Understanding Websites

OSF’s OSSYM Oct 12, virtual

Mark Overmeer/Ronny Lam, Skrodon

Open Infrastructure

- Team Skrodon:
  - Ronny Lam
  - Mark Overmeer
  - Красимир Беров

- Cooperative projects
  1) Crawl Pipeline ➡ Mon
  2) Who Has What
  3) Crawl Planner
  4) Open Console ➡ now

- *Open* Source & Infra

„Skrodon”

- Share real website related data, on full internet scale. Open Infrastructure
  - Share crawled collections
  - Share extracted (meta)data
  - Share computed data
  - Share computation resources

  - Strict EU law & jurisdiction

Collecting

- Combine *all* information about
  - domains
  - websites
  - sub-websites (per supported language, etc)
- Huge!
  - > 1 billion websites
  - thousands of contributors
  - thousands of consumers

Collecting; domains

- From various sources:
  - email black-lists
  - domain ownership
  - jurisdiction of owner
  - age, nr websites, SEO spam-network, ...
- Contributed
- Distributed centrally

Collecting; website

- From various sources:
  - website maintainer
  - take-down notices
  - jurisdiction of content
  - licensing of content
  - used languages, sub-websites, …
  - crawl helpers, like sitemap & robot.txt, keys
  - detected errors, crawl timings, …

Implementation

- We think we know how to make it work
  - Bulk transport interface for automated, batch processes. (Giga-objects scale)
  - ISP / Domain owner / Website maintainer interface.

Google search console

Google Search Console

Een URL in 'http://mark.overmeer.net/' inspecteren

http://mark.overmeer.net/

- Overzicht
- Prestaties
- URL-inspectie

Index

- Dekking
- Sitemaps
- Verwijderingen

Functionaliteit

- Paginafunctionaliteit
- Site-vitaliteit
- Gebruiksgemak op mobiele ap...

Beveiligingsproblemen en handmatige acties

Verouderde tools en rapporten

- Links
- Instellingen

- Feedback sturen
- Over Search Console

Privacy Voorwaarden

Overzicht

Inzicht krijgen in hoe je nieuwe content presteert

SEARCH CONSOLE INSIGHTS

Prestaties

RAPPORT OPENEN >

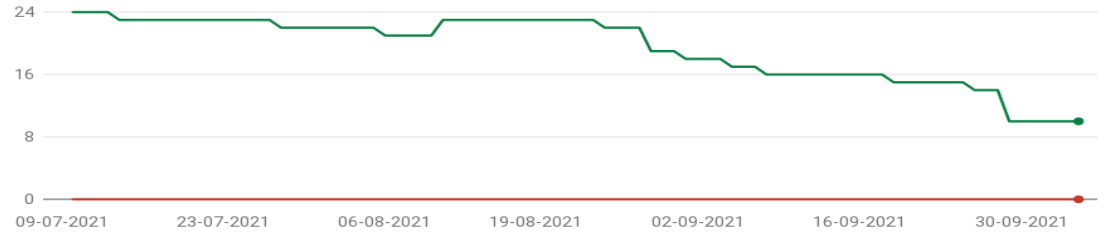
In totaal 8 klikken afkomstig van online zoekopdrachten

Dekking

RAPPORT OPENEN >

0 pagina's met fouten

10 geldige pagina's



Functionaliteit

Information is Power

- Google's strategic advantage over
  - Bing, Yandex, Baidu, … „webmaster" interfaces
  - DuckDuckGo, Infotiger, Plumb.One
  - Pages with indexes, monitoring services,…
- Not everything can be auto-detected (correctly)
- ISPs and Website owners WANT to help: win-win
  - but not in 3+ different consoles

Information is Power

- Much better than sitemaps / robot.txt

- Required by EU law:
  - give license on data
  - right to delete collected information
  - right to correct incorrect information

Presenting: *Open Console*

- Join efforts of website crawling parties to communicate with website owners.

- Shared interface to
  - ISPs (network/hardware maintainers)
  - Domain owner
  - (Sub-)Website maintainer(s)

Presenting: *Open Console*

- Join efforts ~~of website crawling parties~~ to communicate with website owners.

- Shared interface to
  - (email) blacklist reset procedures
  - (external) website classification maintainers, like „porn”, „phishing”, „restaurant”
  - performance monitoring

Organizational

- Very careful design
  - No monopoly on the presentation
  - fair competition
  - legal responsibility
- Association?

Status

Political