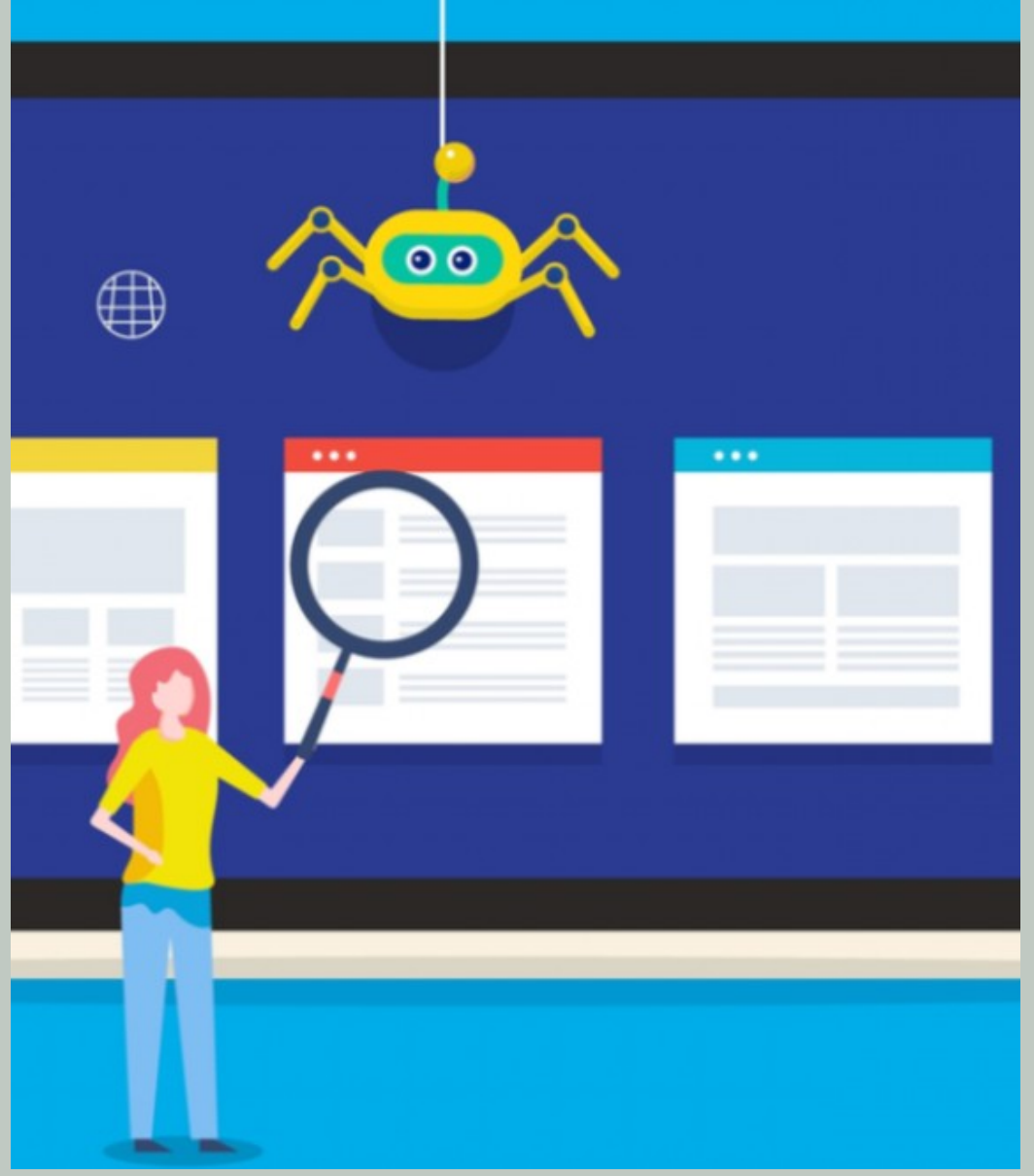
Indeksowanie stron internetowych przez Google



Indeksowanie stron

Gdy roboty indeksujące znajdą stronę internetową, nasze systemy skanują jej treść, podobnie jak przeglądarka. Zwracamy uwagę na kluczowe sygnały – od słów kluczowych po aktualność strony – i rejestrujemy te informacje w indeksie wyszukiwania.

Indeks wyszukiwarki Google zawiera setki miliardów stron internetowych i ma rozmiar ponad 100 000 000 gigabajtów. To jak indeks na końcu książki – ale z wpisami dla wszystkich słów wyświetlanych na każdej indeksowanej stronie. Gdy indeksujemy stronę internetową, przypisujemy ją do indeksu zgodnie ze słowami, które zawiera.

Indeksowanie stron

Proces dodawania stron internetowych do wyszukiwarki.
Podczas indeksacji wszystko zależy od używanego metatagu:
•index
•Noindex

W pierwszym przypadku Robot Google odwiedza stronę, następnie odczytuje kod źródłowy, a później indeksuje.

Metatag noindex oznacza, że strona nie zostanie dodana do indeksu wyszukiwania w sieci. Kiedy korzystamy z wyszukiwania, tak naprawdę przeszukujemy bazę danych Google czyli Indeks.

Indeksowanie stron

Google oferuje Search Console, by właściciele witryn mogli szczegółowo wskazać, jak indeksowane mają być ich strony internetowe.
Właściciele witryn mogą dokładnie określić, jak powinny być przetwarzane strony, poprosić o ponowne zindeksowanie lub całkowicie zrezygnować z indeksowania przy użyciu pliku o nazwie „robots.txt”.

Google nigdy nie pobiera opłat za częstsze indeksowanie witryny – każda strona internetowa ma zapewnione te same narzędzia do indeksowania, w celu zapewnienia użytkownikom wyszukiwarki jak najlepszych wyników.

Web-crawling

Crawler, nazywany także pająkiem, robotem, pełzaczem lub botem to program, który wykorzystują wyszukiwarki. Dzięki niemu badają strukturę, zawartość i kod stron. Później, na tej podstawie wyszukiwarki pokazują użytkownikom strony, które są najbardziej wartościowe w odniesieniu do ich zapytania.

Najbardziej znanym crawlerem jest google crawler, zwany Googlebot.

Web-crawling

Roboty Google sprawdzają wiele czynników na stronie przed indeksacją strony.
Biorą pod uwagę między innymi:

- słowa kluczowe,
- treść,
- poprawny kod,
- element title,
- atrybuty alt.

robots.txt

Istnieje kilka sposobów, aby skłonić Roboty Google do coraz częstszego odwiedzania naszej strony oraz jej zindeksowania. Pierwszym krokiem jest sprawdzenie, czy wcześniej wspomniany plik **robots.txt** umożliwia Robotowi Google prawidłową indeksację strony.

Robots.txt to plik, którego zadaniem jest komunikacja z robotami, które indeksują naszą stronę. Warto zasugerować im, jak powinny to zrobić. Ten plik jest pierwszą rzeczą, jaką sprawdzają boty, które wchodzą na stronę internetową, by dokonać indeksacji w Google.

robots.txt

Składa się on z kombinacji komend zgodnych ze standardem *Robots Exclusion Protocol* – "językiem" zrozumiałym dla botów.

Dzięki temu możemy wpłynąć na kierunek ich ruchu, ograniczając dostęp do zasobów, które w kontekście wyników wyszukiwania są zbędne. Mogą być to pliki graficzne, style, skrypty, a co najważniejsze – określone podstrony naszej witryny.

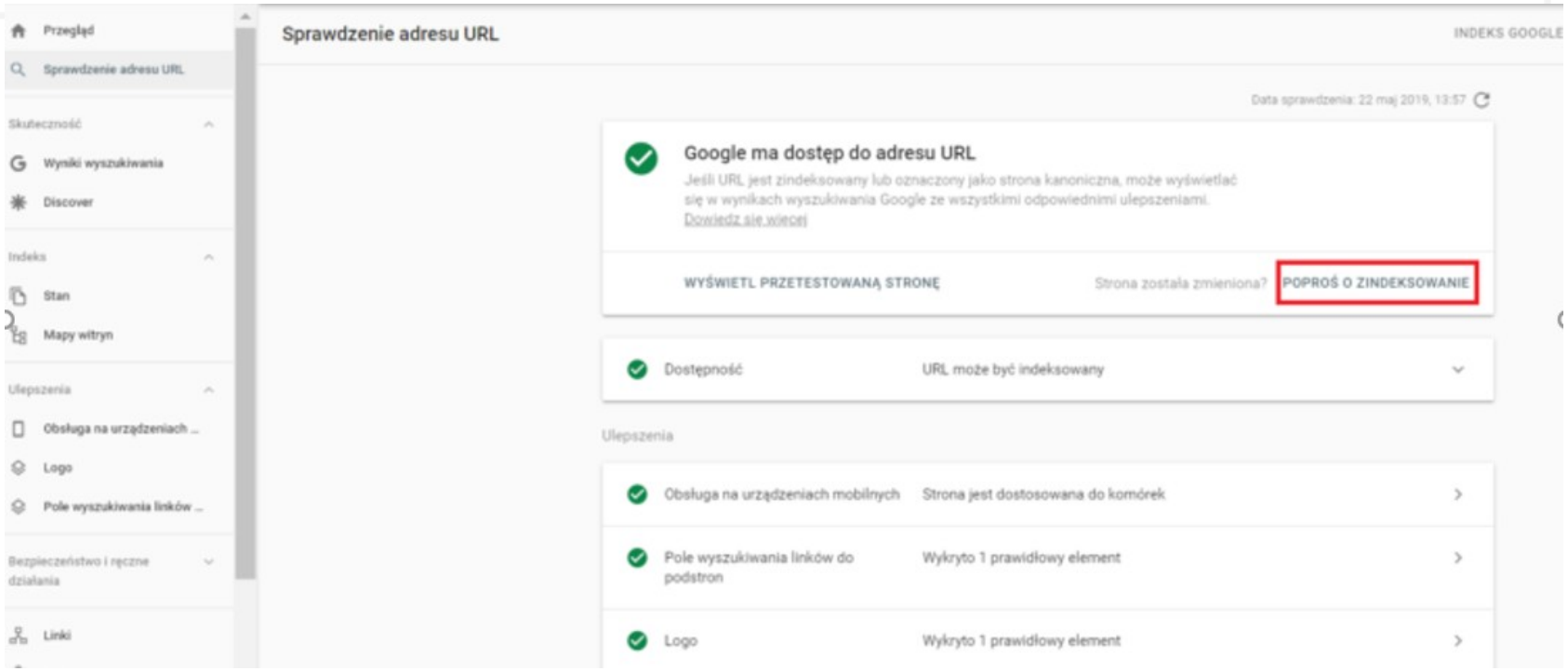
```
User-agent: *
Allow: /environment/cache/images/
Disallow: /application/
Disallow: /environment/
Disallow: /libraries/

Disallow: /*/fav/add/

Disallow: /*/p/comment/add/
Disallow: /*/p/mail/recommend/
Disallow: /*/p/q/
Disallow: /regulamin.html
Disallow: /*/reg/
Disallow: /*/login/
Disallow: /*/basket/
Disallow: /*/searchquery/
Disallow: /*/priceto/
Disallow: /*/pricefrom/
Disallow: /*/index/
Disallow: /*/full/
Disallow: /*/default/
```

Search console

Jest to najszybszy i najprostszy sposób indeksowania strony. Zajmuje od kilku sekund do kilku minut. Po tym czasie Twoja strona staje się widoczna w Google. Wystarczy wkleić pełny adres strony do indeksacji oraz kliknąć → **poproś o zindeksowanie**.



mapy XML

Mapa XML jest przeznaczona dla Robotów Google.
Wszystkie strony powinny ją posiadać, ponieważ **format XML zdecydowanie ułatwia indeksowanie strony w Google**.

Wersja XML jest to **mapa witryny o innej strukturze, która jest przeznaczona dla robotów Google**. Jest to zbiór podstron i informacji o adresach URL. Zawiera między innymi takie szczegóły jak: daty ostatnich aktualizacji podstron, ważność/priorytet URL oraz częstość wykonywanych na nich zmian.

Gdy już uda nam się wygenerować mapę witryny, należy dodać ją do wyszukiwarki Google. Dzięki temu Roboty Google będą wiedzieć gdzie znaleźć daną sitemapę wraz z jej danymi. By przesłać mapę XML do Google należy użyć wcześniej wymienionego **Google Search Console.**

Rozwiązania komercyjne

W przypadku niskiej wiedzy informatycznej lub problemów z samodzielną implementacją rozwiązań do indeksowania można skorzystać z kompleksowych rozwiązań komercyjnych. Istnieje wiele tego typu narzędzi, w większości są one płatne bądź posiadają limitowaną, darmową wersję.

Indeksowanie za pomocą narzędzi online jest istotne dla linków i stron do których nie mamy dostępów. Dzięki dodaniu ich do indeksacji Robot Google będzie mógł swobodnie się po nich poruszać.

Crawl budget

Crawl budget to czynnik, który określa, jak wiele roboty Googla poświęcą, aby zindeksować daną witrynę. Przekłada się to na liczbę podstron dodanych do indeksu wyszukiwarki oraz na częstotliwość indeksowania. Na *współczynnik indeksacji* wpływają m.in:

- *crawl rate limit* – limit liczby podstron odwiedzanych w krótkim czasie przez roboty wyszukiwarek
- *crawl health* – niski czas odpowiedzi serwera, poprawne kody odpowiedzi, szybkość ładowania
- *crawl demand* – zapotrzebowanie na ponowną indeksację, zależne m.in. od częstotliwości aktualizacji treści, popularności witryny
- rozmiar strony, zastosowanie kodu Java Script – strony z JS zużywają więcej budżetu indeksowania

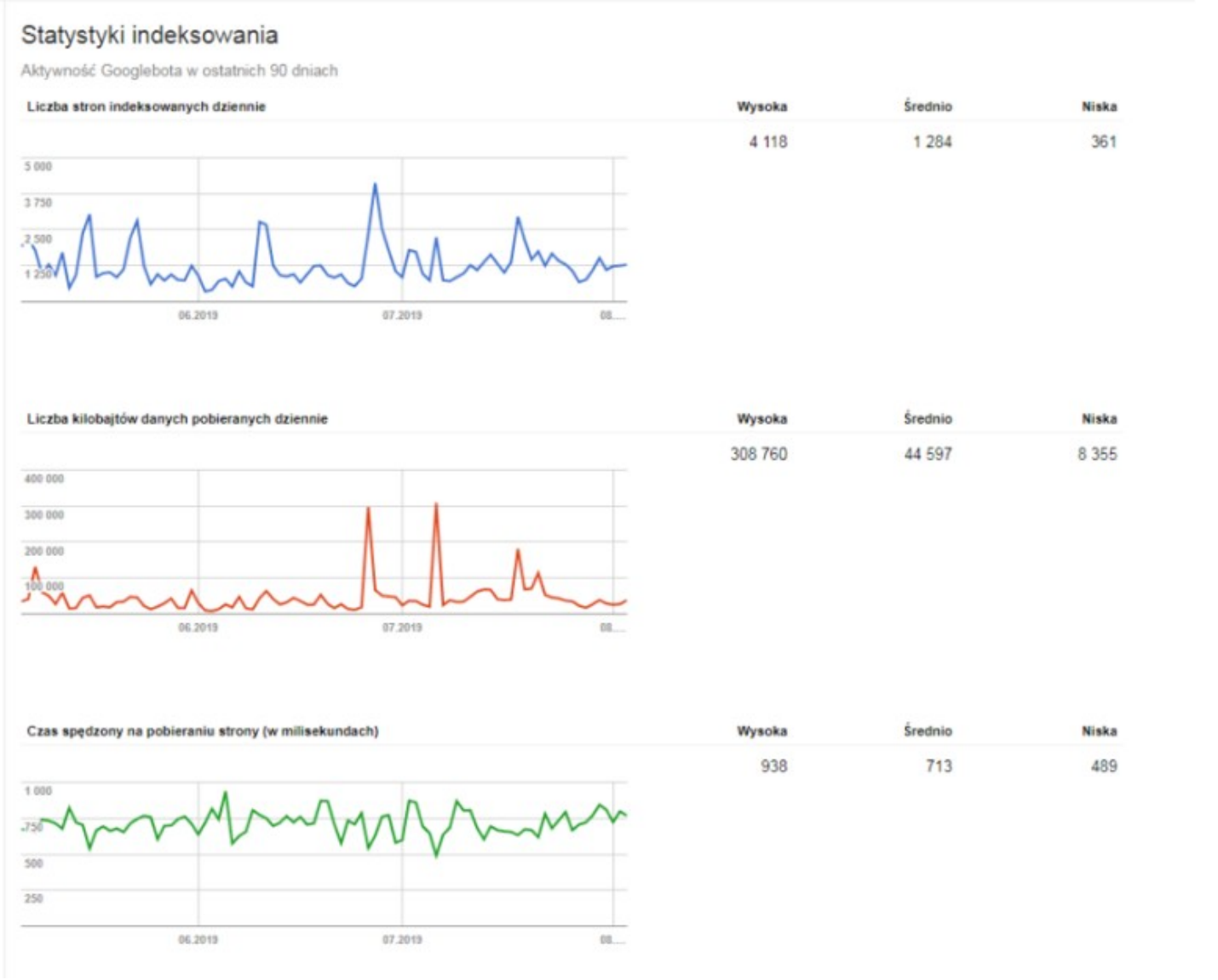
Crawl budget

Największy wpływ na Crawl Budget mają dwa parametry:
- **Crawl Rate Limit**
- **Crawl Demand**

Crawl Rate Limit to limit, który został wprowadzony, żeby Google nie crawlował zbyt dużej ilości stron w danym czasie. Jest to powstrzymanie Google od wysyłania zbyt wielu zapytań, które powodowałyby spowolnienie szybkości strony.
Crawl Demand opiera się na ograniczeniach technicznych. Jeżeli strona jest wartościowa dla potencjalnego użytkownika, Google Robot chętniej będzie ją odwiedzał. Istnieje możliwość, że nawet jeżeli Crawl Rate Limit nie zostanie wykorzystany witryna może nie zostać zindeksowana. Mogą na to wpłynąć dwa czynniki:
- **popularność**, czyli adresy, które posiadają większą liczbę odwiedzin użytkowników są częściej odwiedzane przez roboty Google,
- **aktualność** – algorytmy Google sprawdzają, jak często strona jest aktualizowana.

Crawl budget

Obok możemy zobaczyć dane z 90 dni. W oparciu o nie możemy określić, jak zmienił się nam *crawl budget* w tym czasie.



Dziękujemy za uwagę