THE SUPER SEO GAME

Webinar

Thinking BIG: Big Solutions for Big Sites

oncrawl

GAMEPLAY & RULES

- Earn points by signing up, attending, and participating
- Unlock new levels, earn badges and check our leaderboard
- Use #SuperSEOGame to continue the conversation
- Have fun!

SINGLE PLAYER of the day

Rebecca Berbel

Product Marketing Manager @Oncrawl

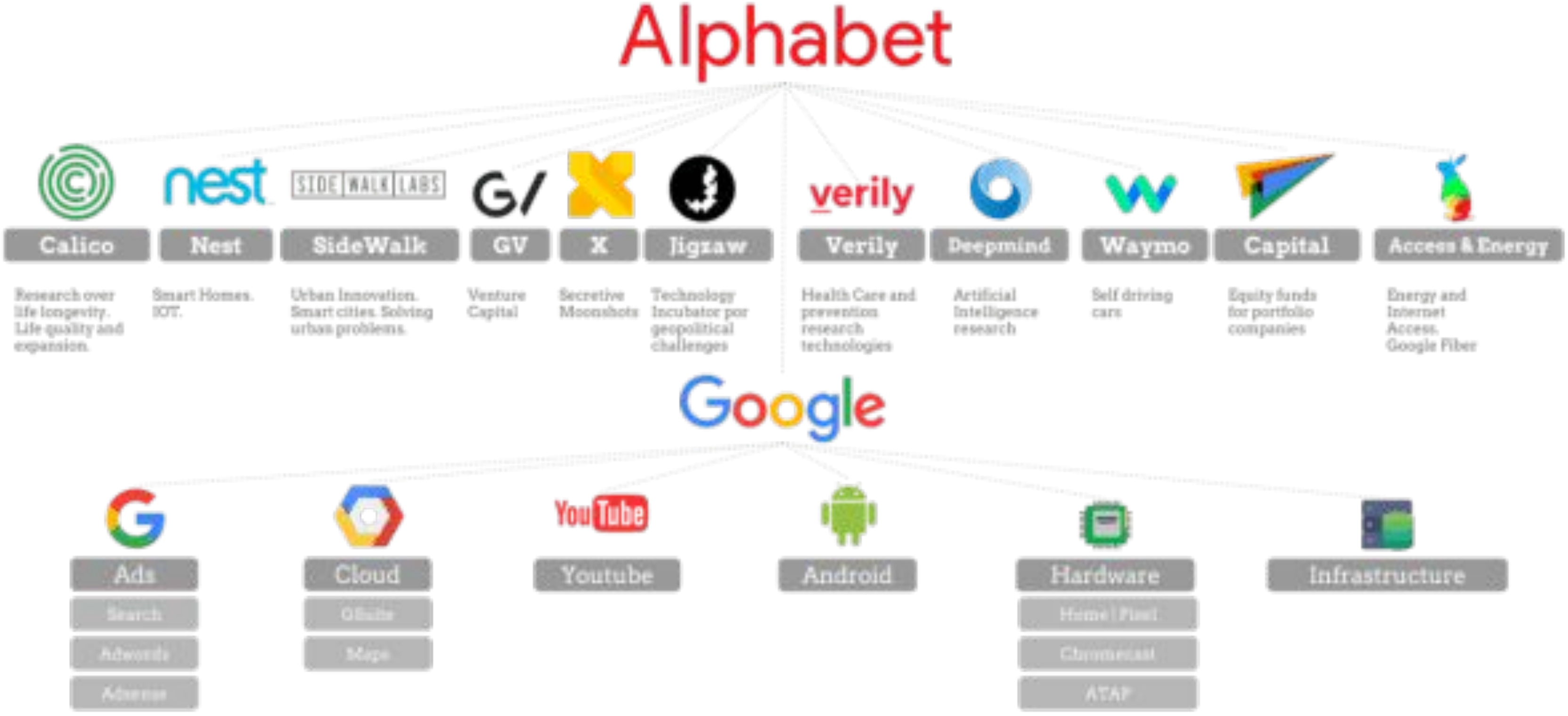
@rebberbel

The BIG picture

1. Definitions & context
2. Technical constraints & solutions
3. Challenges in SEO strategy
4. Focus areas for successful SEO

What is a **big** site?

What do big sites have in **common**?



Alphabet

Calico
Research over life longevity. Life quality and expansion.

Nest
Smart Homes. IOT.

SideWalk
Urban Innovation. Smart cities. Solving urban problems.

GV
Venture Capital

X
Secretive Moonshots

Jigzaw
Technology Incubator por geopolitical challenges

Verily
Health Care and prevention research technologies

Deepmind
Artificial Intelligence research

Waymo
Self driving cars

Capital
Equity funds for portfolio companies

Access & Energy
Energy and Internet Access. Google Fiber

Google

Ads
Search
Adwords
Adsense

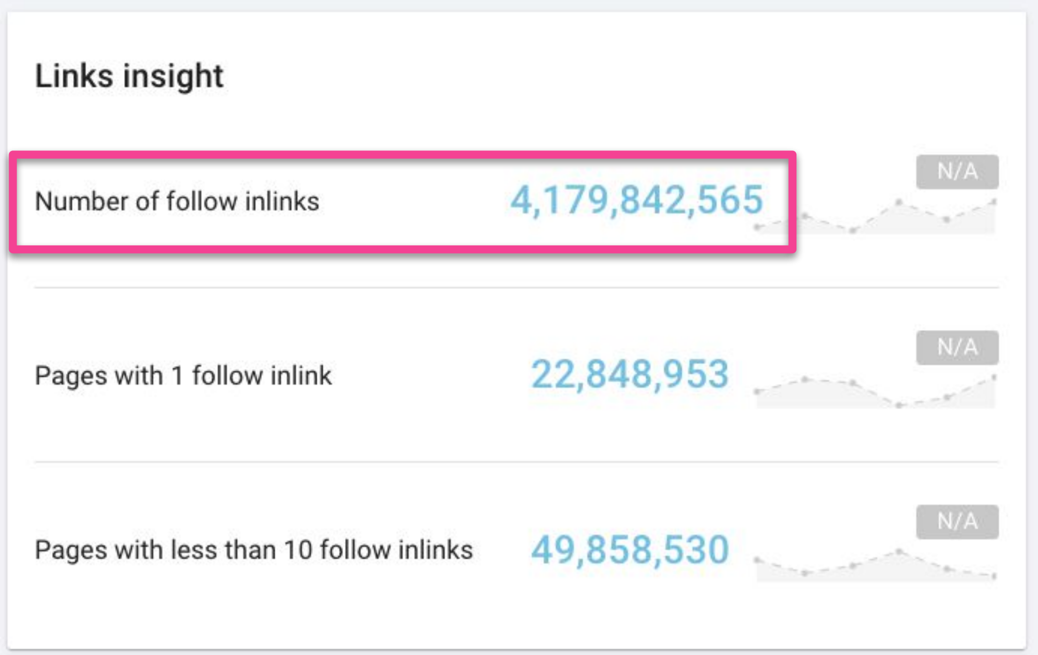
Cloud
GSuite
Maps

Youtube

Android

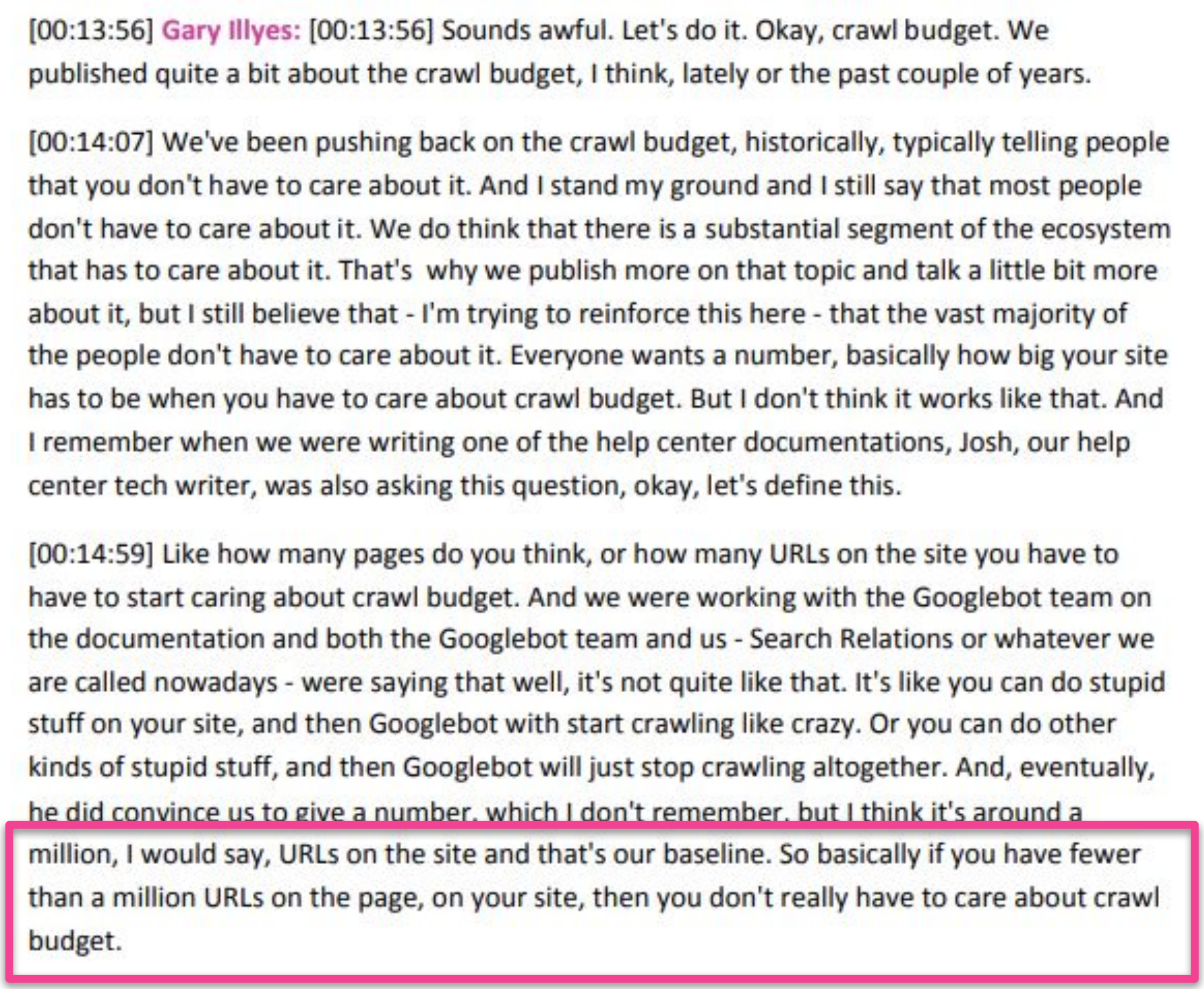
Hardware
Home | Pixel
Chromecast
ATAP

Infrastructure



Links insight

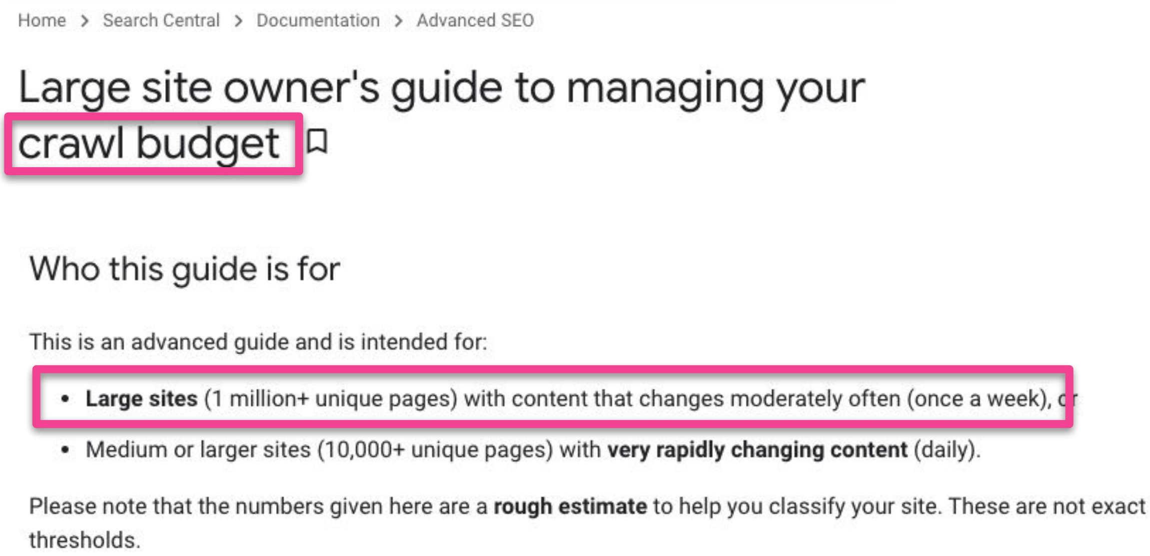
Number of follow inlinks	4,179,842,565	N/A
Pages with 1 follow inlink	22,848,953	N/A
Pages with less than 10 follow inlinks	49,858,530	N/A



[00:13:56] Gary Illyes: [00:13:56] Sounds awful. Let's do it. Okay, crawl budget. We published quite a bit about the crawl budget, I think, lately or the past couple of years.

[00:14:07] We've been pushing back on the crawl budget, historically, typically telling people that you don't have to care about it. And I stand my ground and I still say that most people don't have to care about it. We do think that there is a substantial segment of the ecosystem that has to care about it. That's why we publish more on that topic and talk a little bit more about it, but I still believe that - I'm trying to reinforce this here - that the vast majority of the people don't have to care about it. Everyone wants a number, basically how big your site has to be when you have to care about crawl budget. But I don't think it works like that. And I remember when we were writing one of the help center documentations, Josh, our help center tech writer, was also asking this question, okay, let's define this.

[00:14:59] Like how many pages do you think, or how many URLs on the site you have to have to start caring about crawl budget. And we were working with the Googlebot team on the documentation and both the Googlebot team and us - Search Relations or whatever we are called nowadays - were saying that well, it's not quite like that. It's like you can do stupid stuff on your site, and then Googlebot with start crawling like crazy. Or you can do other kinds of stupid stuff, and then Googlebot will just stop crawling altogether. And, eventually, he did convince us to give a number, which I don't remember, but I think it's around a million, I would say, URLs on the site and that's our baseline. So basically if you have fewer than a million URLs on the page, on your site, then you don't really have to care about crawl budget.



Home > Search Central > Documentation > Advanced SEO

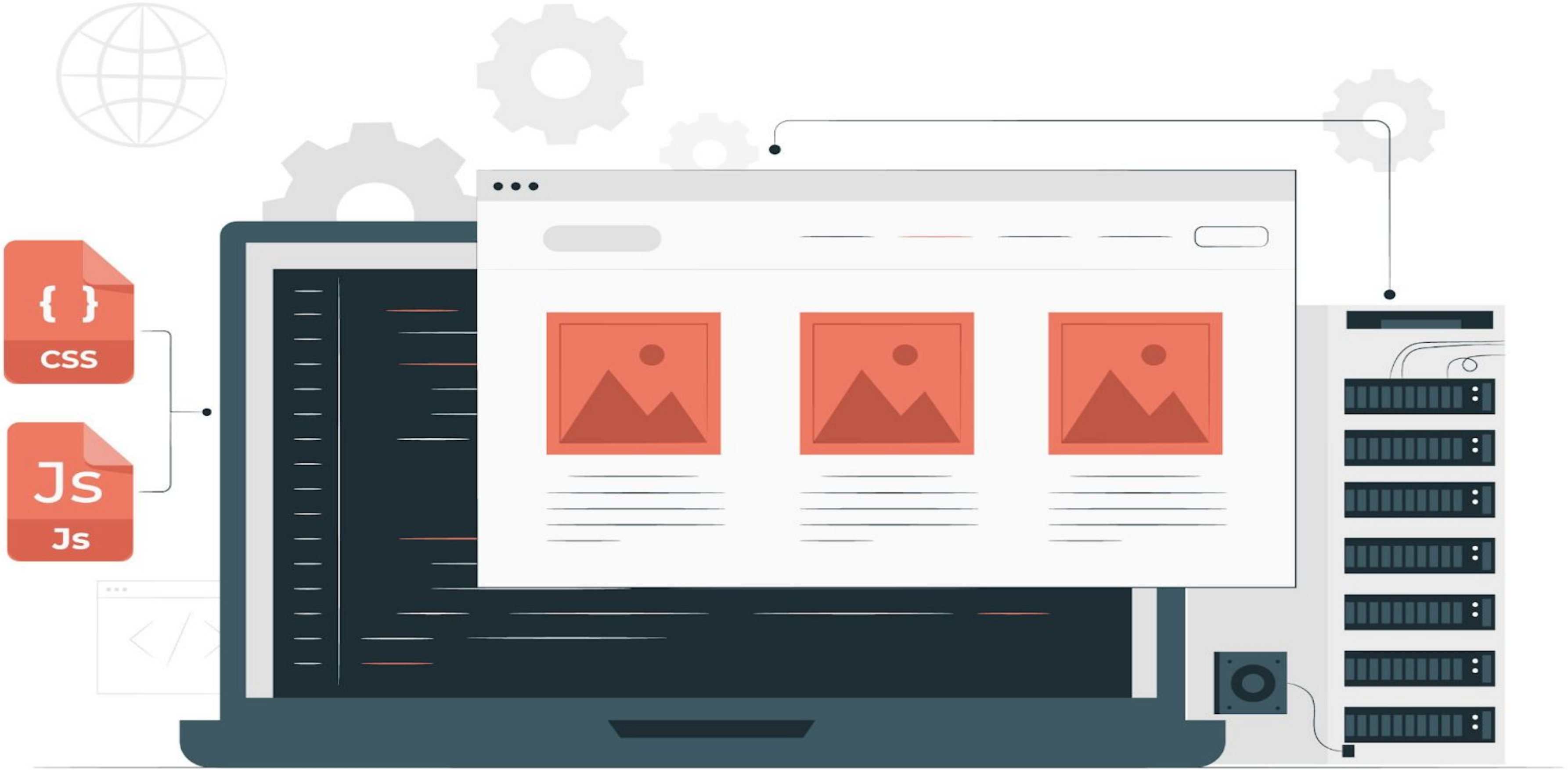
Large site owner's guide to managing your crawl budget

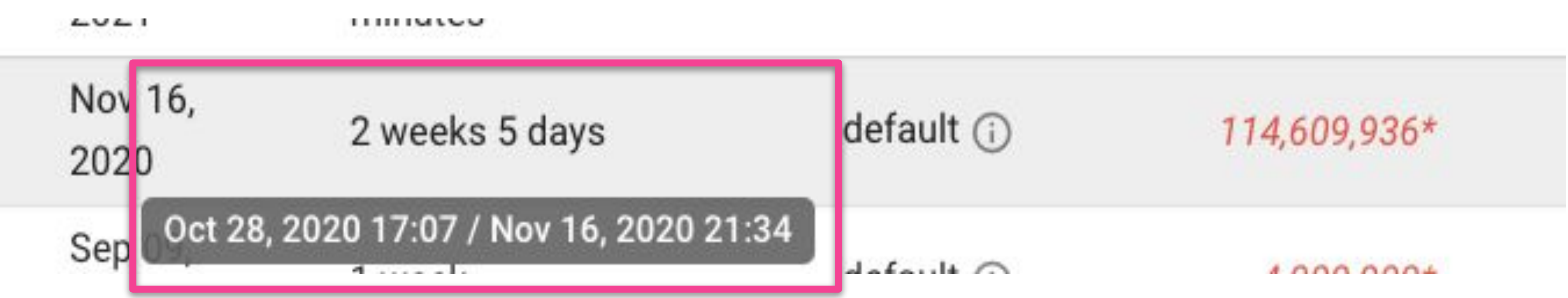
Who this guide is for

This is an advanced guide and is intended for:

- **Large sites** (1 million+ unique pages) with content that changes moderately often (once a week), or
- Medium or larger sites (10,000+ unique pages) with **very rapidly changing content** (daily).

Please note that the numbers given here are a **rough estimate** to help you classify your site. These are not exact thresholds.


{ }
CSS
Js
Js



Nov 16, 2020 | 2 weeks 5 days | default | 114,609,936*

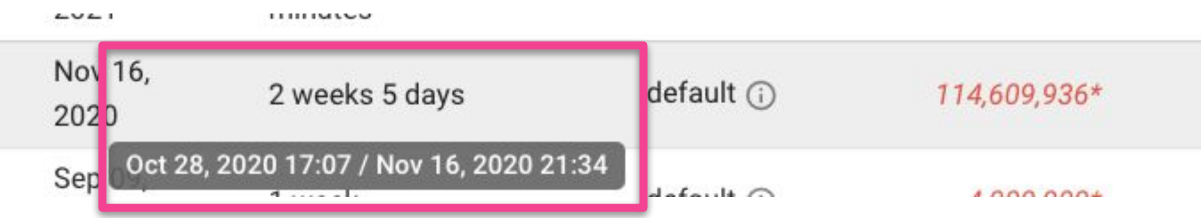
Oct 28, 2020 17:07 / Nov 16, 2020 21:34

Sep

How can see the **big** picture?

1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data
5. Retrieving data
6. Interpreting data

Why is it hard to get a view of the whole site?



Nov 16, 2020	2 weeks 5 days	default ⓘ	*114,609,936**

Oct 28, 2020 17:07 / Nov 16, 2020 21:34

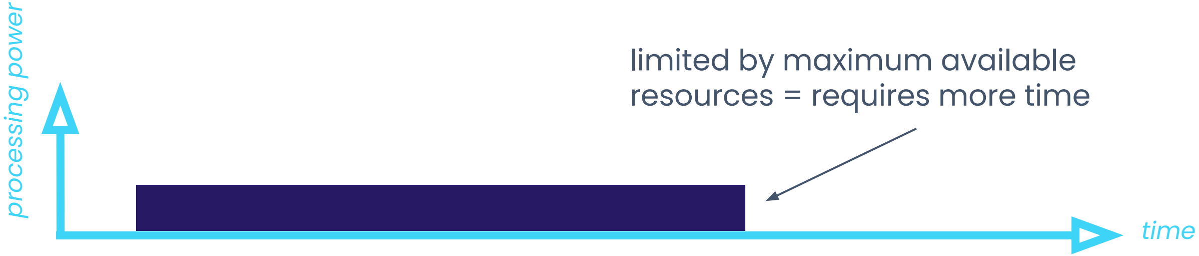
How do you reduce the time it takes to crawl?

⚠️ **Site infrastructure**

⚠️ **Processing power**

1 Collecting data

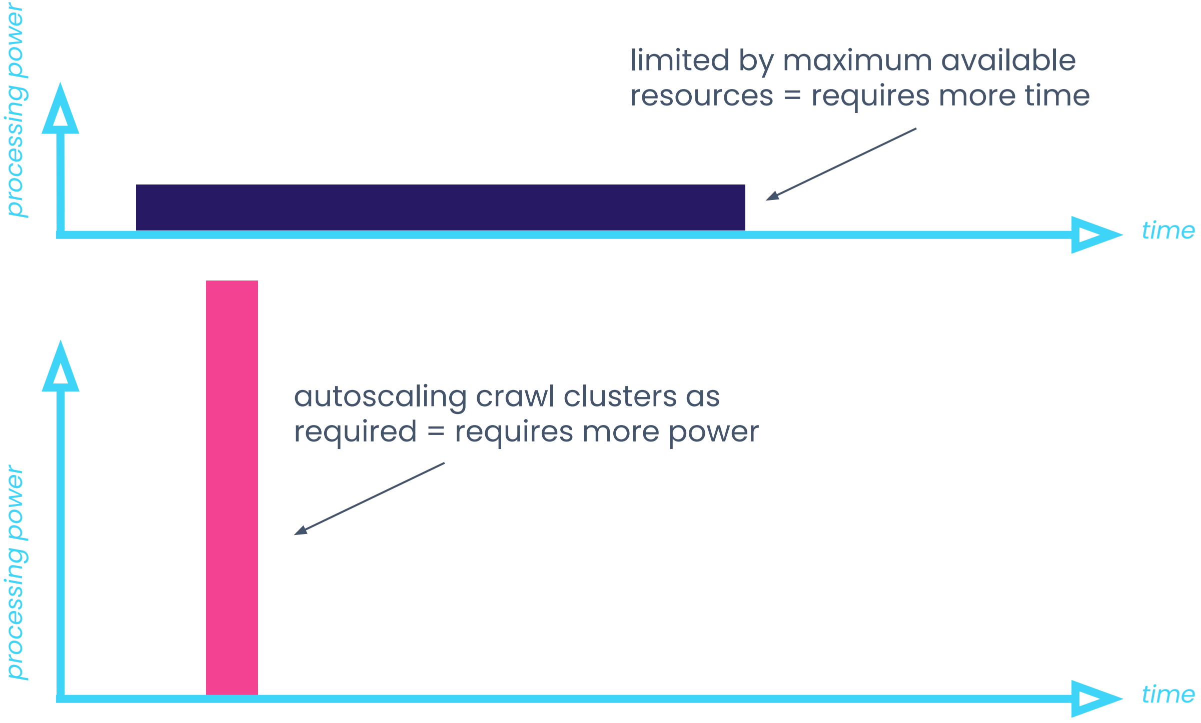
How do you reduce the time it takes to crawl?



limited by maximum available resources = requires more time

1 Collecting data

How do you reduce the time it takes to crawl?



limited by maximum available resources = requires more time

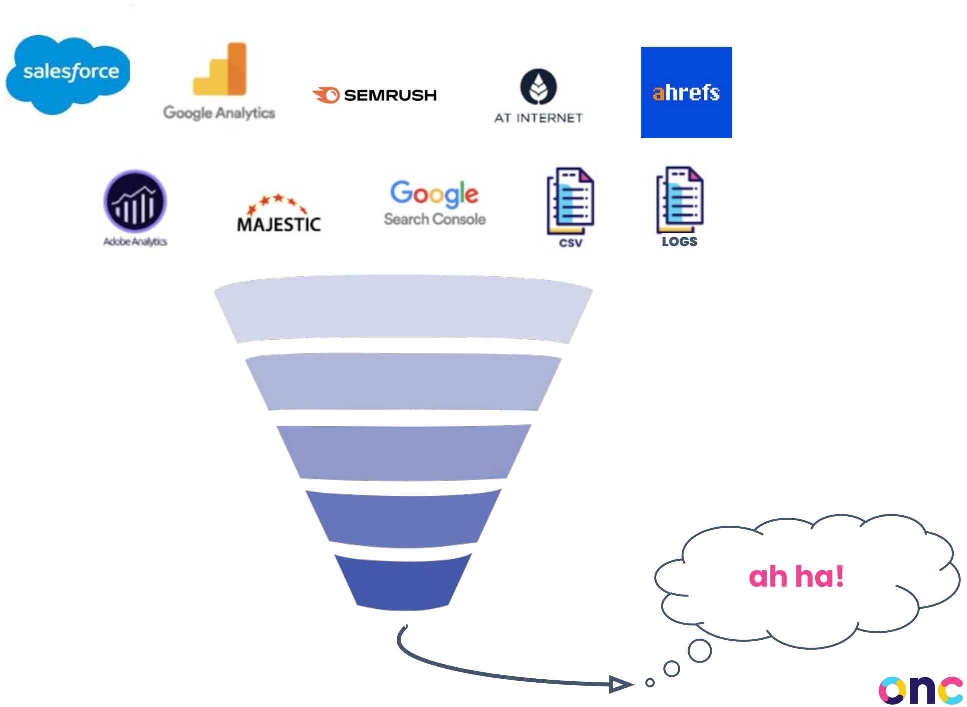
autoscaling crawl clusters as required = requires more power

> “Google Kubernetes Engine enables us to tackle gigantic projects immediately, allowing our customers to crawl even **hundreds of millions of JavaScript pages** without prior notice. We’ve seen our JavaScript crawler scale up from 10 to 750 pods in just a few minutes.”

—Philippe David, Chief Technology Officer, Oncrawl

1 Collecting data

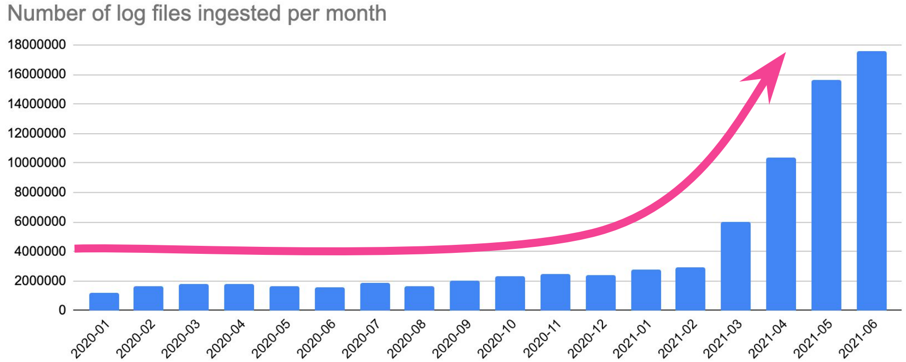
Why is it hard to get a view of the whole site?



1 Collecting data

Why is it hard to get a view of the whole site?

LOGS



1 Collecting data

2 Preparing data

Managing data for the full site

- **Extracting**
- **Structuring**
- **Sorting**
- **Filtering**
- **Grouping**

1 Collecting data

2 Preparing data

100 000 000 (pages)

× **100** (links per page)

10 000 000 000 (links)

1 Collecting data

2 Preparing data

100 000 000 (pages)

$$= (100\,000\,000 \times 99\,999\,999) \div 2$$

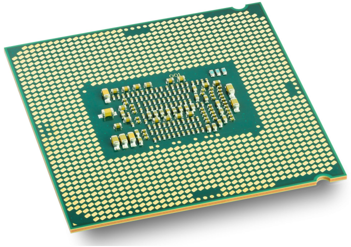
$$\cong 100\,000\,000^2$$

10 000 000 000 000 000

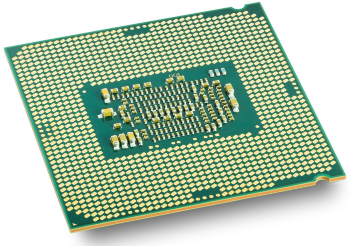
(duplication analyses)

1 Collecting data

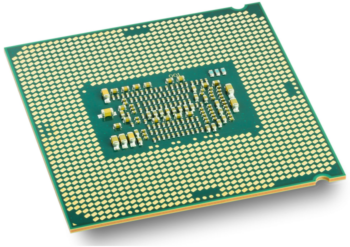
2 Preparing data



× 1



× 2



× 3

3 300 000 000 000 000 000

1 Collecting data

2 Preparing data

BIG data

= sorting + hash + logarithms

~~3 300 000 000 000 000 000~~

800 000 000

1 Collecting data

2 Preparing data

3 Analyzing data

Metrics that add **BIG** value

- ✓ **Inrank**
- ✓ **Duplicate content**

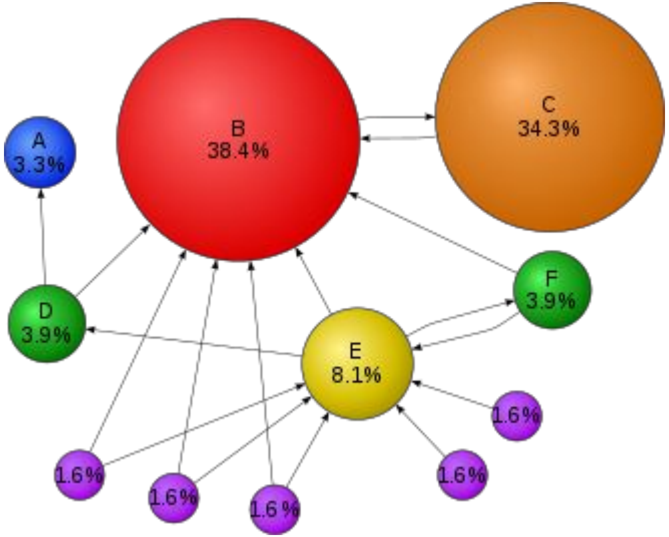
oncrawl

1 Collecting data

2 Preparing data

3 Analyzing data

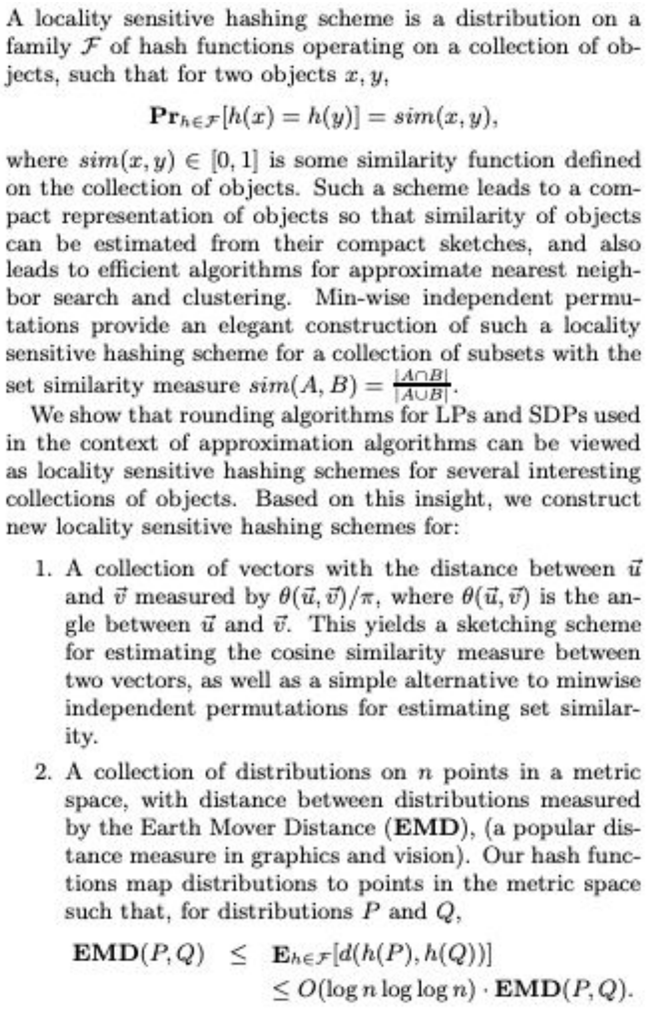
$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$



1 Collecting data

2 Preparing data

3 Analyzing data



A locality sensitive hashing scheme is a distribution on a family $\mathcal{F}$ of hash functions operating on a collection of objects, such that for two objects x, y,

$$\mathbf{Pr}_{h\in\mathcal{F}}[h(x) = h(y)] = sim(x, y),$$

where $sim(x, y) \in [0, 1]$ is some similarity function defined on the collection of objects. Such a scheme leads to a compact representation of objects so that similarity of objects can be estimated from their compact sketches, and also leads to efficient algorithms for approximate nearest neighbor search and clustering. Min-wise independent permutations provide an elegant construction of such a locality sensitive hashing scheme for a collection of subsets with the set similarity measure $sim(A, B) = \frac{|A\cap B|}{|A\cup B|}$.

We show that rounding algorithms for LPs and SDPs used in the context of approximation algorithms can be viewed as locality sensitive hashing schemes for several interesting collections of objects. Based on this insight, we construct new locality sensitive hashing schemes for:

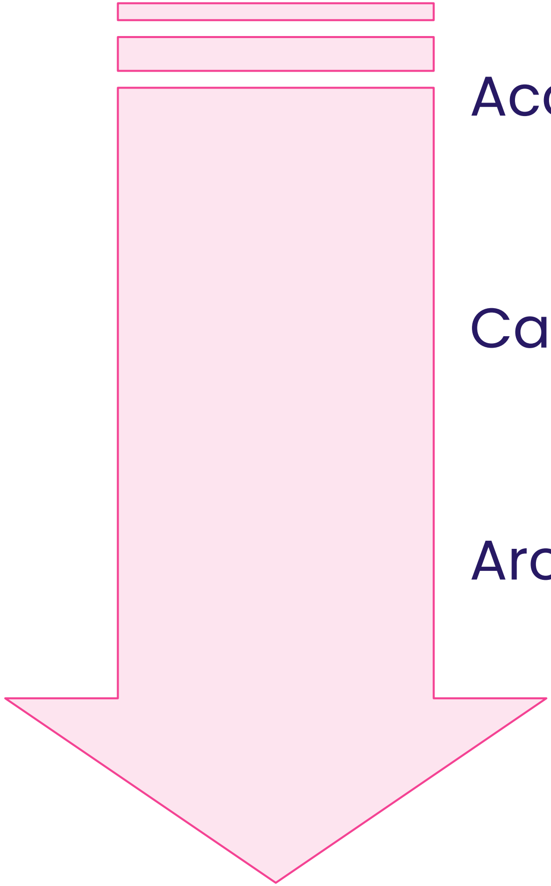
1. A collection of vectors with the distance between $\vec{u}$ and $\vec{v}$ measured by $\theta(\vec{u}, \vec{v})/\pi$, where $\theta(\vec{u}, \vec{v})$ is the angle between $\vec{u}$ and $\vec{v}$. This yields a sketching scheme for estimating the cosine similarity measure between two vectors, as well as a simple alternative to minwise independent permutations for estimating set similarity.

2. A collection of distributions on n points in a metric space, with distance between distributions measured by the Earth Mover Distance (**EMD**), (a popular distance measure in graphics and vision). Our hash functions map distributions to points in the metric space such that, for distributions P and Q,

$$\begin{aligned}\mathbf{EMD}(P, Q) &\leq \mathbf{E}_{h\in\mathcal{F}}[d(h(P), h(Q))] \\ &\leq O(\log n \log\log n) \cdot \mathbf{EMD}(P, Q).\end{aligned}$$

SimHash (Moses S. Charikar)

1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data



Accessible information

Calculatable information

Archived information

1 Collecting data

2 Preparing data

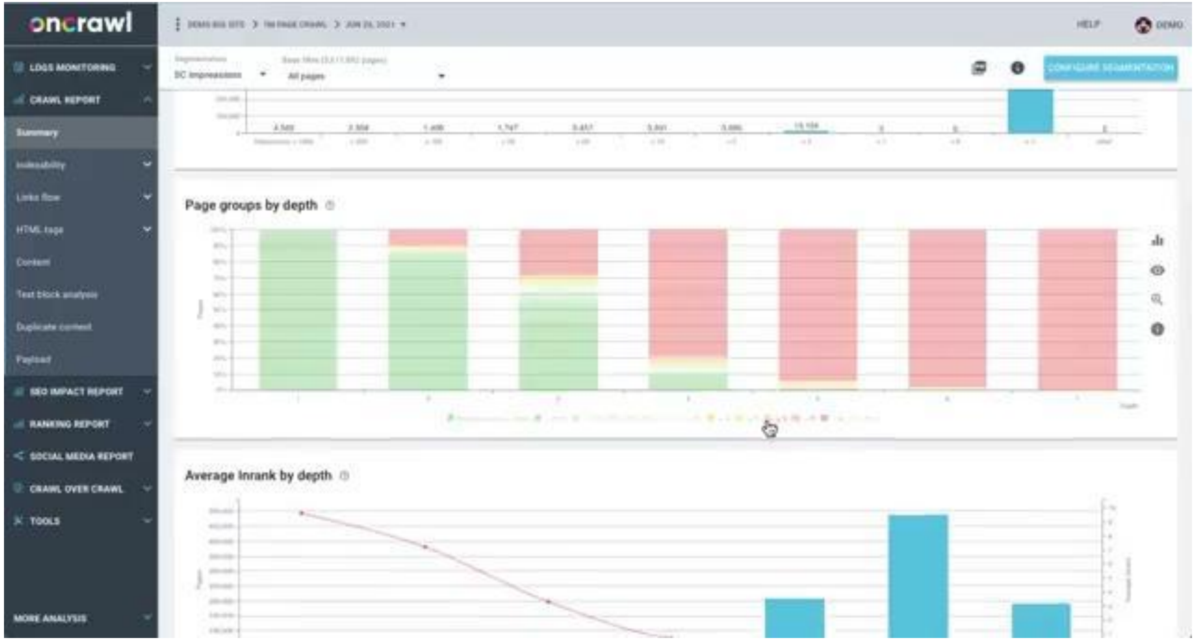
3 Analyzing data

4 Storing data

> If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. **The contents of the data lake stream in from a source** to fill the lake, and various users of the lake can come to **examine**, **dive in**, or **take samples**.

—James Dixon, then Chief Technology Officer, Pentaho

1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data
5. Retrieving data



1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data
5. Retrieving data

1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data
5. Retrieving data

> We use one of the best systems – ElasticSearch – to store, index and retrieve your data, for **over 1400 native metrics** as well as additional connectors, ingested or scraped data, etc. Our goal is **fast and reliable retrieval**.

—Vincent Terrasi, Data SEO Educator & Product Manager, Oncrawl

1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data
5. Retrieving data

Accessibility

oncrawl

1 Collecting data

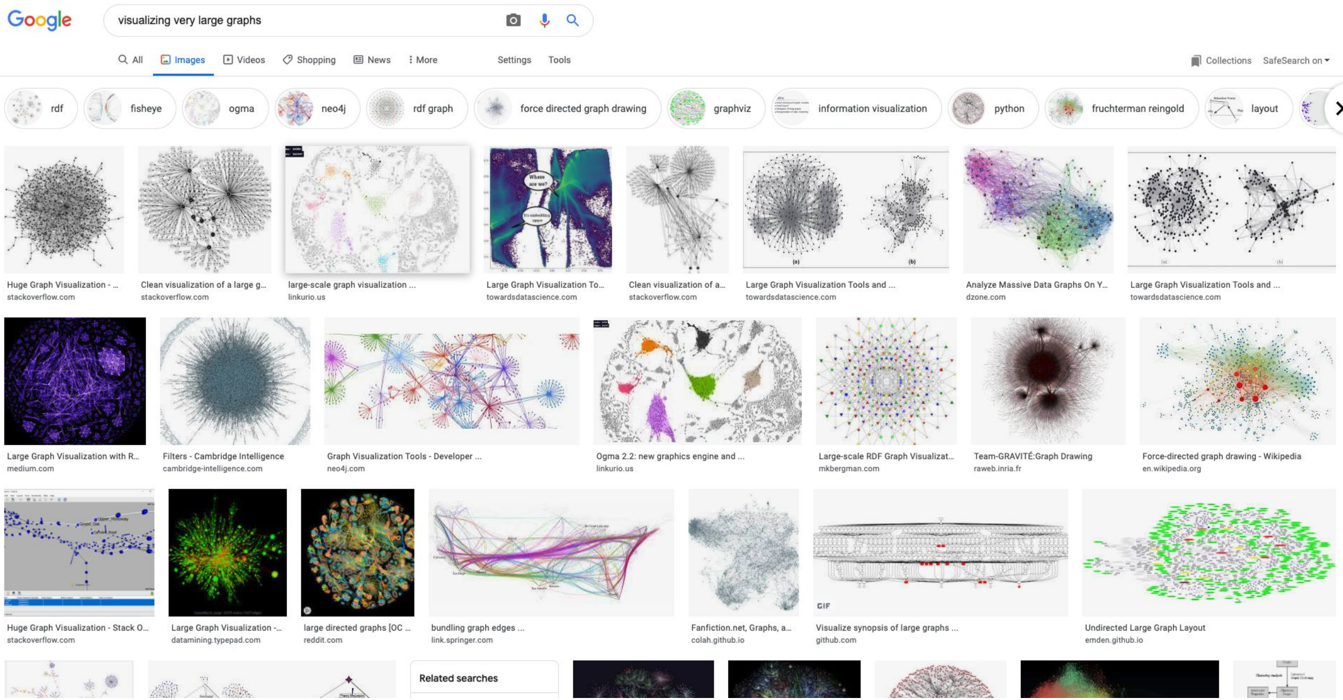
2 Preparing data

3 Analyzing data

4 Storing data

5 Retrieving data

6 Interpreting data



visualizing very large graphs

Huge Graph Visualization - ... stackoverflow.com

Clean visualization of a large g... stackoverflow.com

large-scale graph visualization ... linkurio.us

Large Graph Visualization To... towardsdatascience.com

Clean visualization of a... stackoverflow.com

Large Graph Visualization Tools and ... towardsdatascience.com

Analyze Massive Data Graphs On Y... dzone.com

Large Graph Visualization Tools and ... towardsdatascience.com

Large Graph Visualization with R... medium.com

Filters - Cambridge Intelligence cambridge-intelligence.com

Graph Visualization Tools - Developer ... neo4j.com

Ogma 2.2: new graphics engine and ... linkurio.us

Large-scale RDF Graph Visualizat... mkbergman.com

Team-GRAVITÉ:Graph Drawing raweb.inria.fr

Force-directed graph drawing - Wikipedia en.wikipedia.org

Huge Graph Visualization - Stack O... stackoverflow.com

Large Graph Visualization -... datamining.typepad.com

large directed graphs [OC ... reddit.com

bundling graph edges ... link.springer.com

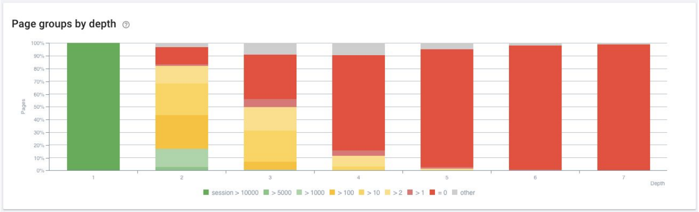
Fanfiction.net, Graphs, a... colah.github.io

Visualize synopsis of large graphs ... github.com

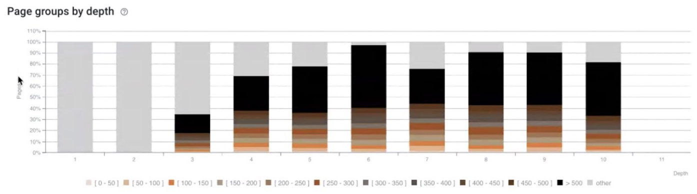
Undirected Large Graph Layout emden.github.io

Related searches

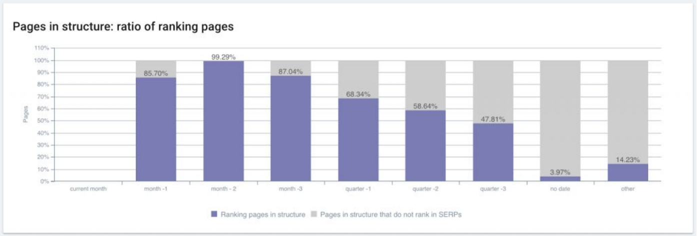
1. Collecting data
2. Preparing data
3. Analyzing data
4. Storing data
5. Retrieving data
6. Interpreting data



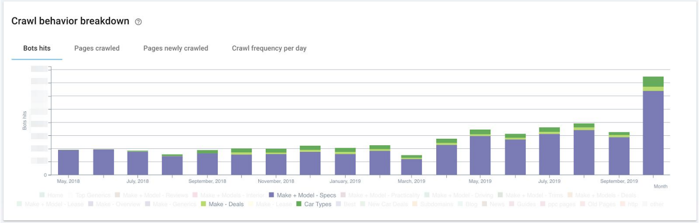
By number of GA sessions
green = pages with most sessions



By profit margin
darker = greater margin



By publication date
lefthand columns = most recent



By page template margin
each color = different template

1 Collecting data

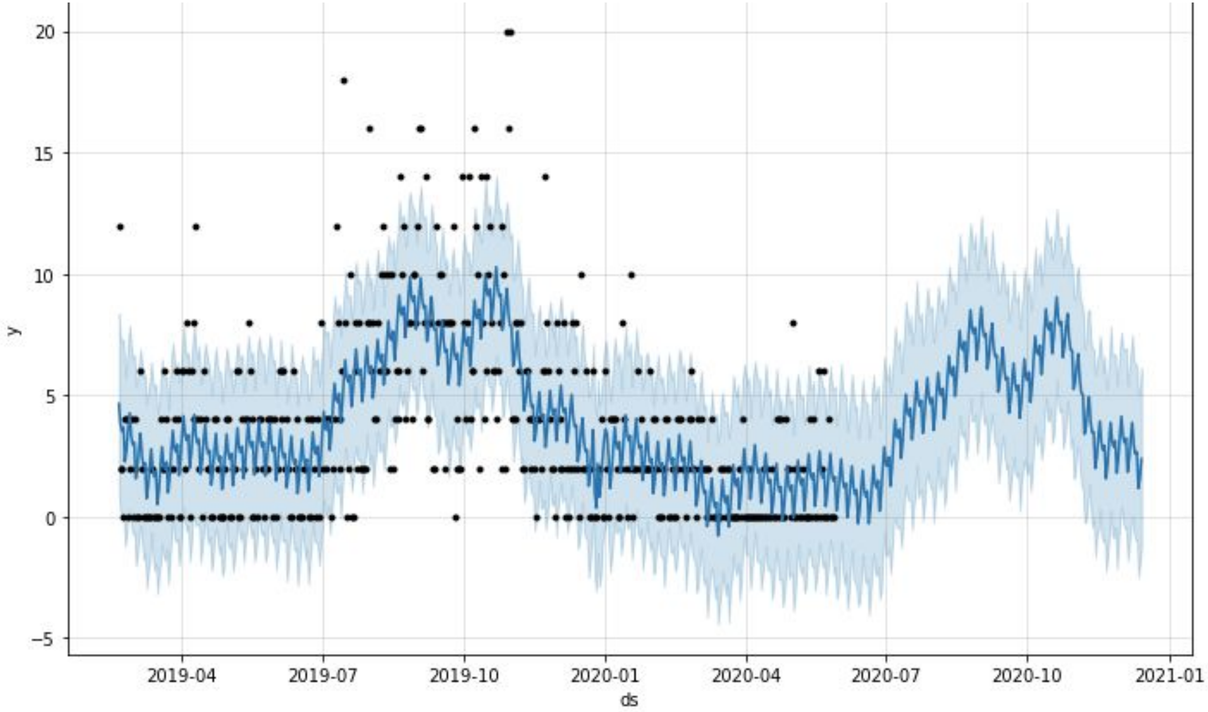
2 Preparing data

3 Analyzing data

4 Storing data

5 Retrieving data

6 Interpreting data



oncrawl

What does it mean to think **big**?

Get the right support for SEO

Use data to illustrate:

- **Low risk**: the project is not complex or hard to predict
- **Low cost**: the technical solution envisioned is reasonable
- **Easy to measure/monitor**: quantitative data is available to measure the actions and the results
- **Impactful**: there is a business benefit to be gained from the success of this project

I could pitch this project as the first step to **improve the ROI of all of our future SEO projects**. At the **C-level** when I present my ideas, I can tell that there’s a lot of confidence. I’d still like to bring the data to the table, because you can’t make good decisions without the data to back it up. But when I say, ‘here’s what I’d like to do next,’ the response now is mostly, ‘that sounds like **a good direction** to move in.’

— Murat Yatağan, VP of Growth, Brainly

Define strategy adapted to your budget

Set up a sustainable crawl strategy

- **Infrequent massive crawls**: Obtain and maintain a view of the full site
- **Regular smaller crawls**: Monitor and improve
  - Top levels of the site
  - Limited to a certain template
  - Limited by parameters
  - Lists of important pages
  - Lists of pages with known problems
  - Limited to areas where changes have occurred

Set up a sustainable crawl strategy

- **Representative crawls**: Can you generalize? What are the limitations?
- **Setting up small(er) crawls**: How do you guide a crawler? Why?
  - URL lists
  - Links that are "followed" by the crawler
  - Subdomains
  - Robots.txt rules

Where are the **big** wins?

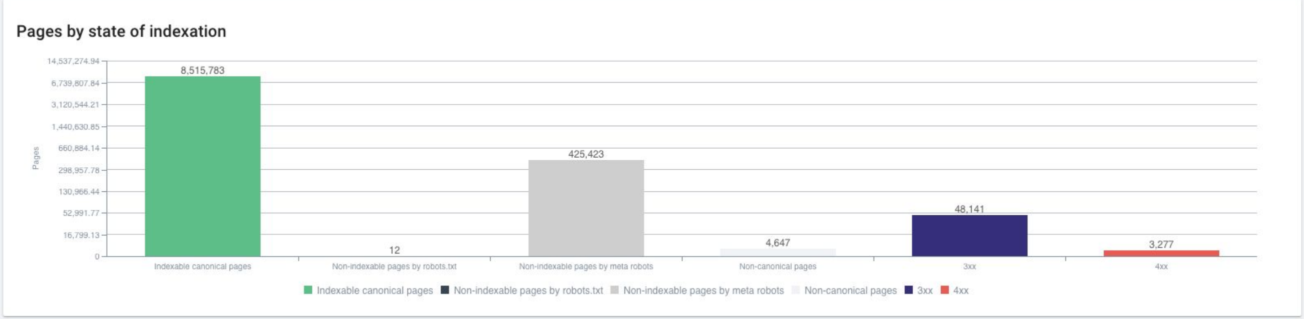
"I'd probably start by looking at **indexability** and the **canonical strategy**."

— Mickaël Serantes, Technical SEO Strategist, Oncrawl

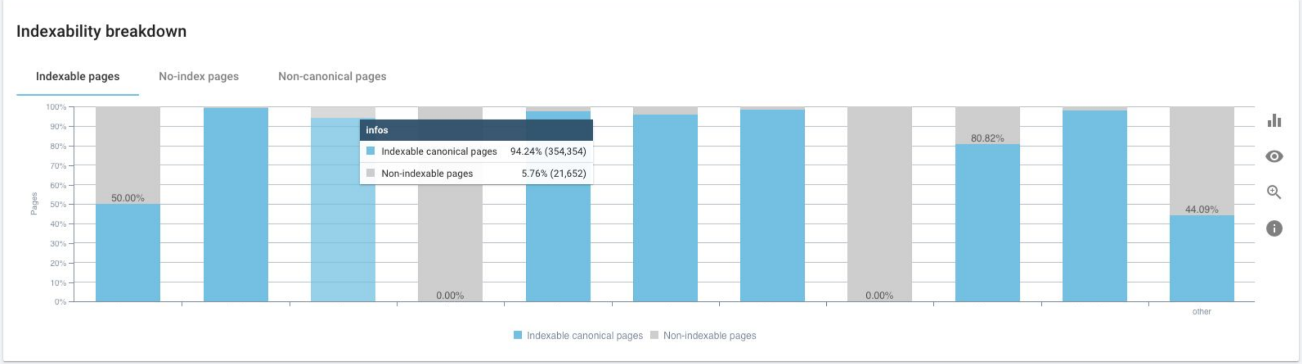
"Particularly with ecommerce sites with parameters, I'd focus on **indexability** and **crawl budget**."

— Adeline Lecellier, SEO Strategist, Oncrawl

Indexing & Strategy



Pages by state of indexation



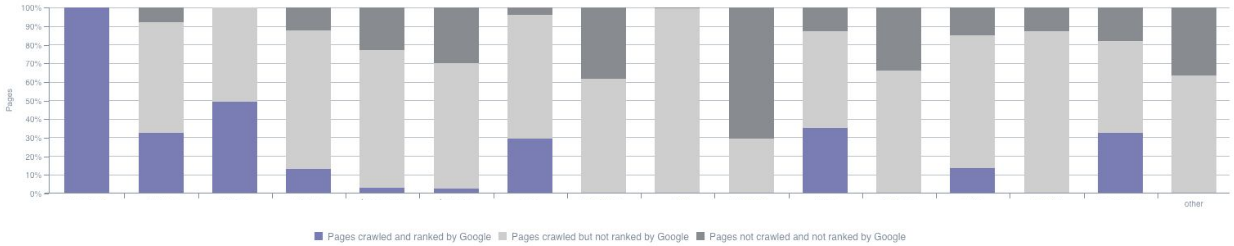
Indexability breakdown

infos

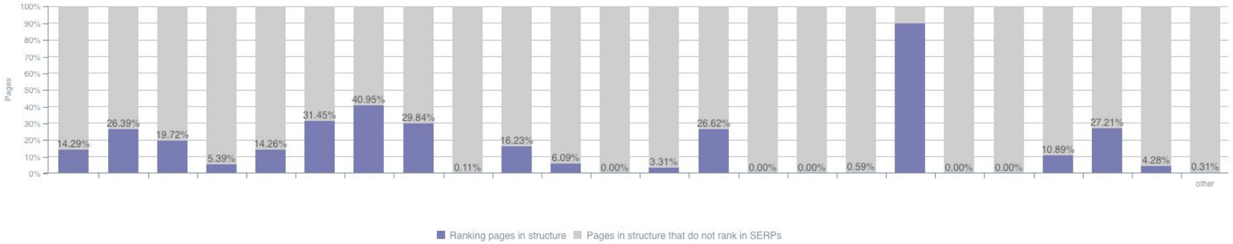
■ Indexable canonical pages	94.24% (354,354)
■ Non-indexable pages	5.76% (21,652)

Indexing & Ranking

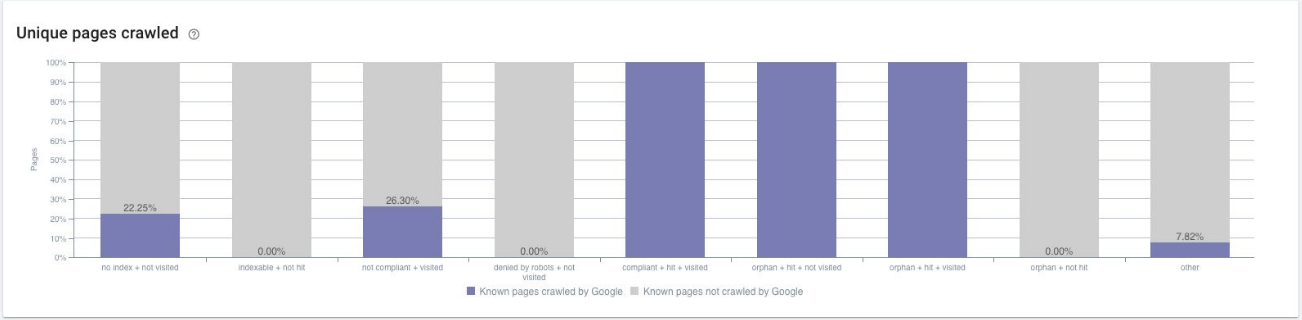
Pages crawled by Google: ratio of ranking pages



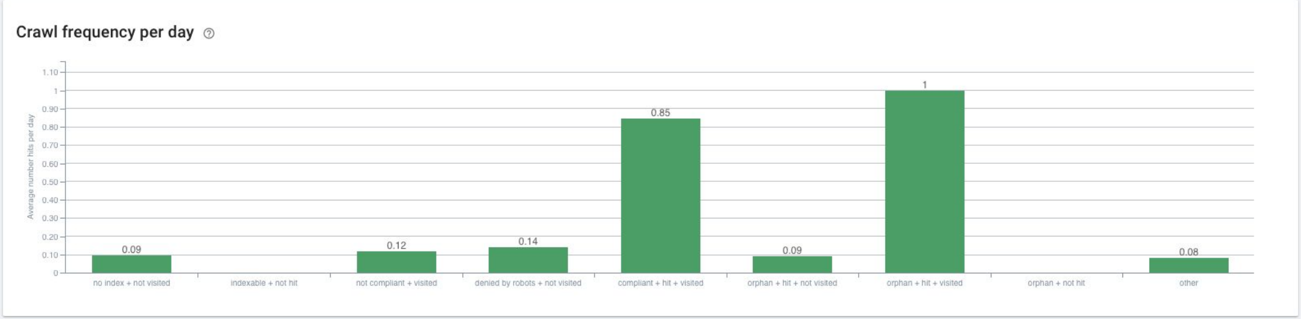
Pages in structure: ratio of ranking pages



Indexing & Crawl budget



Unique pages crawled



Crawl frequency per day

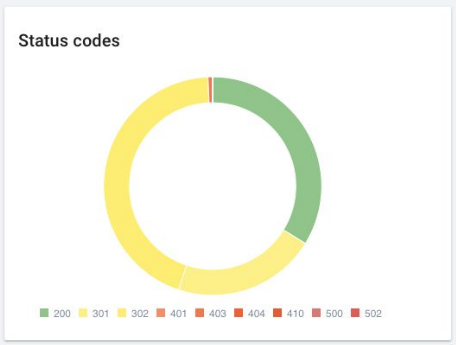
"It's important for big sites to look at **overall health**, particularly **status codes**. Often if they've been using less enterprise-level tools, they have **never seen a full view** of their site health before."

— Janaina Barreto-Romero, Customer Success Manager, Oncrawl

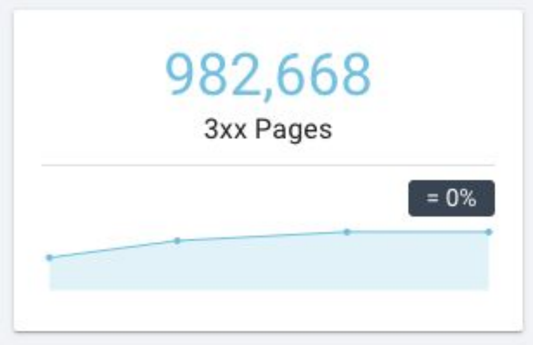
"A good first focus is to check **sanity**. This works well with a sample of pages that is **representative of page templates**."

— Jérôme Salomon, Technical SEO Strategist, Oncrawl

Sanity checks & 3xx redirects

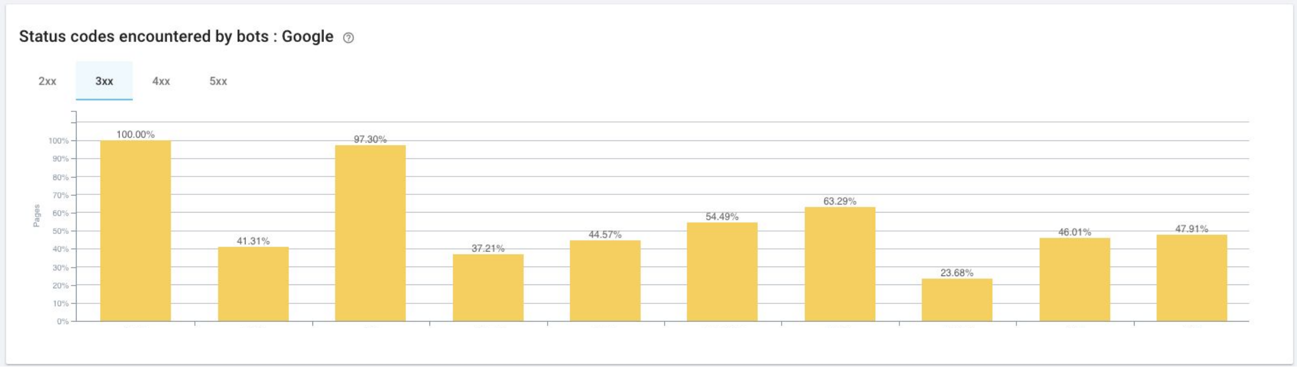


Status codes



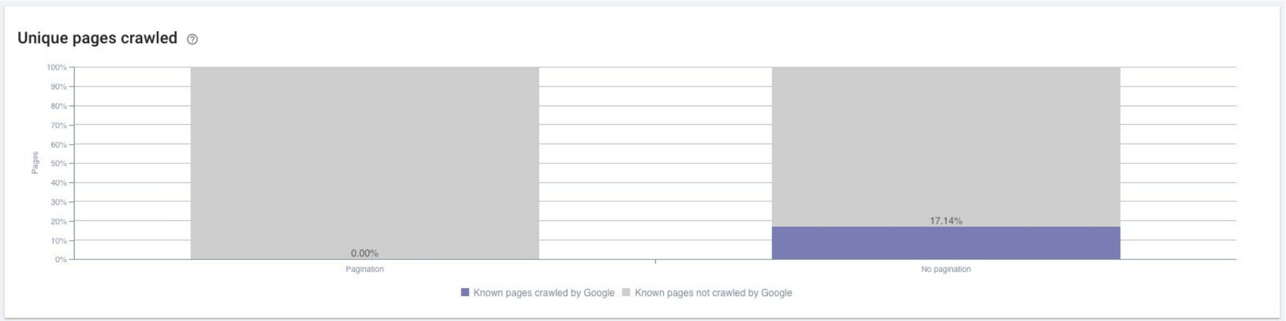
982,668

3xx Pages

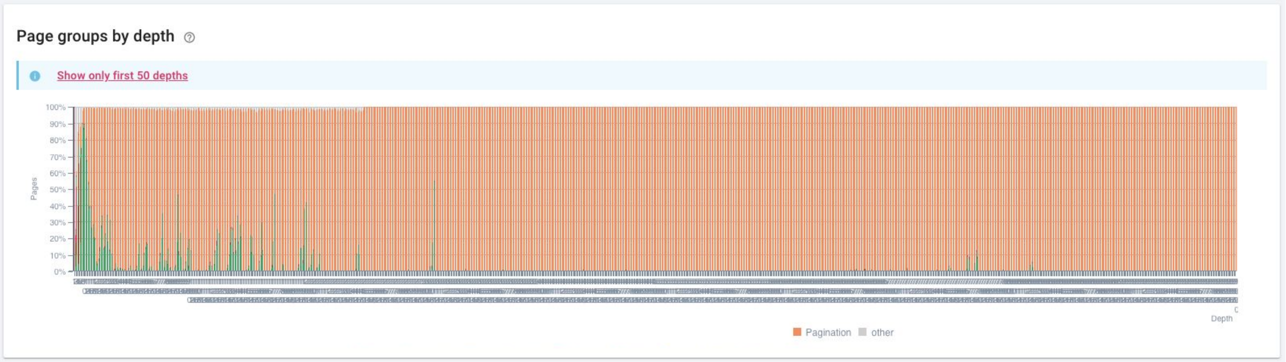


Status codes encountered by bots : Google

Pagination



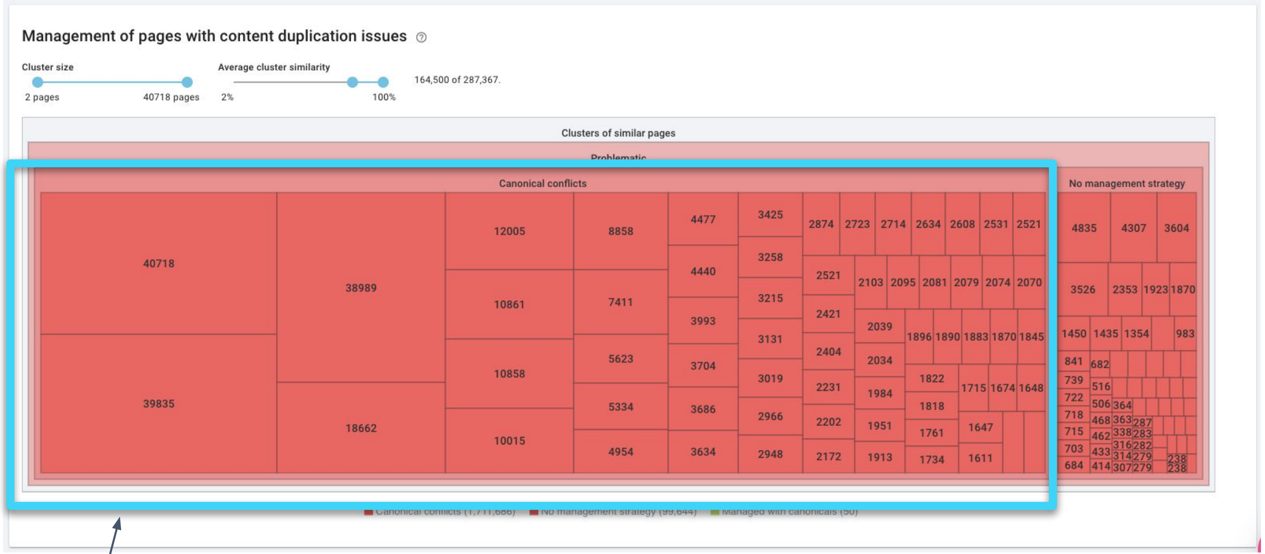
Unique pages crawled



Page groups by depth

Show only first 50 depths

Duplicate content



Management of pages with content duplication issues

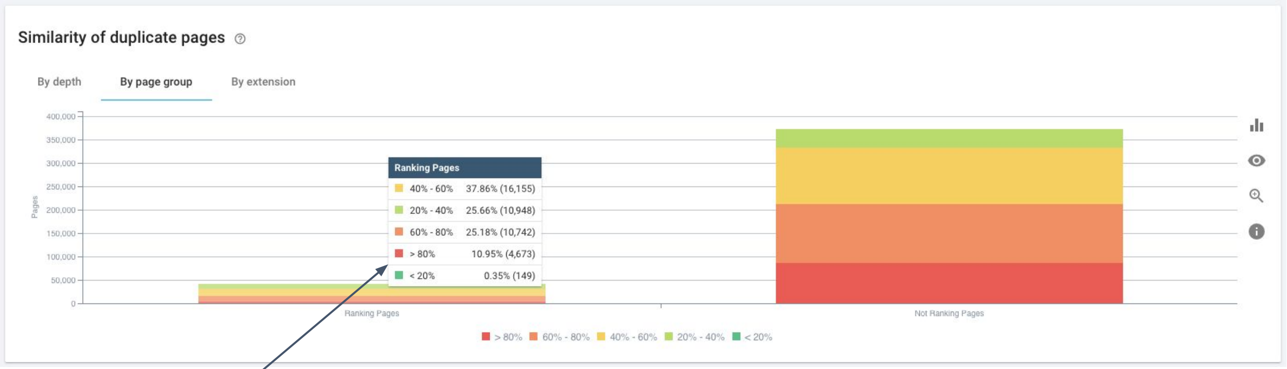
Cluster size: 2 pages – 40718 pages

Average cluster similarity: 2% – 100%

164,500 of 287,367.

pages in these clusters declare a canonical URL...
BUT there's **more than one canonical URL** per cluster

Duplicate content



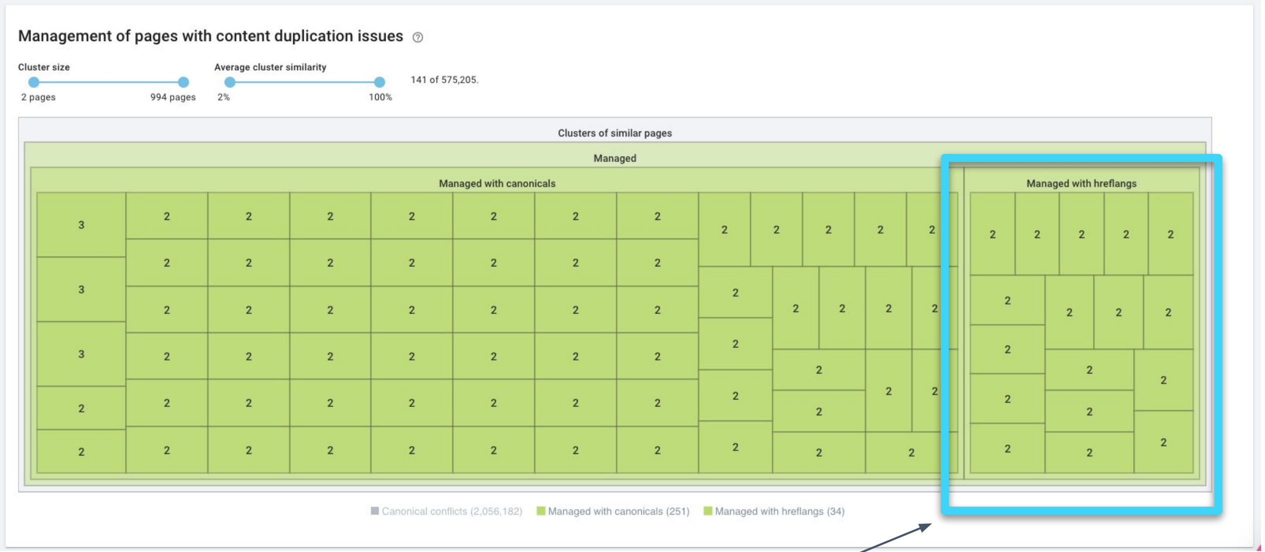
Similarity of duplicate pages

By depth | By page group | By extension

Ranking Pages	
🟨 40% - 60%	37.86% (16,155)
🟩 20% - 40%	25.66% (10,948)
🟧 60% - 80%	25.18% (10,742)
🟥 > 80%	10.95% (4,673)
🟩 < 20%	0.35% (149)

the pages that rank
that are also in a cluster with >80% similarity

Hreflang



Management of pages with content duplication issues

Cluster size: 2 pages – 994 pages

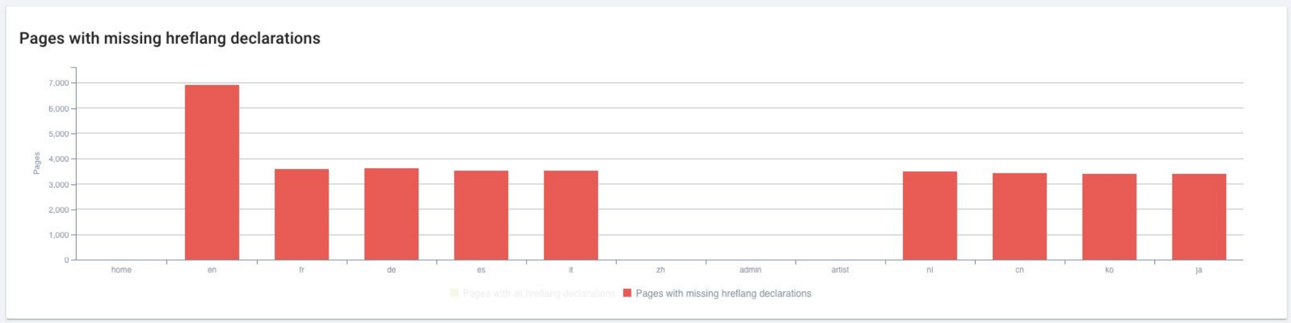
Average cluster similarity: 2% – 100%

141 of 575,205.

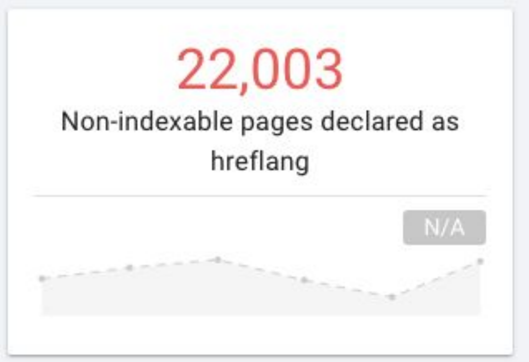
🟩 Canonical conflicts (2,056,182) 🟩 Managed with canonicals (251) 🟩 Managed with hreflangs (34)

pages in these clusters use hreflang declarations to explain the similarities between the pages

Hreflang



Pages with missing hreflang declarations



22,003

Non-indexable pages declared as hreflang

N/A

"It's also important to **ingest additional information**, like exports from Ahrefs or SEMrush."

— Janaina Barreto-Romero, Customer Success Manager, Oncrawl

"What's important? Sanity, indexing, and logs. And you should also be **ingesting conversion data** if you really want to improve SEO on a big site."

— Julien Picot, Product Manager, Oncrawl

Big sites are

different.

- Insist on what you need
- Use the right tools
- Focus
- Segment
- Prioritize
- "Quick wins"

Grab your controller, it’s time for Q&A!