params** from the canonical
  - E.g. a webshop facet filter on price that modifies the URL
  - **Don't** omit *paging* parameters (unless you have a "view-all" page)

RAPID-FIRE BEST PRACTICES #4

Also:

- Canonicals should never point to /node/ URLs but rather to "clean" URLs
- /node/ URLs should probably not be indexable either
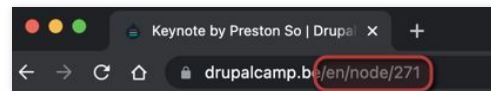
RAPID-FIRE BEST PRACTICES   #5

Multilingual sites should be translated 100% whenever possible

- Is it possible to **translate everything?**
- **Deny access to untranslated pages** (content_language_access module)

RAPID-FIRE BEST PRACTICES   #5

This is not ideal for a couple of reasons...

It could scare away non-English speakers.

Confused Googlebot is confused.

RAPID-FIRE BEST PRACTICES   #6

Follow-up on the amount of pages indexed by Google.

via Google Search Console
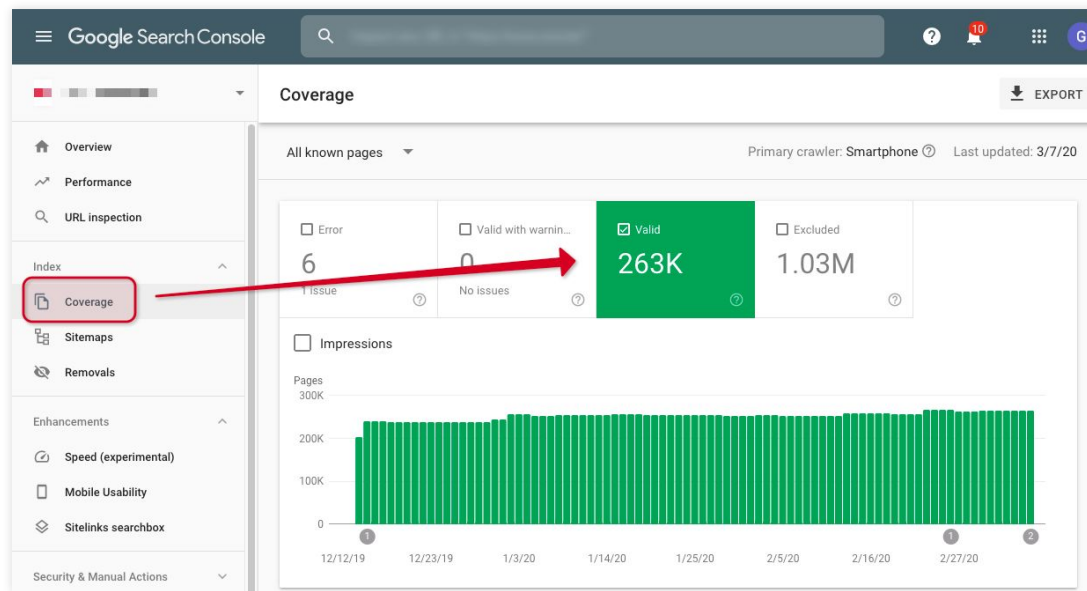
RAPID-FIRE BEST PRACTICES   #6

Seems **too high?**

Maybe some rabbithole setup is needed to remove separate pages from the index.

Seems **too low?**

Maybe important pages are no-indexed for some reason?

RAPID-FIRE BEST PRACTICES   #7

Make sure your base url is not *"http://default/"* in your sitemap XML,
we've seen this issue occur **a lot** on live sites.

Setup your cron job correctly, more info:
https://www.drupal.org/project/xmlsitemap/issues/1944518

RAPID-FIRE BEST PRACTICES   #7

Submitted sitemaps

Sitemap	Type	Submitted ↓	Last read	Status	Discovered URLs	
/sitemap.xml	Sitemap	Feb 17, 2017	Mar 10, 2020	Submitted URL not found (404)	100	

Rows per page: 10 ▾   1-1 of 1   < >

81

RAPID-FIRE BEST PRACTICES   #7
The best ways to fix this issue

- Set this correctly for the sitemap module

Default base URL

https://www.myawesomesite.be

On some hosting providers it is impossible to pass parameters to cron to tell
Drupal which URL to bootstrap with. In this case the base URL of sitemap links
can be overridden here.
Example: *https://www.myawesomesite.be*

- Updating your cron job to include your website url

```
drush cron --uri=http://www.example.com
```

RAPID-FIRE BEST PRACTICES   #8

Use the Google Analytics module (google_analytics) or the Google Tag Manager module (google_tag).

Not both.

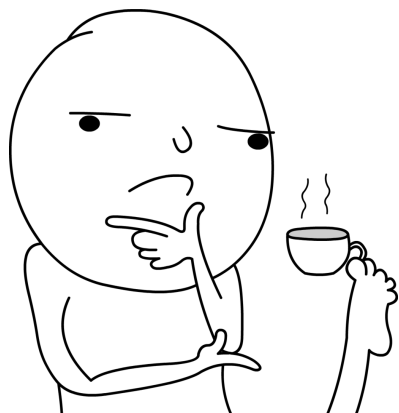# RAPID-FIRE BEST PRACTICES   #8

Using both modules makes your tracking setup **prone to errors.**

RAPID-FIRE BEST PRACTICES #8

I wonder what happened here…

QUESTIONS?

Contact Brent
brent@dropsolid.com
https://www.linkedin.com/in/brentgees/
https://www.drupal.org/u/brentg
https://github.com/brentgees/phpstorm-live-templates-drupal

Contact Wouter
wouter@dropsolid.com
https://www.linkedin.com/in/wouterdebruycker/

Thank you

To our wonderful sponsors, our awesome community and fantastic volunteers!
Platinum sponsors

Thank you

Gold sponsors

Stay in touch

#ddd2022 on Drupal slack

@drupaldevdays

/drupaldevdays