

# Open Search with Skrodon

**Mark Overmeer Msc**  
**MARKOV Solutions, NL**  
**May 10, 2022 NLUUG**

# NLUUG NJ95

- DHP = Dutch Home Page
- 1994 -- 1998!

January 1994





# NLUUG NJ95

- DHP = Dutch Home Page
- 1994 -- 1998!

There is a very good reason that “The Yellow Pages” is flat, not a tree.

➔ only tagging works

August 1998

Iedere week de  
nieuwste sites  
per E-mail?  
Klik hier!



[Tekstueel](#)



## Dutch Home Page

[Welkom bij de DHP v2.4](#)

### Thematisch zoeken

Vanaf deze pagina vindt u de sites geordend naar onderwerp.  
[Uitleg.](#)

#### Onderverdeeld in

- [Actueel](#) (85)  
[Events](#) [Nieuws](#) [Weer](#) ..
- [Computers](#) (2295)  
[Hardware](#) [Internet](#) [Software](#) ..
- [Cultuur](#) (1581)  
[Geloof](#) [Geschiedenis](#) [Kunst](#) ..
- [Geen categorie](#) (1)
- [Gezondheid](#) (517)  
[Milieu](#) [Patienten](#) [Zorg](#) ..
- [Media en Naslag](#) (661)  
[Bibliotheken](#) [Radio](#) [Tijdschriften](#) ..
- [Onderwijs](#) (808)  
[Middelbare scholen](#) [Primair onderwijs](#)  
[Studentenorganisaties](#) ..
- [Overheid en Regionaal](#) (596)  
[Gemeenten](#) [Politiek](#) [Steden](#) ..
- [Transport](#) (525)  
[Auto's](#) [Boten](#) [Motoren](#) ..
- [Vrije tijd](#) (2086)  
[Sport](#) [Uitgaan](#) [Vakantie](#) ..
- [Wetenschap](#) (246)  
[Astronomie](#) [Biologie](#) [Medisch](#) ..
- [Winkelen](#) (1518)  
[Erotiek](#) [Horeca](#) [Makelaars](#) ..
- [Zakelijk](#) (2307)  
[Banen](#) [Geld](#) [Industrie](#) ..

Bedrijven zijn opgesplitst naar hun activiteit: in [Winkelen](#) gaat het om bedrijven die openstaan voor publiek, die in [Zakelijk](#) leveren producten en diensten voornamelijk aan andere bedrijven.



[Heini Withagen](#) ([heiniw@dhp.nl](mailto:heiniw@dhp.nl)) en [Mark Overmeer](#) ([markov@ATComputing.nl](mailto:markov@ATComputing.nl))  
18 augustus 1998, 12805 sites (183 nieuw) in 223 categorieën; [disclaimer](#).

# Indices

- DHP taught me:
  - too many sites to maintain manually.
  - localized depicting websites is unnatural.
  - hierarchical indexes do not work.

➔ Text Search Engines?

The screenshot shows the AltaVista search engine interface. At the top, the logo reads "alta vista: SEARCH" with the tagline "smart is beautiful" on the right. A navigation bar includes links for "Search", "Live", "Shopping", "Local", "Free Access", and "Email". Below this is a "Go Live" button and a promotional message: "Own your favorite music at CDNow!". The main search area features a search bar with the text "Find this:", a "Search" button, and a language dropdown menu set to "any language". A tip below the search bar says: "Tip: Search for pages in a foreign language only." Below the search bar are radio buttons for "Find Results on:" with options for "The Web" (selected), "News", "Discussion Groups", and "Products". The main content area is divided into several columns of links and featured content. On the left, there's a section for "AltaVista Live!" with sub-sections like "Get Inside" (Money, News, Sports, Translation, Travel, Careers, Health, Entertainment, Local, Portfolio, Email, Chat, Alert), "Today's Features" (Build your homepage on AltaVista, Live Web Events - by Yack.com, Shopping: Give a Gift, Get up to \$50!), "Top Stories" (World Prepares for Millennium Bashes, Vanuatu Quake Kills Eight, Injures 100), and "Technology News" (Y2K Bug Shows Up in Philadelphia, Fiber Optic Streets Cause Problems). Below this is a "Search for..." section with categories like Farming, Denise Richards, and Books. The middle column lists various categories such as "Arts & Entertainment", "Autos", "Business & Finance", "Computers", "Games", "Health & Fitness", "Home & Family", "Internet", and "News & Media". The right column lists "Recreation & Travel", "Reference", "Regional", "Science", "Shopping", "Society & Culture", and "Sports". At the bottom right, there's a section for "AltaVista Shopping.com" with features like "Find the Lowest Prices" (listing Palm V, Nikon Coolpix 950, Canon Mini DV Camcorder), "Compare Features" (DVD Players, Camcorders, Cordless Phones, Televisions), and "Read Reviews" (Desktop Computers, Notebook Computers, Printers, Digital Cameras). The footer contains various links like "About AltaVista", "Help", "Contact Us", "Advertise With Us", "Business Solutions", "Job Openings", "Press Room", "Privacy", "Terms of Use", "A CMGI Company", "Shopping", "Money", "News", "Sports", "Travel", "Careers", "Health", "Entertainment", and a copyright notice: "©1999 AltaVista Company. AltaVista® is a registered trademark and Smart is Beautiful and the AltaVista logo are trademarks of AltaVista".

# I saw a problem

- (By example) English about has 300k words.
- Native speakers *know*\* 10k to 40k of them.
- Non-natives max 3k.

\*) sources report different ranges.

# I saw a problem

- (By example English) has 300k words.
- Native speakers know 10k to 40k of them.
- Non-natives max 3k.
- Estimated #web-pages:
  - 1998: 2.4 **m**illion pages,  $2.4e+6$
  - 2020\*: 5.5 **b**illion pages,  $5.5e+9$
- Estimate number of words per page: 400-800

\*) source [worldwidewebsize.com](http://worldwidewebsize.com)

# I saw a problem

- (By example English) has 300k words.
- Native speakers know 10k to 40k of them.
- Non-natives max 3k.
- Estimated #web-pages:
  - 1998: 2.4 million pages,  $2.4e+6$
  - 2020\*: 5.5 billion pages,  $5.5e+9$
- Estimate number of words per page: 400-800

Google

search engine



All

Images

Videos

News

Maps

More

Settings

Tools

About 1.660.000.000 results (0,60 seconds)

# other options?



## TimeWalker (4)



tcp-dump  
45000 events

src ip-address

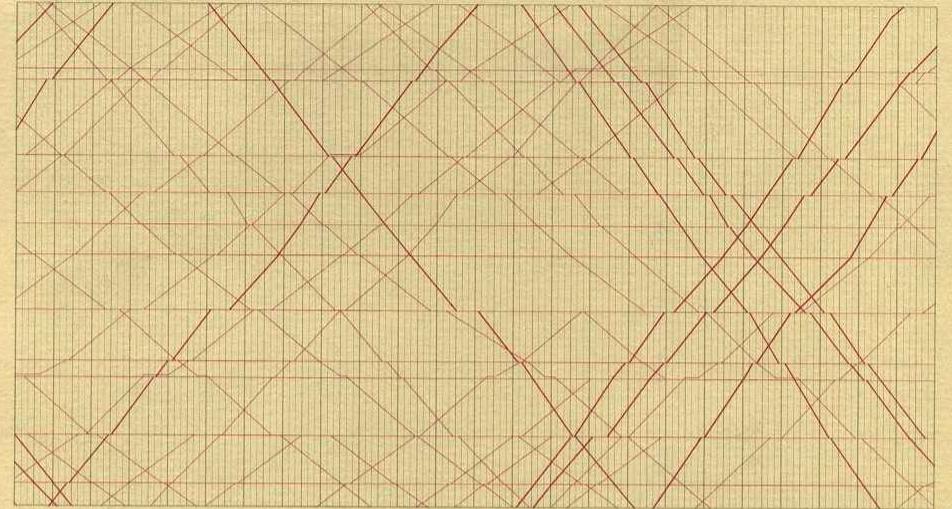
srcPort \* out

anau \* out

qtype \* dur

collapsed levels

8 secs = 160 \* 50 msecs

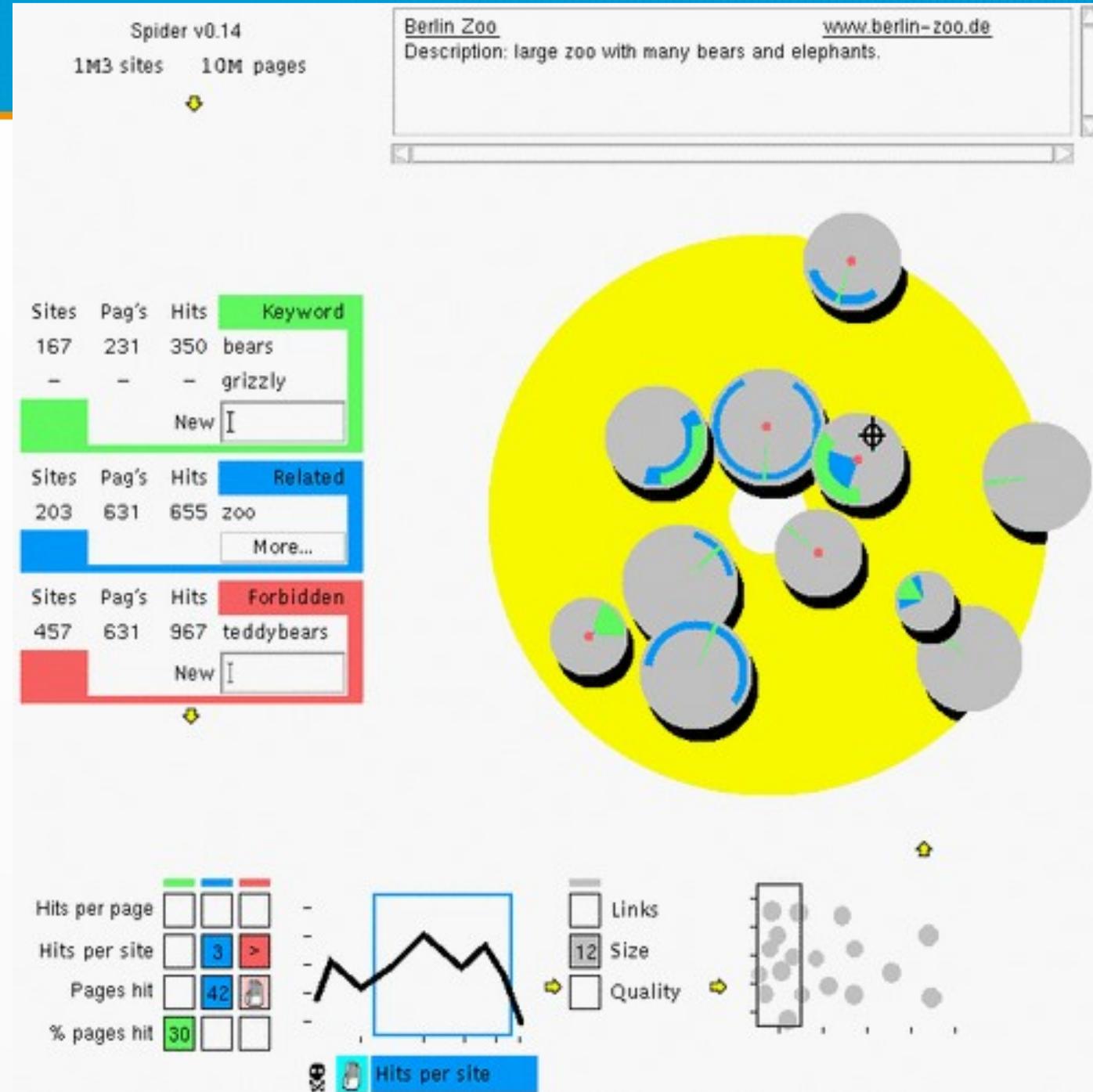


## The Visual Display of Quantitative Information

EDWARD R. TUFTE

# ... do better?

- We do not have enough words.
- Presented at Terena conference, Lund **1999**  
(Two talks and two papers published in IEEE31 “Transactions on Circuits and Systems”)

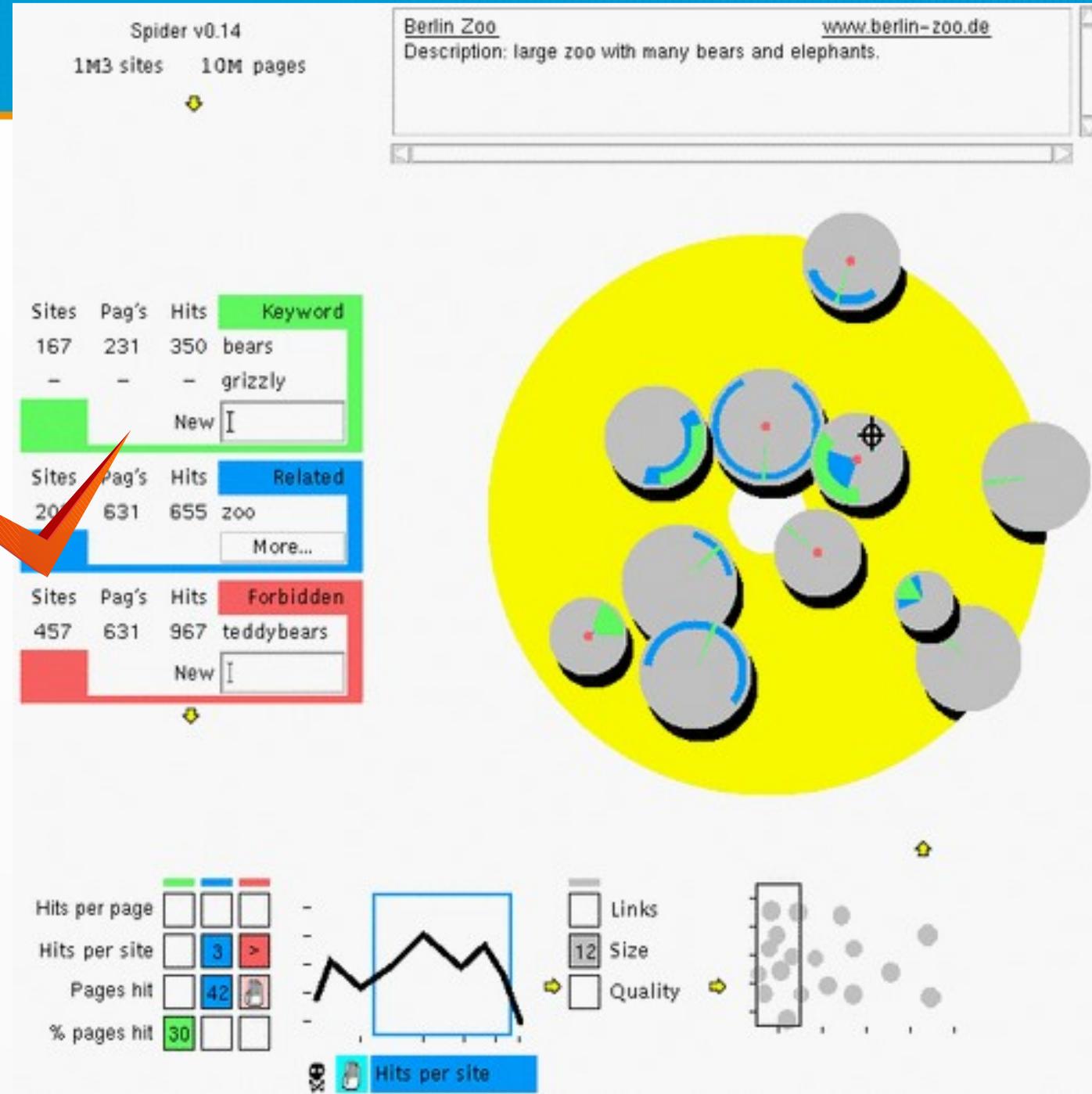


written in Java1 on “Linux” 0.95, 14k4 modem  
on PC 486DX2/66, 16MB RAM, 80MB disk (3000€)

# ... do better?

- We do not have enough words.
- Presented at Terena conference, Lund **1999**  
(Two talks and two papers published IEEE31 “Transactions on Circuits and Systems”)

No data to explore  
my ideas!

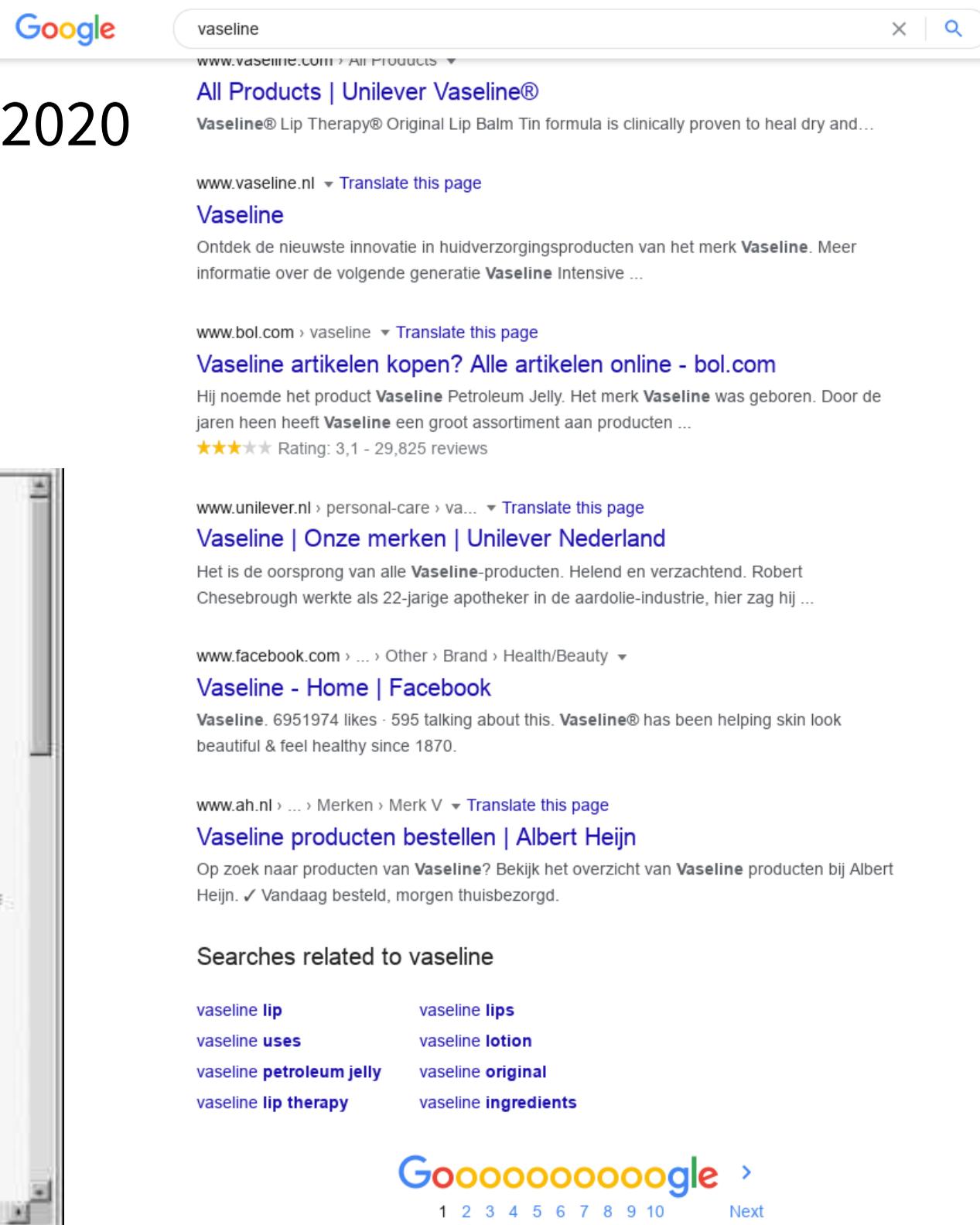


# Use case 1

2020

# Evolved?

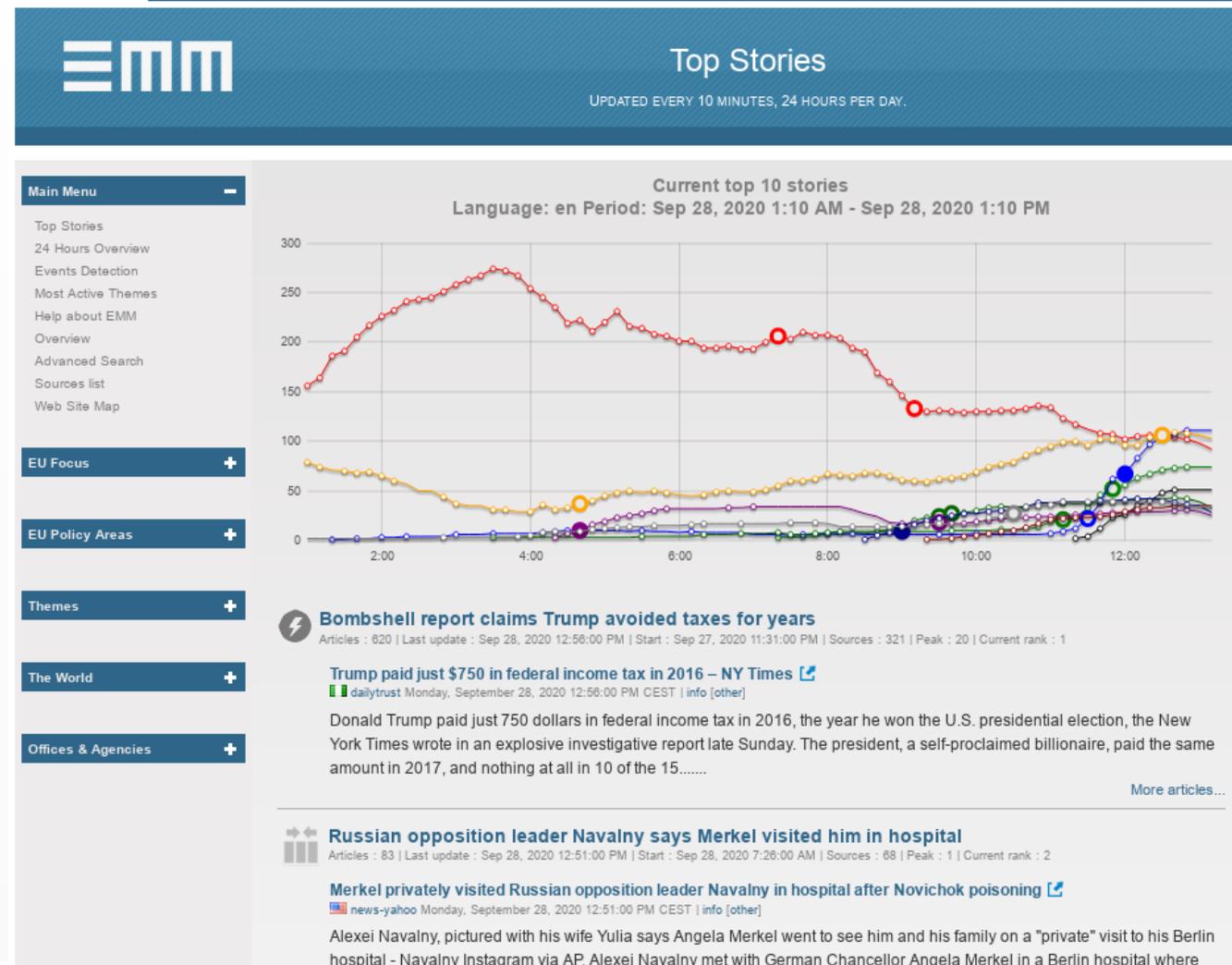
1996



# Many specialized Search Engines

- EMM: The European **Media Monitor** (EU JRC)

- collect articles from >20k media.
- often (daily?)
- not stored: collects trends about content
- contract based access.



# Example 2

- NCSC-NL Taranis
  - collects **articles** from XX sources, via webpages, RSS-feeds, email
  - more than once per hour
  - subject grouping
  - some contracts (like Twitter)
  - long-term storage

The screenshot displays the Taranis web interface. At the top, there is a navigation bar with links for 'Taranis Configuration', 'Statistics', 'User settings', 'About', and 'logout'. The main header features the Taranis logo (Vulnerability Management) and the logo of the Nationaal Cyber Security Centrum (Ministerie van Justitie en Veiligheid). Below the header is a red navigation bar with tabs for 'Assess', 'Analyze', 'Write', 'Publish', 'Dossier', 'Report', and 'Tools'. A search bar is located on the right side of this bar. Below the navigation bar, there are statistics for 'News' (1293) and 'Security vuln' (1492). The main content area includes a search form with fields for 'Search', 'From date', and 'To date'. There are also dropdown menus for 'Category' and 'Source', and a 'Sorting order' dropdown set to 'Date/Time newest (default)'. A 'Hits/pg' dropdown is set to '100'. Below the search form are buttons for 'U', 'R', 'I', and 'W', each with a checkbox. A 'Search!' button is also present. Below the search form, there are buttons for 'C', 'U', 'Markas read', and 'Mark as important'. The results section shows '100 of 2785 results' and buttons for 'Mail items', 'Bulk analysis', and 'Multiple analyses'. The results are displayed in a table with columns for 'Timestamp', 'Source', and 'Title / description'. Each row includes a checkbox, a timestamp, a source logo (APNIC), a star rating, a title, a snippet of the article, and a set of icons for actions like copy, share, and favorite.

Timestamp	Source	Title / description
18-08-2020 13:00:51	APNIC	<b>20 burn-out symptomen die aangeven dat je op de rem moet trappen - Metronieuws.nl</b> 20 burn-out symptomen die aangeven dat je op de rem moet trappen Metronieuws.nl Corona als katalysator bij burn-out: Hoe herken je de signalen op tijd? RTV Noord 'Miljoenen werkenden lopen risico op burn-out' Metronieuws.nl Door de coronacrisis zijn er veel meer 'verborgen burn-outs' in Nederland Leeuwarder Courant Hele verhaal bekijken via Google Nieuws
18-08-2020 13:00:51	APNIC	<b>'Zandvoort lijkt nieuwe coronabrandhaard'   Binnenland - Telegraaf.nl</b> 'Zandvoort lijkt nieuwe coronabrandhaard'   Binnenland Telegraaf.nl Corona breidt zich uit in Zandvoort, meeste nieuwe besmettingen in Amsterdam wnl.tv Meeste nieuwe besmettingen vastgesteld in Amsterdam, Zandvoort een van de snelste stijgers Trouw Zandvoort nieuwe brandhaard voor corona: 52 procent meer besmettingen Hartvannederland.nl Corona breidt zich uit in Zandvoort Zeelandnet Nieuws Hele verhaal bekijken via ...
18-08-2020 13:00:51	APNIC	<b>'Spookvoetballer' Bernio Verhagen geeft fraude én vervalsing toe in Deense rechtbank - AD.nl</b> 'Spookvoetballer' Bernio Verhagen geeft fraude én vervalsing toe in Deense rechtbank AD.nl Avontuur Nederlandse 'spookvoetballer' eindigt in Deense cel NOS Spookvoetballer Verhagen erkent schuld aan fraude Voetbal International Nederlandse spookspeler bekend schuld voor Deense rechtbank FCUpdate mobiel 'Spookvoetballer bekend fraude voor rechtbank' Crimesite Hele verhaal bekijken via Google Nieuws
18-08-2020 13:00:51	APNIC	<b>Bob Jungels voor twee jaar naar AG2R Citroën - Wielerflits</b> Bob Jungels voor twee jaar naar AG2R Citroën Wielerflits Bob Jungels verlaat Deceuninck-Quick-Step en tekent bij AG2R NU.nl Luxemburgse kampioen Jungels verruigt Deceuninck - Quick-Step voor AG2R AD.nl Lilian Calmejane over transfer naar AG2R: "Voelt als een nieuwe carrière" Wielerflits Sport kort: Jungels weg bij Deceuninck - Quick-Step Telegraaf.nl Hele verhaal bekijken via Google Nieuws

# Example 3

- KB, National Library of the Netherlands
- “collect **digital history** of NL”
- contracts per 13k websites, mainly political, cultural and media
- collect by manual screenshots, twice a year
- only access with physical presence in the building (GDPR)
- Use-case: Internet Achive

The logo for KB nationale bibliotheek is located in the top right corner. It consists of a light blue rectangular box with a white border. Inside the box, the letters 'KB' are positioned to the left of a right-facing curly bracket. To the right of the bracket, the words 'nationale' and 'bibliotheek' are stacked vertically in a sans-serif font. The right side of the box is a darker blue shape that tapers to a point on the right.

KB } nationale  
bibliotheek

# Example 4

- KB, National Library of the Netherlands
- Research to the development of sites written in the Frisian language.
  - where are the pages in Frisian?
  - expensive to implement an own crawler (based on existing software)
- Use-case: get **language corpus** for dictionary maintenance or language research.

The logo for the National Library of the Netherlands (KB) is displayed in a light blue rectangular box with a gold-colored arrow pointing to the right on its right side. The text 'KB' is on the left, followed by a right-facing curly bracket, and then the words 'nationale' and 'bibliotheek' stacked vertically on the right.

KB } nationale  
bibliotheek

# Examples 5, 6, 7, 8, ...

- A **copyright holder** want to check images for violations
- The **politician** want to see where his **name** appears
- A company wants to see where its **brand promotion** appears
- A company wants to see when **competitors** change prices
- Police wants to survey **criminal** content
- 
-

# Offering information

- Fast libraries **want to be found**
  - National Libraries
  - CERN
  - Facebook, Twitter, YouTube
  - Webshops
- Parties want to **influence** the information retrieval
  - Webmasters
  - ISPs
  - Governments
  - Law enforcement
  - Fake news patrol
  - End-users
  - ...

# Merging Search Engine Needs

- People have **different** opinions about what “searching the web” is.



# Merging Search Engine Needs

- People have **different** opinions about what “searching the web” is.
- Many, many **overlapping** components between the use-cases.

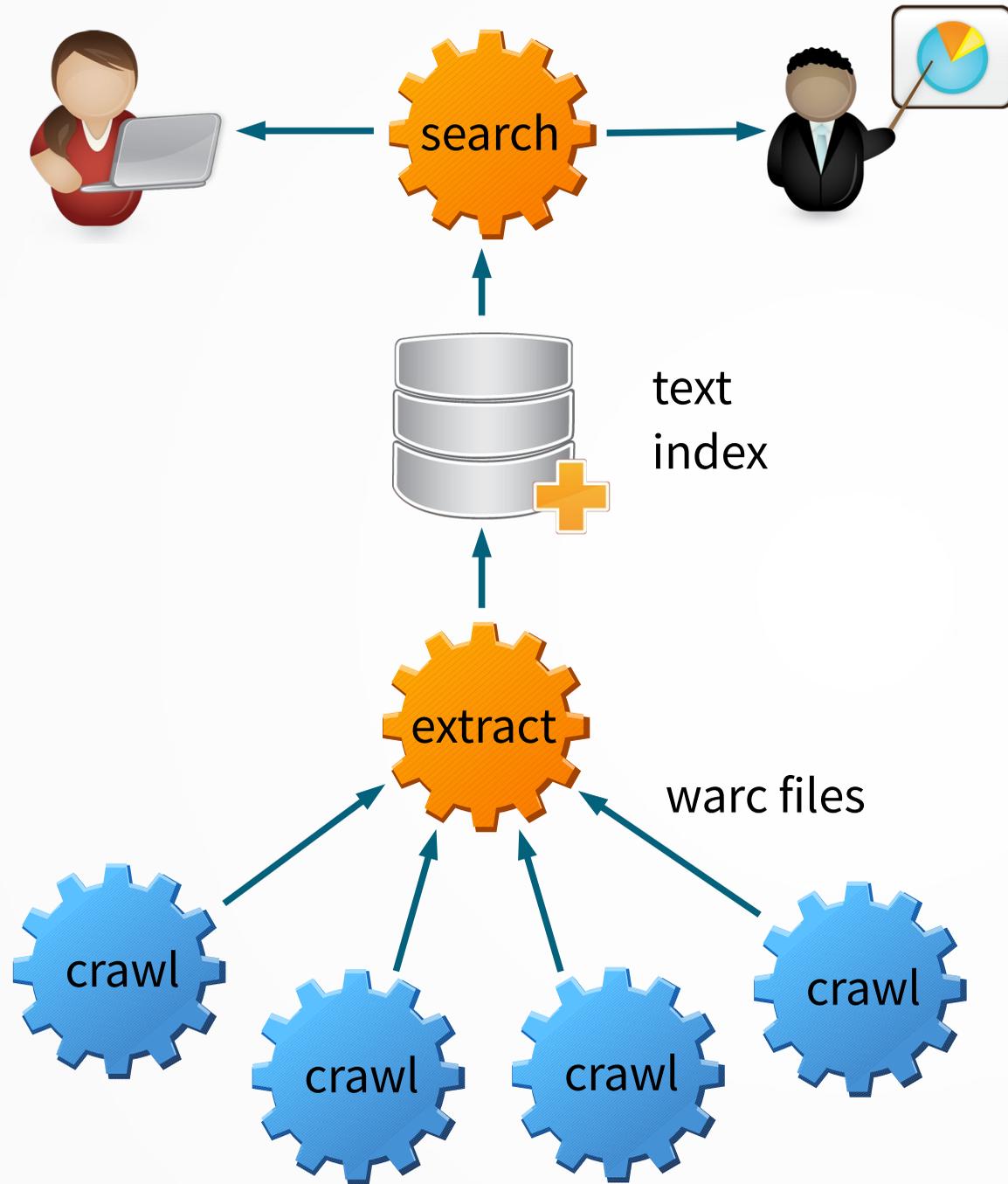


# Merging Search Engine Needs

- People have **different** opinions about what “searching the web” is.
- Many, many **overlapping** components between the use-cases.
- Collection of web-based information at the moment is **very primitive, ad-hoc** and **expensive**.

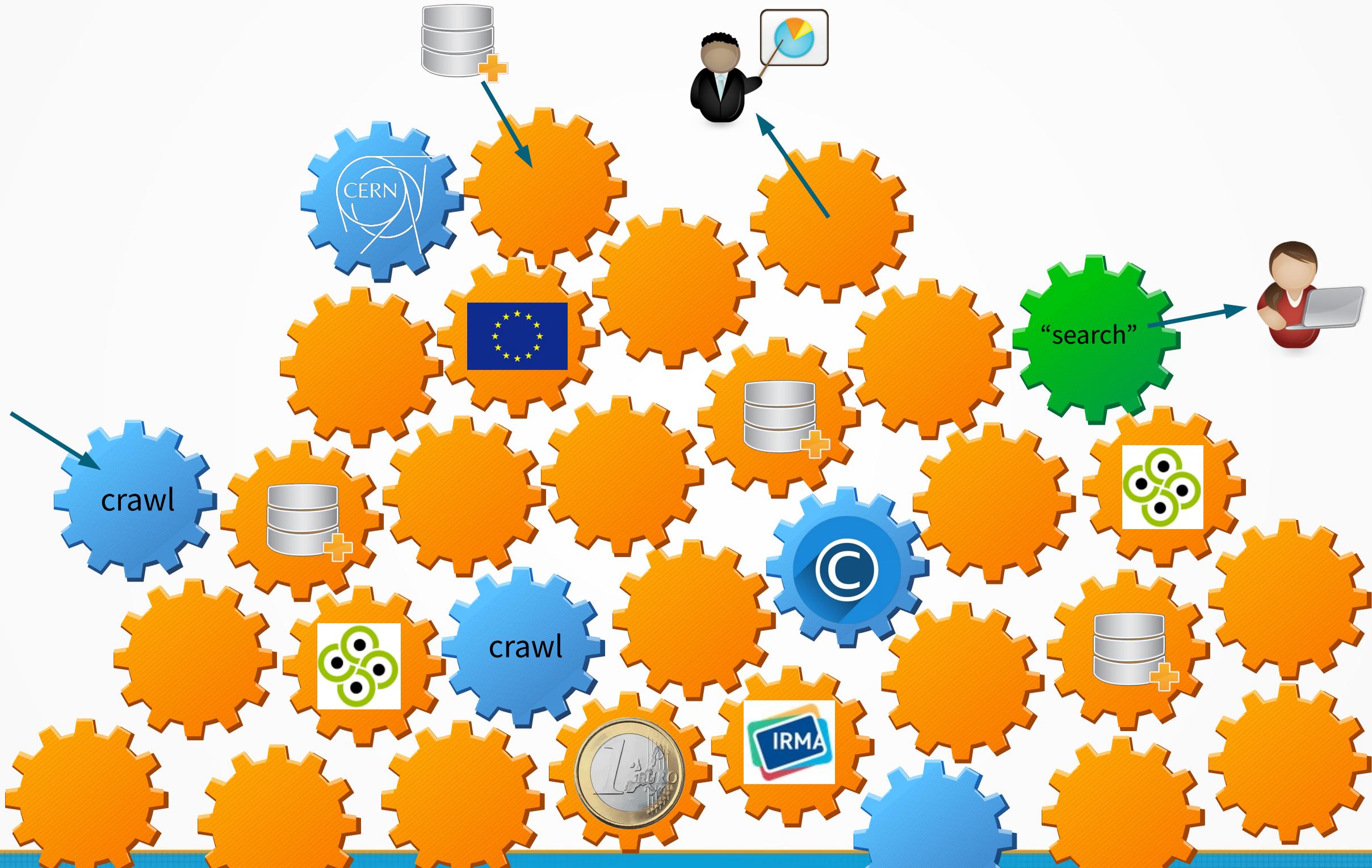
→ Let's evolve!





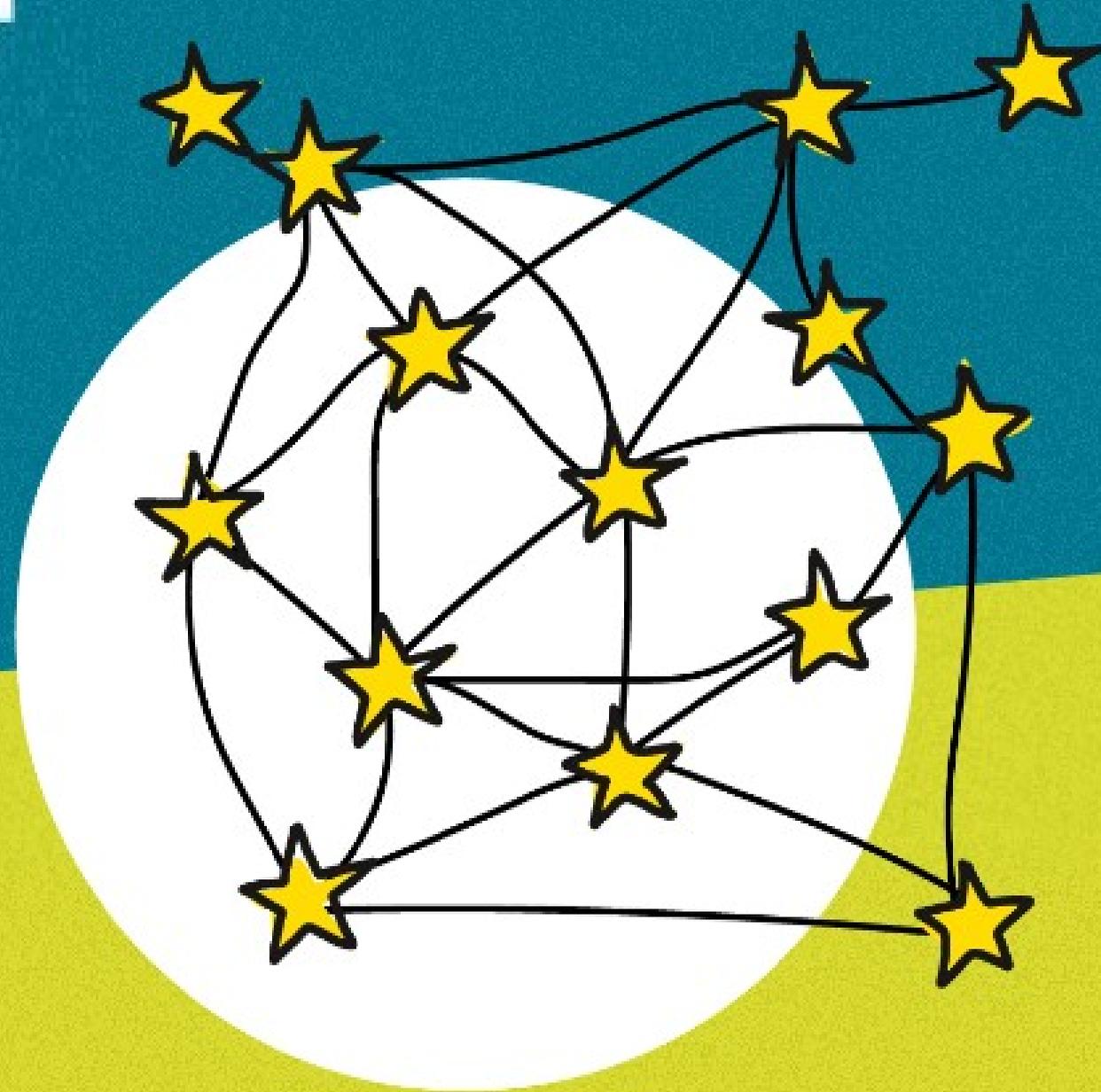
This is **not** how a “Google scale” can be made to work.

For most use-cases, the (text) index will **lack crucial data**.





open search  
foundation



# Generic Needs

- Crawl data
- Meta-data about websites
- Understanding languages, geography, synonyms, etc.
- Compute power
- Ranking / User Interfaces / Usage
- Good behavior

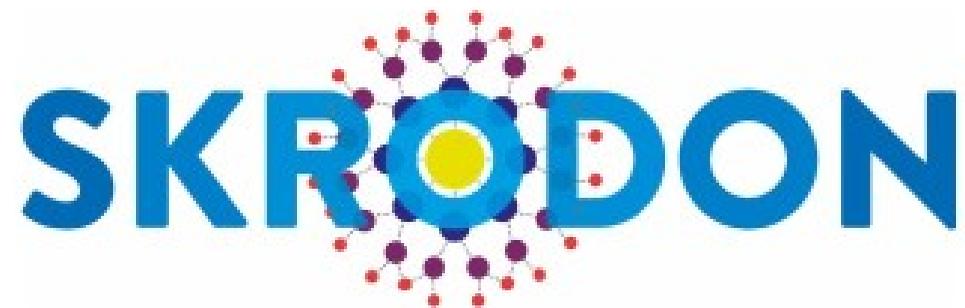


# Good Behavior

- Privacy
- Open Participation
- Open interfaces
- Fair Behavior
- Shared Resources
- EU law



?



-- established 2022 --

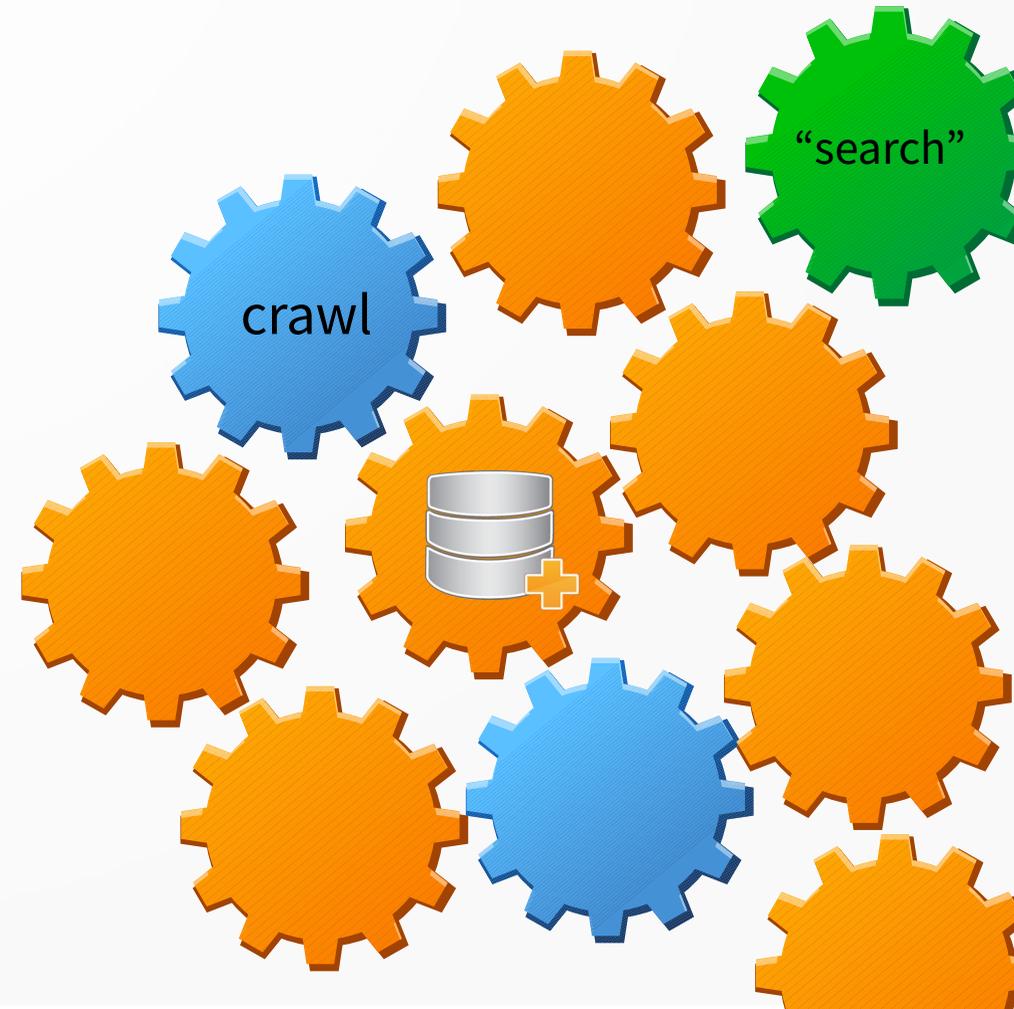
!

# Good Behavior

**SKRODON**

-- established 2022 --

=  $\cap$



# First components in Skrodon



- **Crawl Pipeline**
- Crawl Planner
- Crawl results locator
- **Open Console**

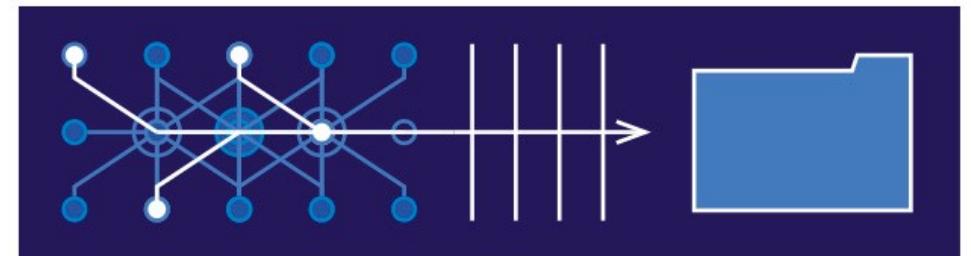


- Meshy Space

# Crawl Pipeline



- Read static (WARC) crawled data: reused!
- Filter
  - select
  - extract
  - repackage
- 10TB input per day
- CommonCrawl monthly collection:
  - 3G pages
  - 320TB req/resp HTTP data
  - 64000 files of 1GB.gz on AWS



PIPELINE  
SHARED - SEARCH . EU

# Crawl Pipeline



- Selectors:

- response codes
- content type
- language, text size
- domain
- full word match
- pattern match text

- Extractors:

- http-headers
- plain text content
- from html
  - normalized urls
  - <link>
  - <meta>
  - href's
  - OpenGraph

REUSE of effort!

## 1. Search Engine **Crawlers** need knowledge about websites

- urls of pages (sitemaps)
- access rights (robots.txt)
- page update frequency
- website popularity
- abuse: spam, phishing, SEO-networks
- provided languages
- geographical location
- network location, ISP

# Open Console



1. Search Engine Crawlers need knowledge about websites
2. Website **owners** want to help automated users
  - publish page update triggers
  - authorization to get page summaries which usually require payment
  - multi-language site/page descriptions
  - optimal crawl moments
  - contact information, location
  - legal information, like jurisdiction, license, and owner

# Open Console



1. Search Engine Crawlers need knowledge about websites
2. Website owners want to help automated users
3. Automated **processes** want to inform websites
  - Access errors, performance
  - Coverage, frequency
  - Additional services on offer

# Open Console



1. Search Engine Crawlers need knowledge about websites
2. Website owners want to help automated users
3. Automated processes want to inform websites
4. Automated **services** are required to implement correction processes
  - repair incorrect information
  - take-down notices

# Open Console



[...]

**5. ISPs** want to help hosted websites

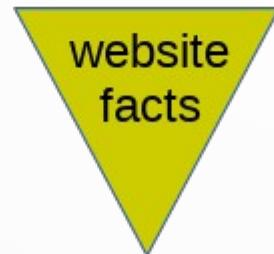
**6. Visitors** want to rate / describe websites

- flag explicit content
- inform authorities about illegal content
- flag probably unwanted content, f.i. by Amnesty International or Child protection.
- wikipedia style neutral descriptions
- ...

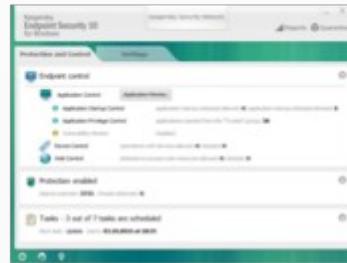
**Design big,  
Implement SMALL**

# Open Console

- discovered
- new websites
  - spam senders
  - SEO networks
  - meta-data
  - phishing sites
- Producers*



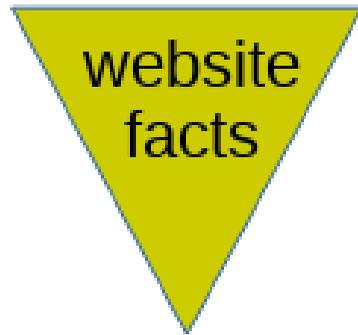
- applications
- crawlers
  - spam filters
  - indexing
  - research
  - protection
- Consumers*



  
website owner  
domain owner  
network owner



# Open Console, Facts

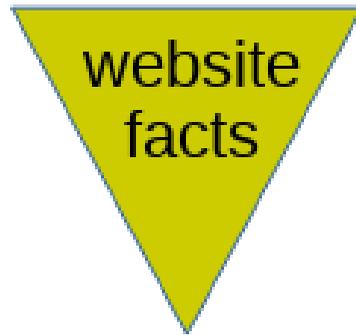


- Publisher maintains **facts**:  
( namespace-of-publisher ,  
normalized-url ,  
key-constant-by-publisher ,  
values ,  
expire  
)



# Open Console, Facts

**On the PetaFact scale!**

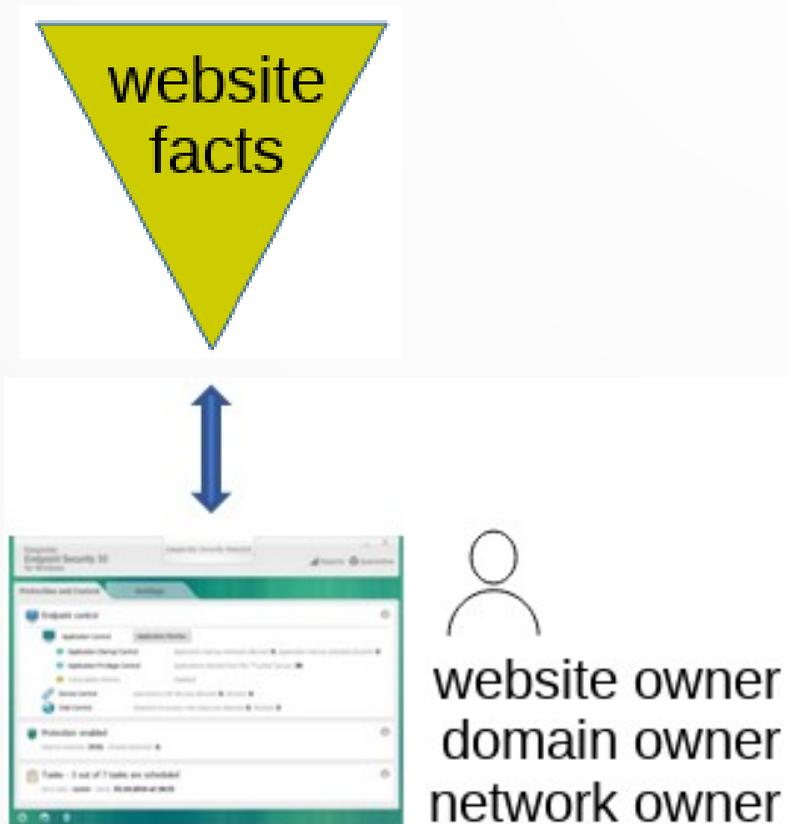


Consumer:

- sets-up filters
  - registers its **interest**
  - gets informed via **push**
- interprets key/values



# Open Console, User Interface



- Both producer & consumer
- Facts may be forms
- “Google Search Console”-alike
- Alternative implementations

# @Google

- Site ownership
- Crawl optimization
- Feedback

**Major competitive advantage**

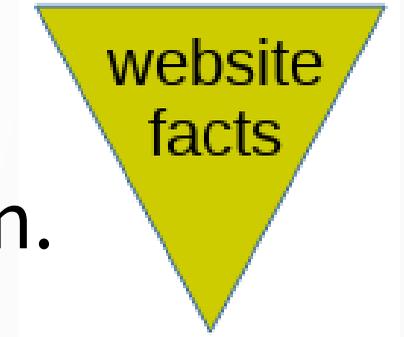
The screenshot displays the Google Search Console interface for the URL `http://mark.overme...`. The left sidebar contains a navigation menu with the following items: Overzicht, Prestaties (highlighted), URL-inspectie, Index (with sub-items: Dekking, Sitemaps, Verwijderingen), Functionaliteit (with sub-items: Paginafunctionaliteit, Site-vitaliteit), Beveiligingsproblemen en handmatige acties, Verouderde tools en rapporten, Links, and Instellingen.

The main content area shows the 'Prestaties' (Performance) section. At the top, it indicates the search type is 'Web' and the date range is 'Afgelopen 3 maand...'. Below this, three summary cards are visible: 'Totaal aantal klikken' (Total clicks) with a value of 0, 'Totaal aantal vert...' (Total conversions) with a value of 14, and 'Gem...' (Average position) with a value of 0%. A line chart titled 'Aantal klikken' (Number of clicks) shows data points for the dates 05-02-2022, 17-02-2022, 01-03-2022, and 13-03-2022. The chart shows a peak of 2 clicks on 17-02-2022 and 13-03-2022, and a peak of 1 click on 05-02-2022 and 01-03-2022.

At the bottom of the interface, there are three tabs: 'ZOEKOPDRACHTEN' (Queries), 'PAGINA'S' (Pages), and 'LANDEN' (Countries). The 'ZOEKOPDRACHTEN' tab is currently selected.

# @Open

- Join efforts of Bing, Baidu, ~~Yandex~~, DuckDuckGo, ...
- Public configuration
- Site owner proof, but also
  - domain owner
  - network owner / ISP
  - organization owner
- Cooperation in namespace design.
- Config form abstraction
- Structured ownerships and permission distribution



**>1G entities!**

# Open Console

- Not only for Crawling:
  - publishing email blacklists (and undo listing)
  - publishing “objective” rating on website content
  - offering commercial services to website owners

Generic interface towards the maintainers of the web, to

- distribute facts,
- battle monopolies, and
- implement law.

# Hard nuts

- Namespace ownership
  - hierarchy
  - use as OpenID
- Participation rules
- Juridical implications of owning Facts
- Attracting participation
- Commercial participation
- Fact ownership
- Fact distributing filters
- Fact cluster database
- UI battle of services
- UI form fields abstraction
- Translations
- Rolling upgrades of data

# Status



- First experiments have run @procoliX
- Requires ~15 cores to process 10TB/day

Open Console:

- Website
- Specs for forms
- Meshy Space
- Looking for participation!

