Open Search with Skrodon

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

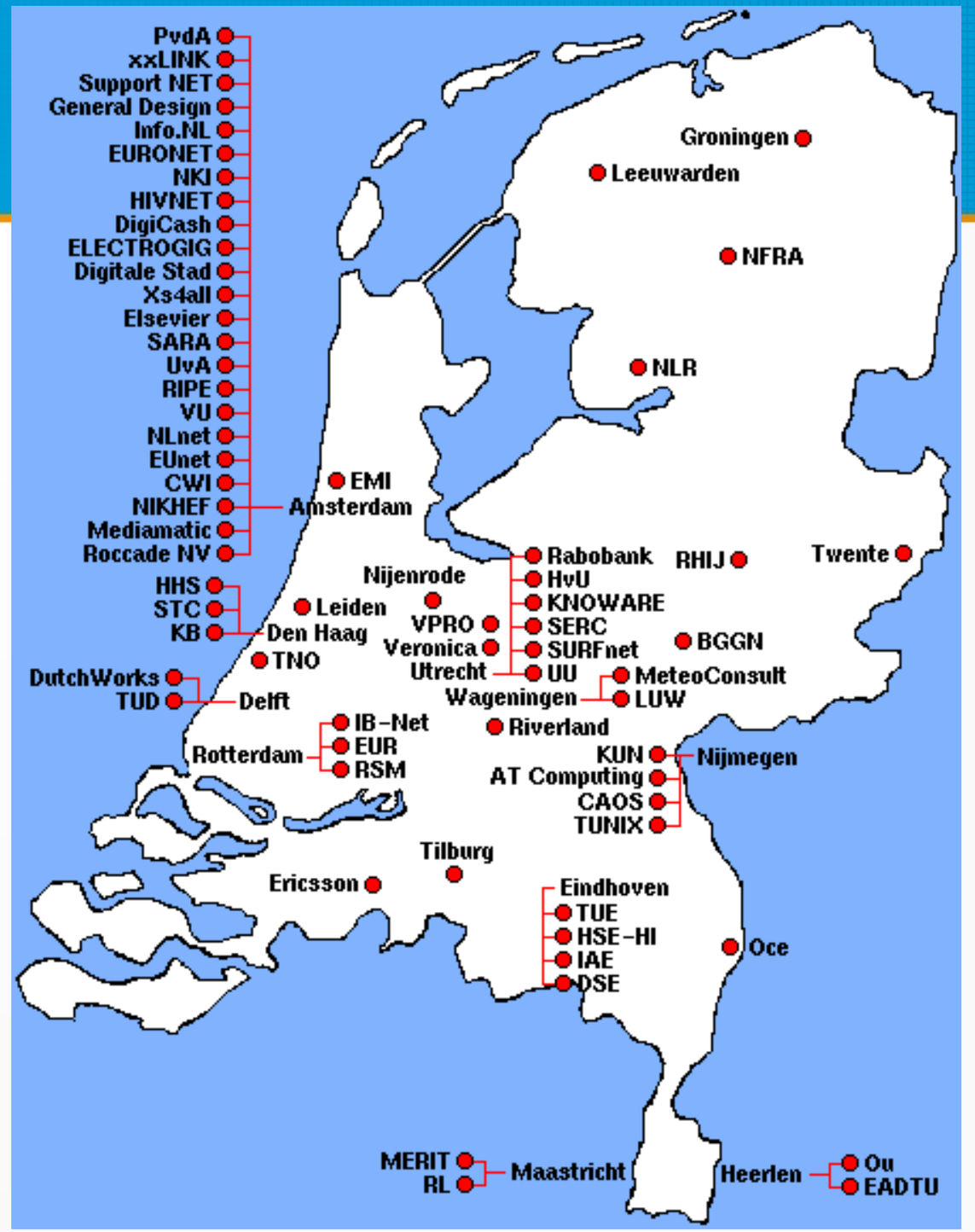
NLUUG NJ95

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

January 1994

NLUUG NJ95

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



June 1994

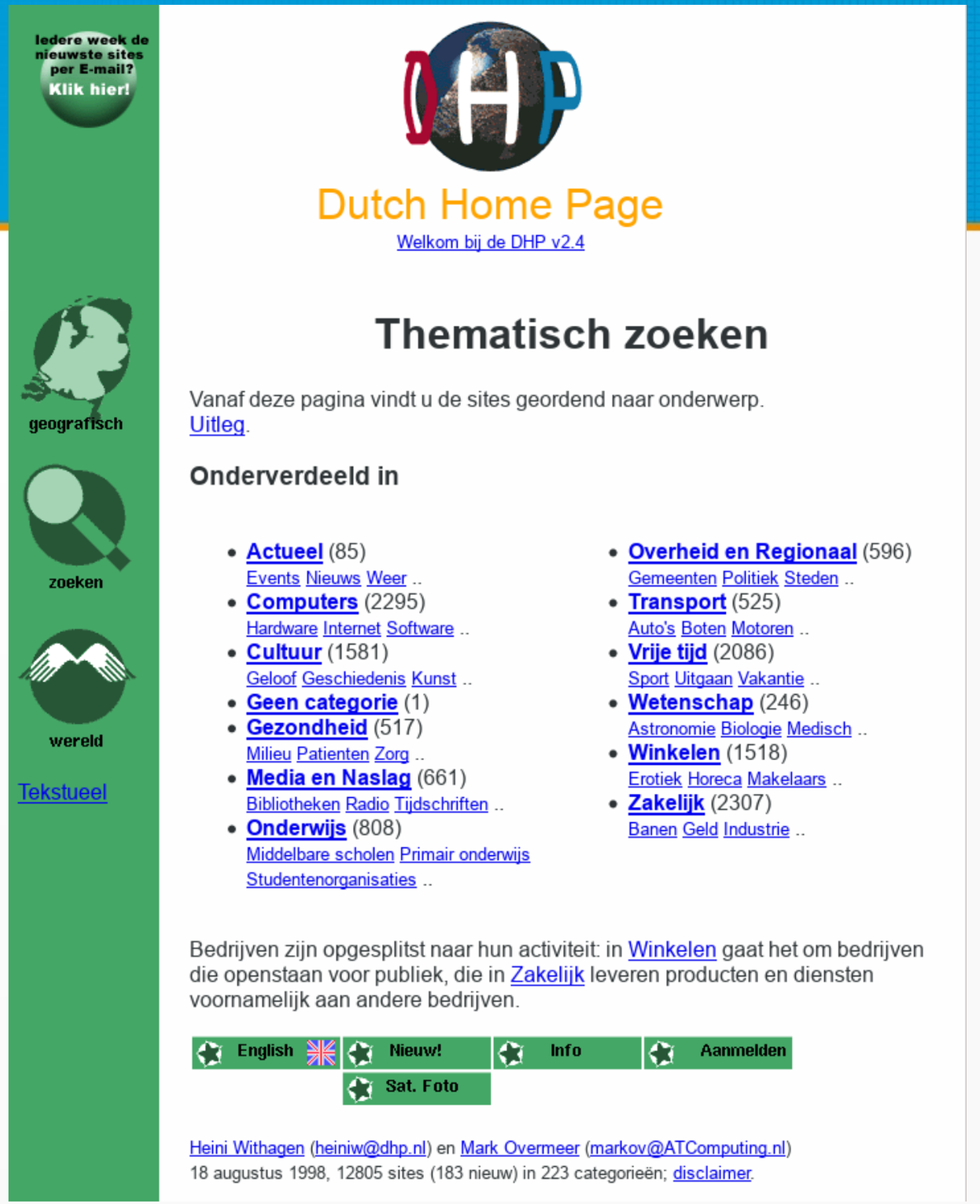
NLUUG NJ95

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

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

→ only tagging works

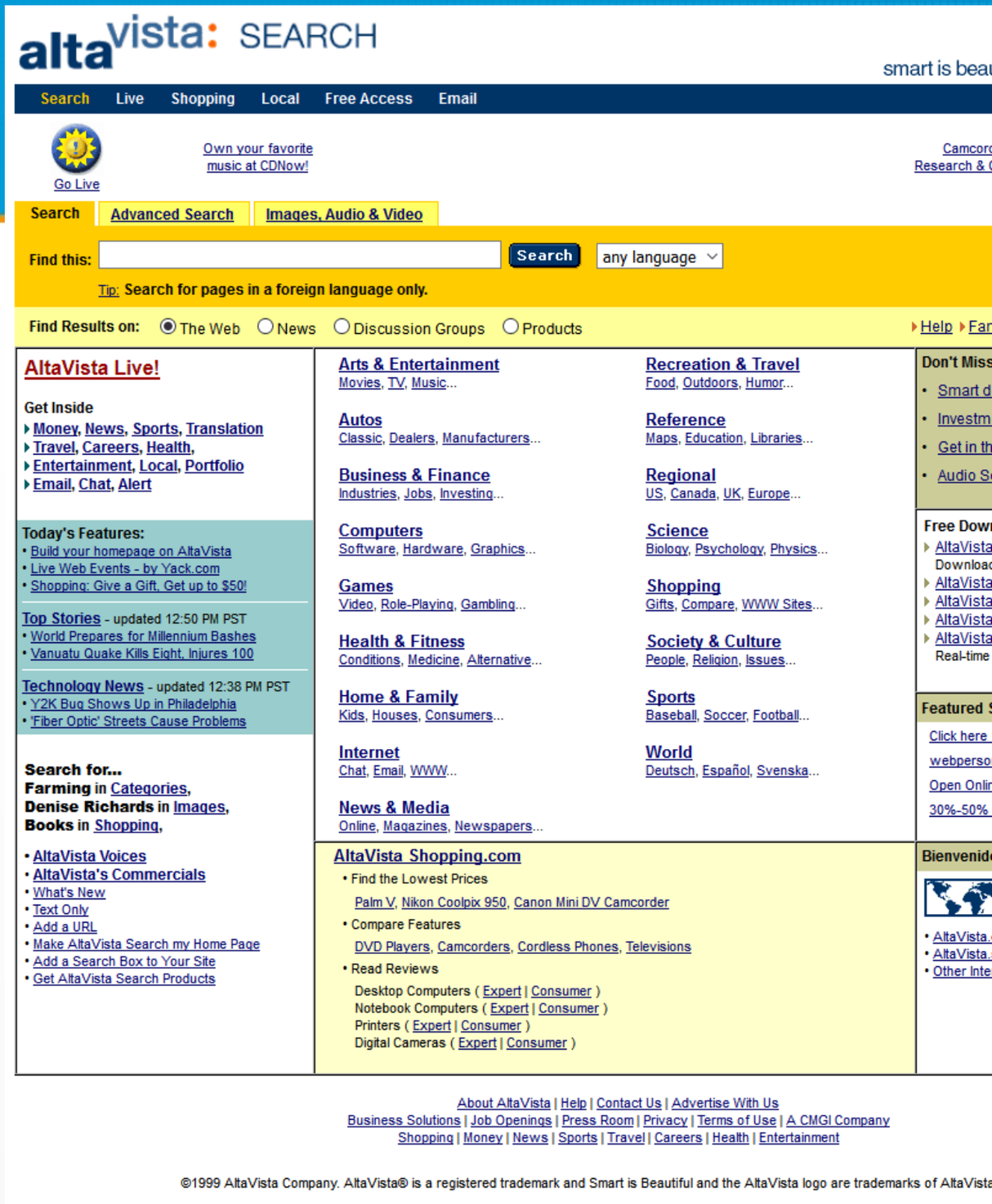
August 1998



Indices

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

➡ Text Search Engines?



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.

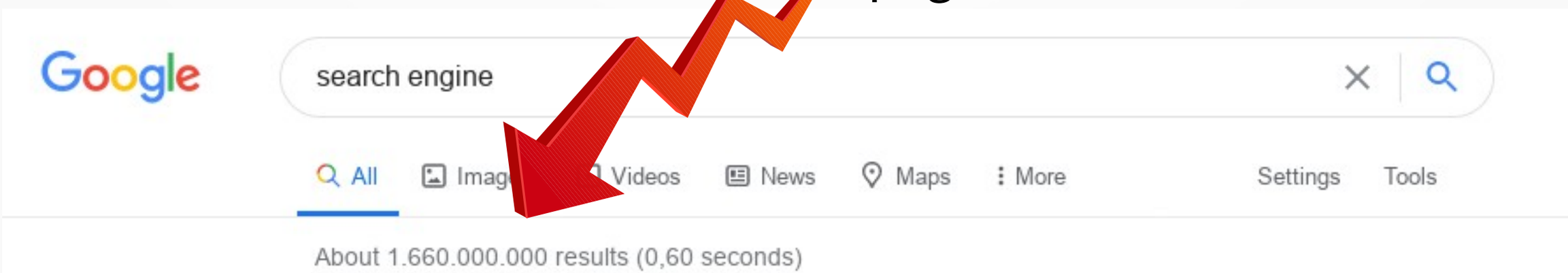
- Estimated #web-pages:
  - 1998: 2.4 **m**illion pages, 2.4e+6
  - 2020*: 5.5 **b**illion pages, 5.5e+9
- Estimate number of words per page: 400-800

*) source worldwidewebsize.com

I saw a problem

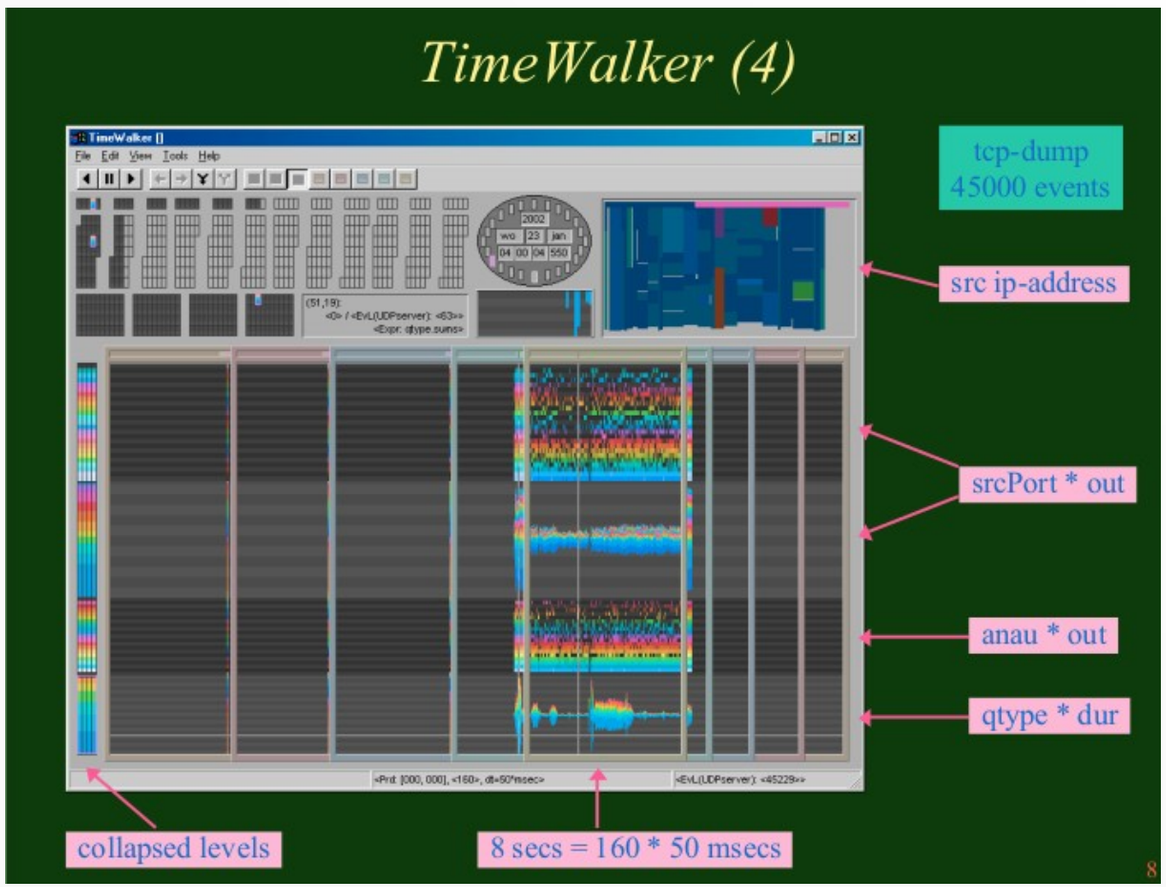
- (By example English) has 300k words.
- Native speakers know 10k to 40k of them.
- Non-natives max 3k.

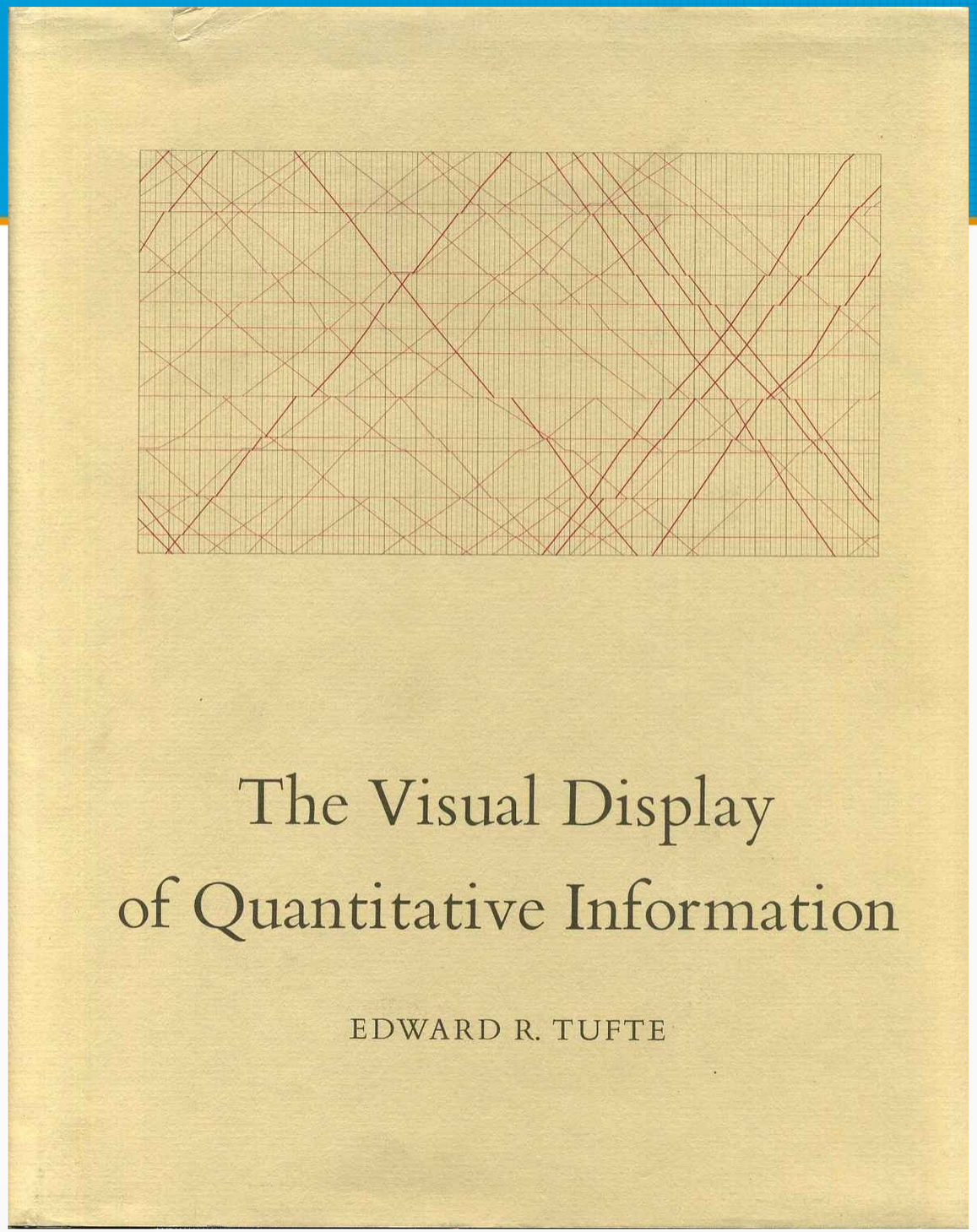
- Estimated #web-pages:
  - 1998: 2.4 million pages, 2.4e+6
  - 2020*: 5.5 billion pages, 5.5e+9
- Estimate number of words per page: 400-800



other options?

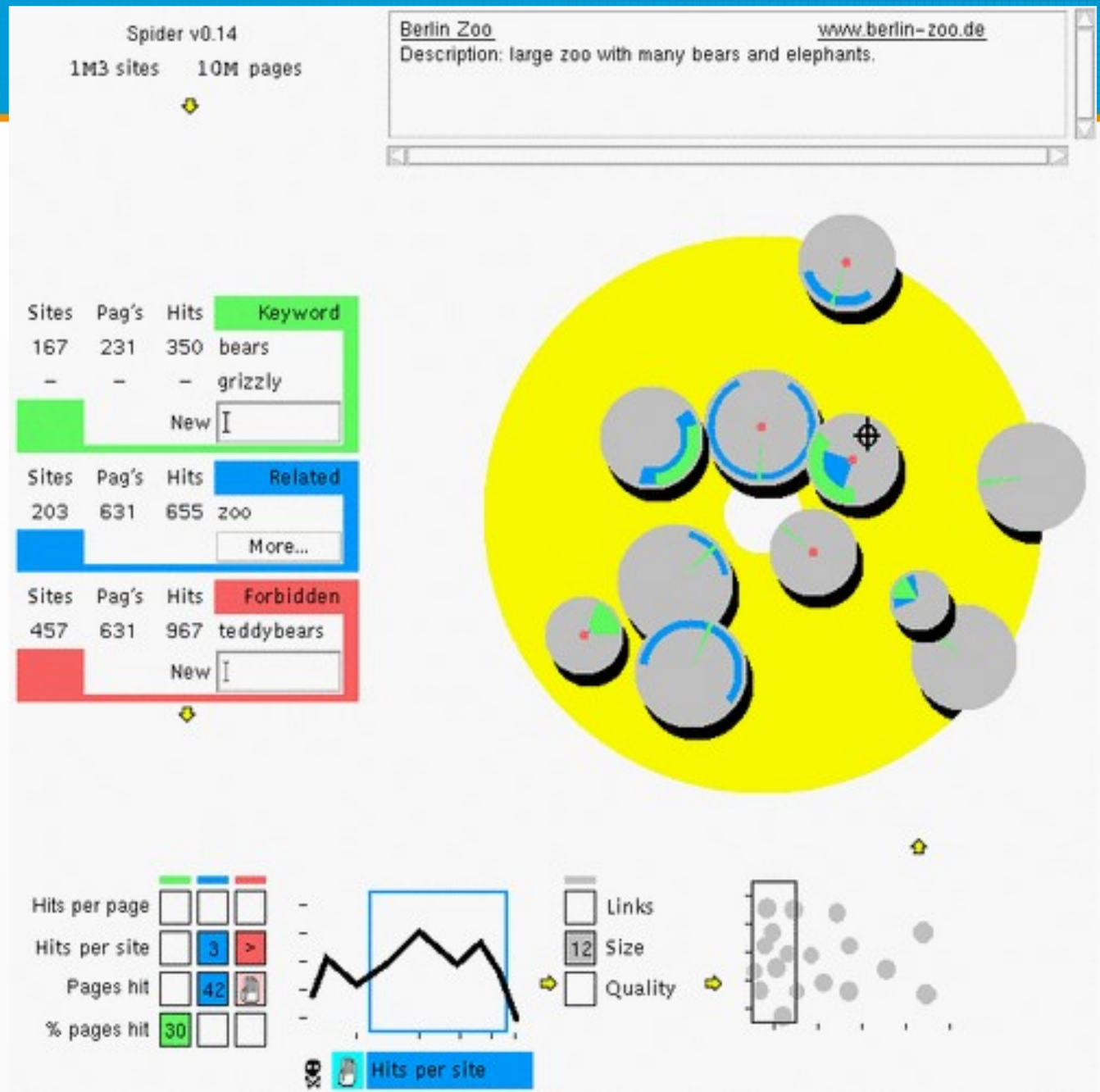
nlnet FOUNDATION





. . . 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”)

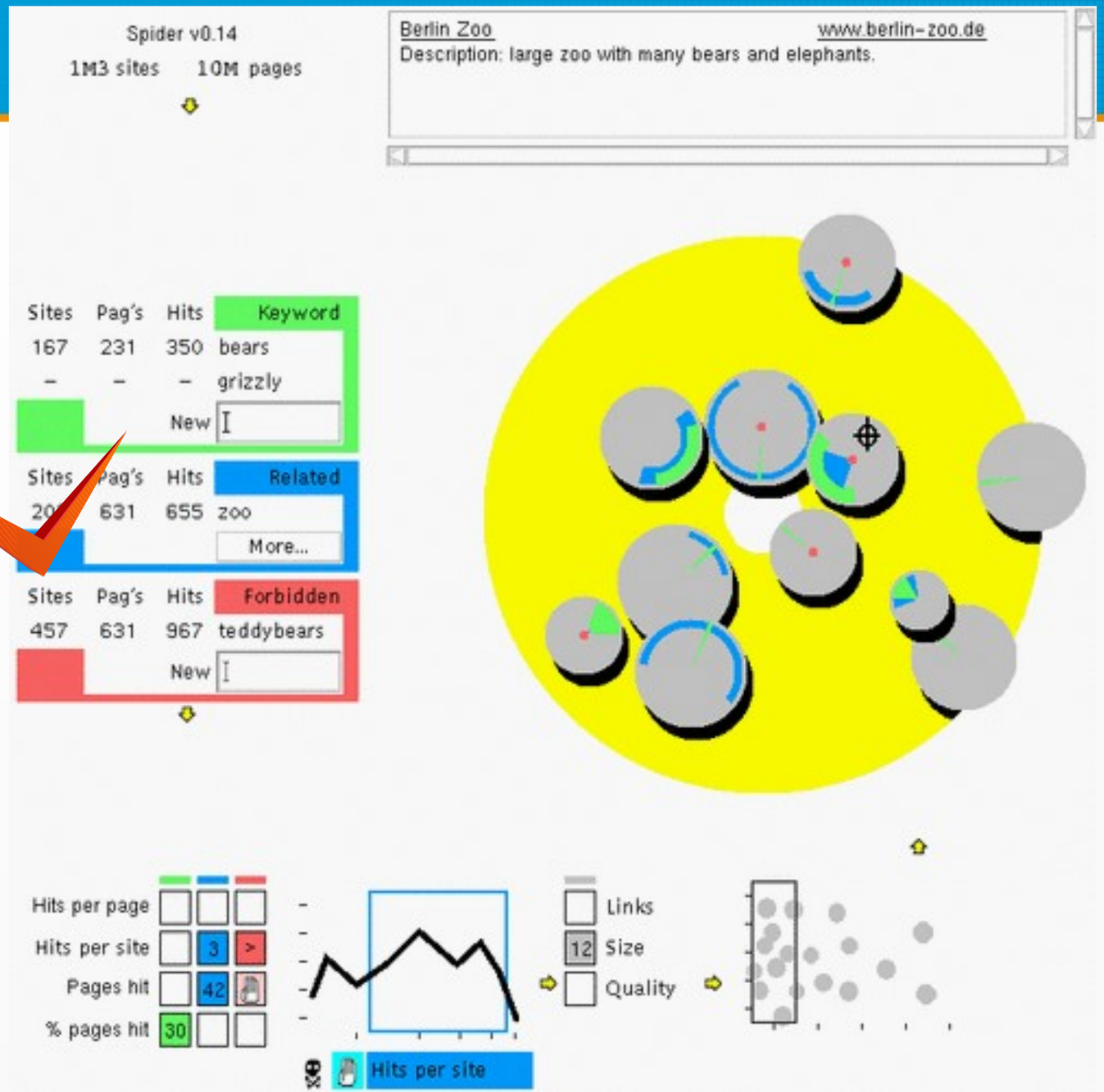


written inJava1 on “Linux” 0.95, 14k4 modem
on PC 486DX2/66, 16MB RAM, 80MB disk (3000€)

. . . do better?

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

No data to explore my ideas!



Use case 1

Evolved?

1996

2020

Google

vaseline

www.vaseline.com › All Products

All Products | Unilever Vaseline®

Vaseline® Lip Therapy® Original Lip Balm Tin formula is clinically proven to heal dry and…

www.vaseline.nl ▾ Translate this page

Vaseline

Ontdek de nieuwste innovatie in huidverzorgingsproducten van het merk **Vaseline**. Meer informatie over de volgende generatie **Vaseline** Intensive ...

www.bol.com › vaseline ▾ Translate this page

Vaseline artikelen kopen? Alle artikelen online - bol.com

Hij noemde het product **Vaseline** Petroleum Jelly. Het merk **Vaseline** was geboren. Door de jaren heeft **Vaseline** een groot assortiment aan producten ...

★★★☆☆ Rating: 3,1 - 29,825 reviews

www.unilever.nl › personal-care › va... ▾ Translate this page

Vaseline | Onze merken | Unilever Nederland

Het is de oorsprong van alle **Vaseline**-producten. Helend en verzachtend. Robert Chesebrough werkte als 22-jarige apotheker in de aardolie-industrie, hier zag hij ...

www.facebook.com › ... › Other › Brand › Health/Beauty ▾

Vaseline - Home | Facebook

Vaseline. 6951974 likes · 595 talking about this. **Vaseline**® has been helping skin look beautiful & feel healthy since 1870.

www.ah.nl › ... › Merken › Merk V ▾ Translate this page

Vaseline producten bestellen | Albert Heijn

Op zoek naar producten van **Vaseline**? Bekijk het overzicht van **Vaseline** producten bij Albert Heijn. ✓ Vandaag besteld, morgen thuisbezorgd.

Searches related to vaseline

vaseline **lip**	vaseline **lips**
vaseline **uses**	vaseline **lotion**
vaseline **petroleum jelly**	vaseline **original**
vaseline **lip therapy**	vaseline **ingredients**

Goooooooooogle ›

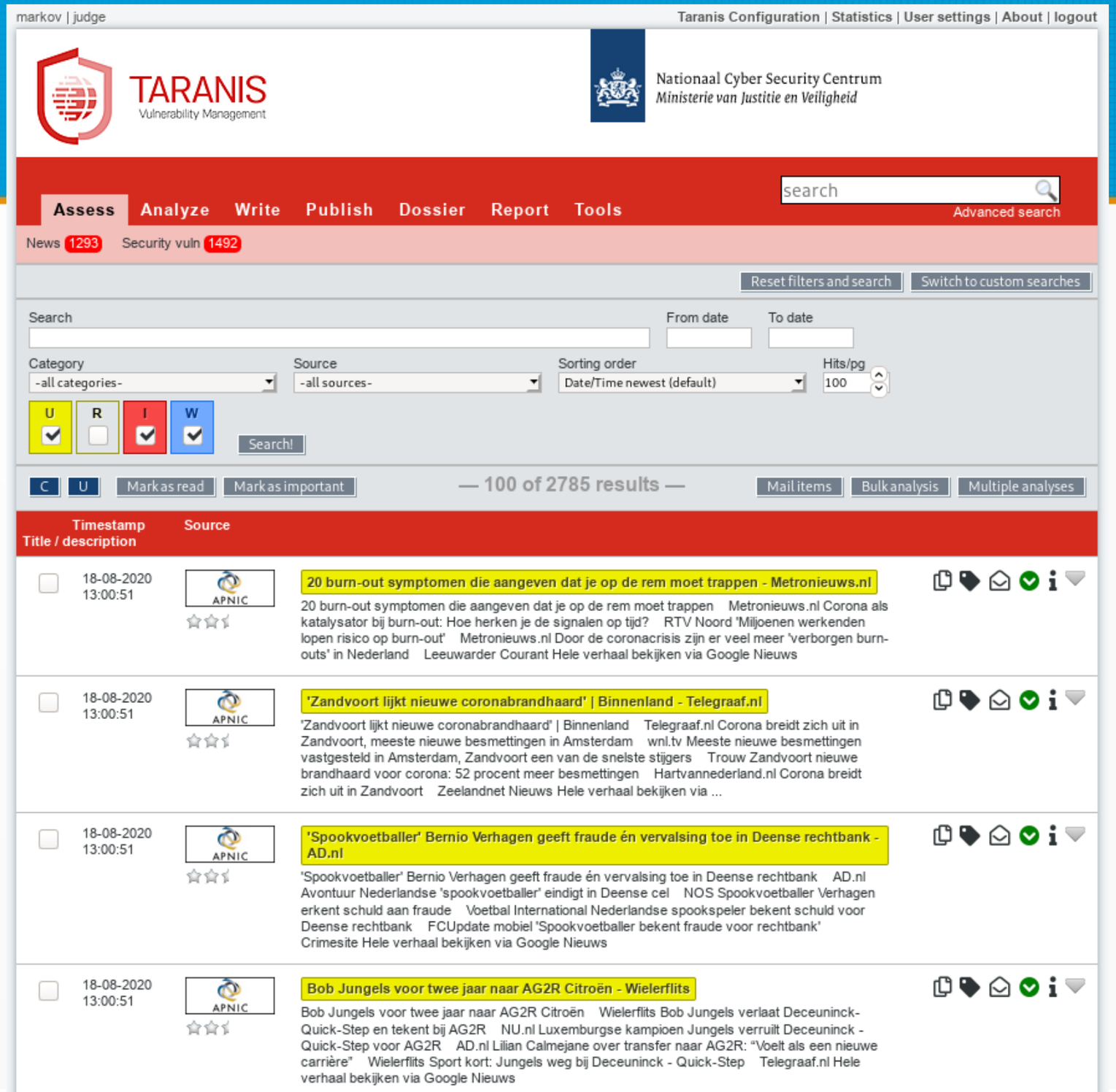
1 2 3 4 5 6 7 8 9 10 Next

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



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

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.

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.

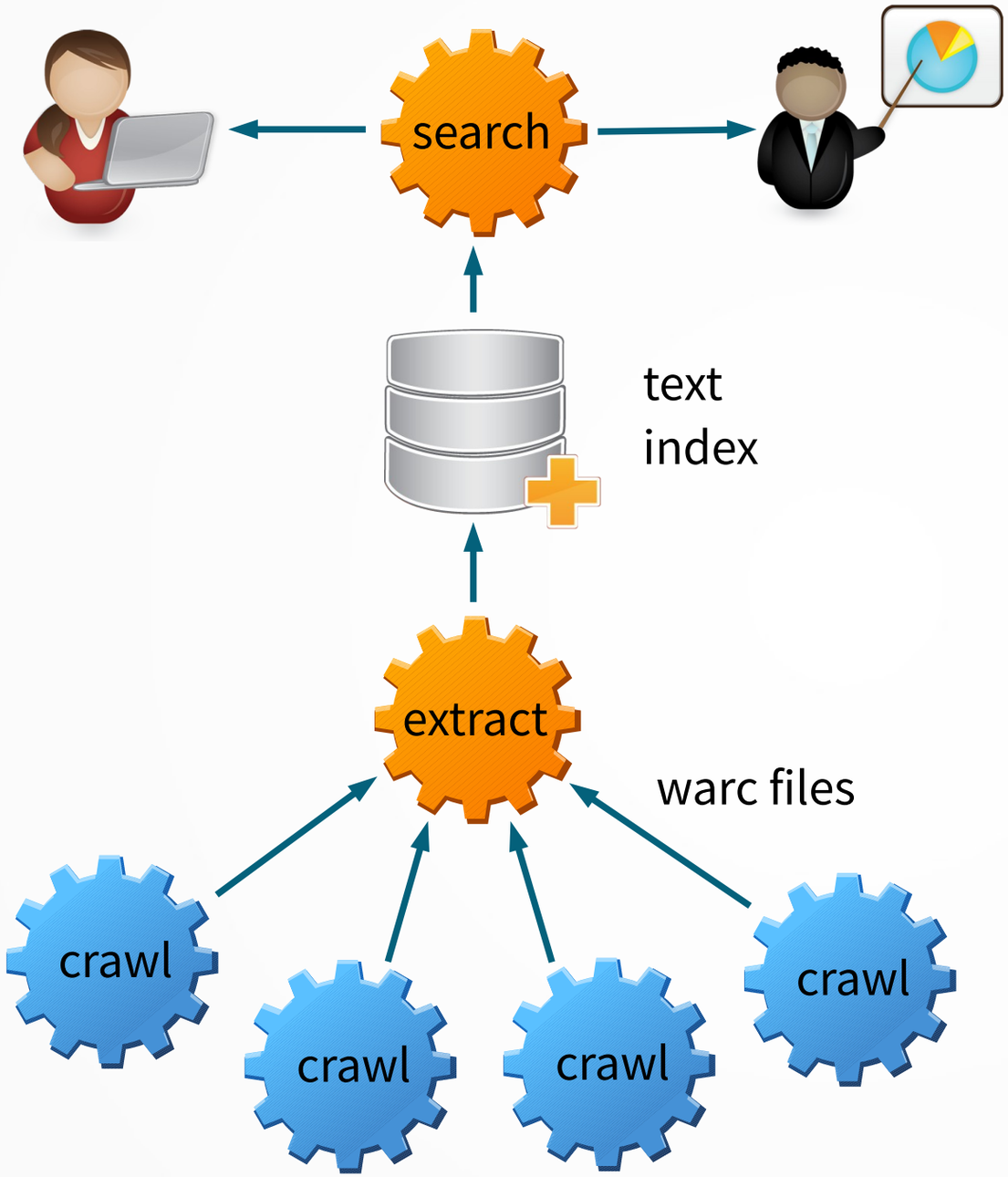
Merging Search Engine Needs

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

Merging Search Engine Needs

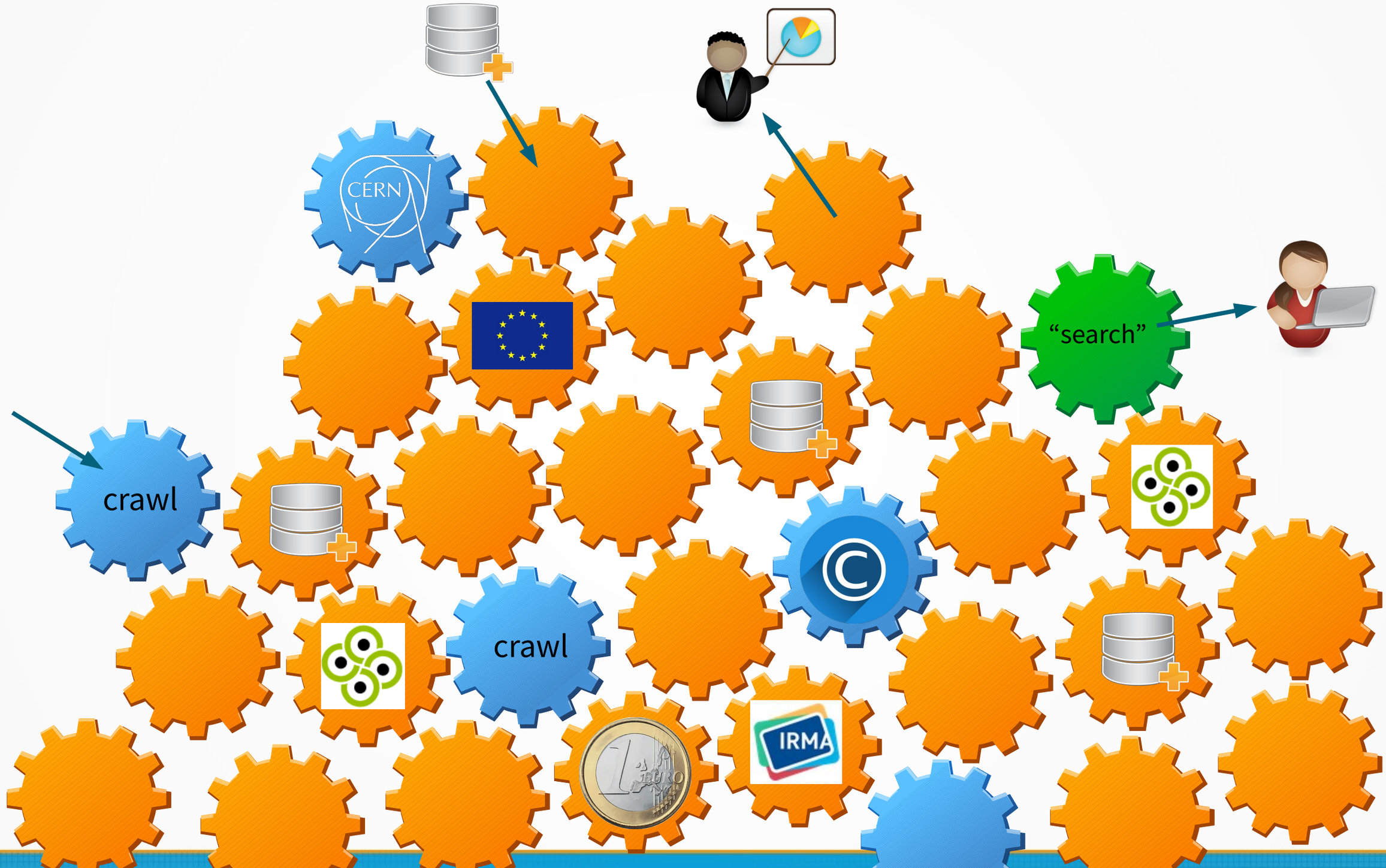
- People have **different** opinions about what “searching the web” is.
- Many, many **overlapping** components between the use-cases.
- Collection of web-based information at the moment is **very primitive**, **ad-hoc** and **expensive**.

➔ Let’s evolve!



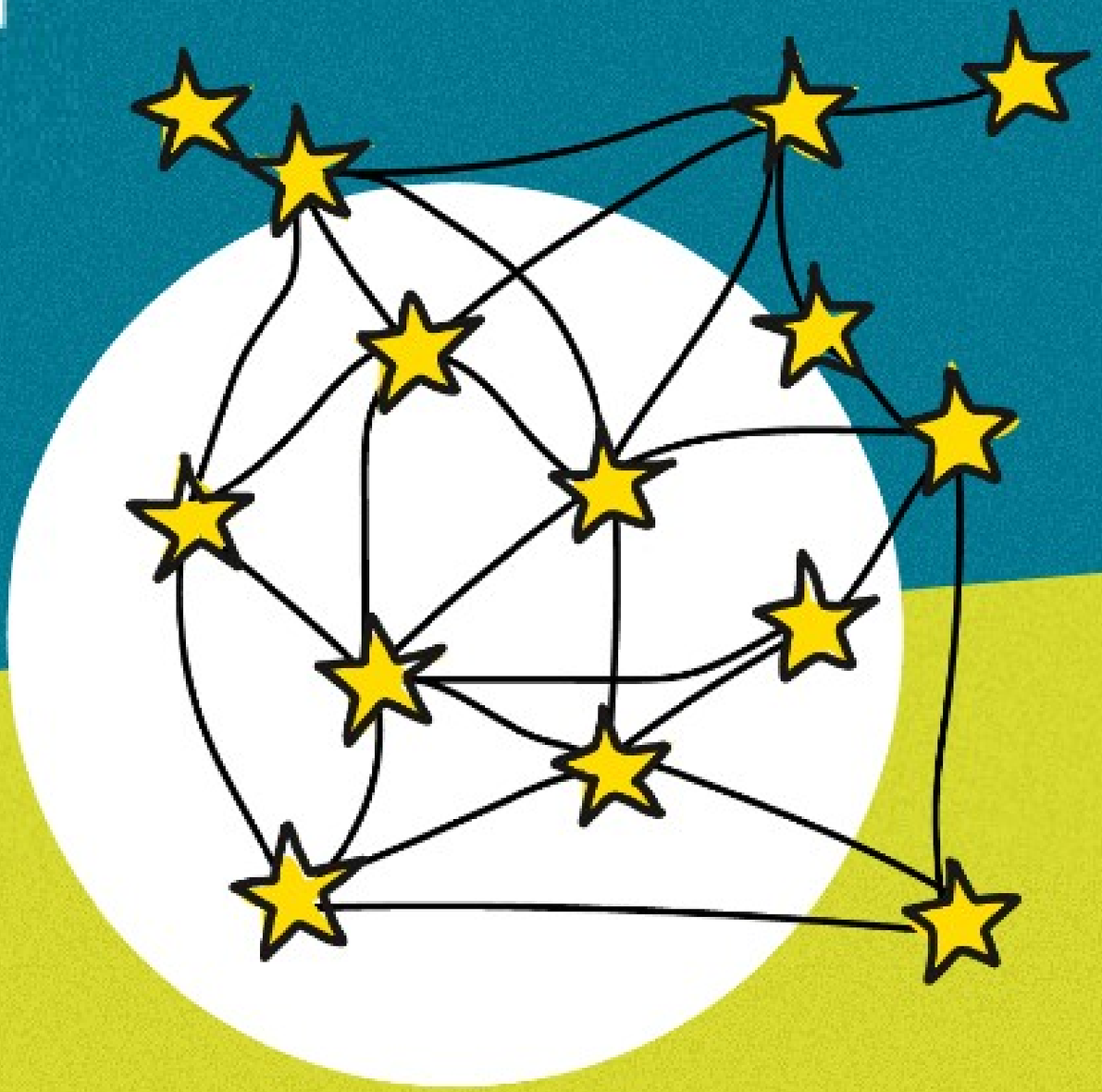
This is **not** how a "Google scale" can be made to work.

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


CERN
“search”
crawl
crawl
IRMA

open search foundation

standardize
cooperation
(like shops)



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

?

SKRODON

-- established 2022 --

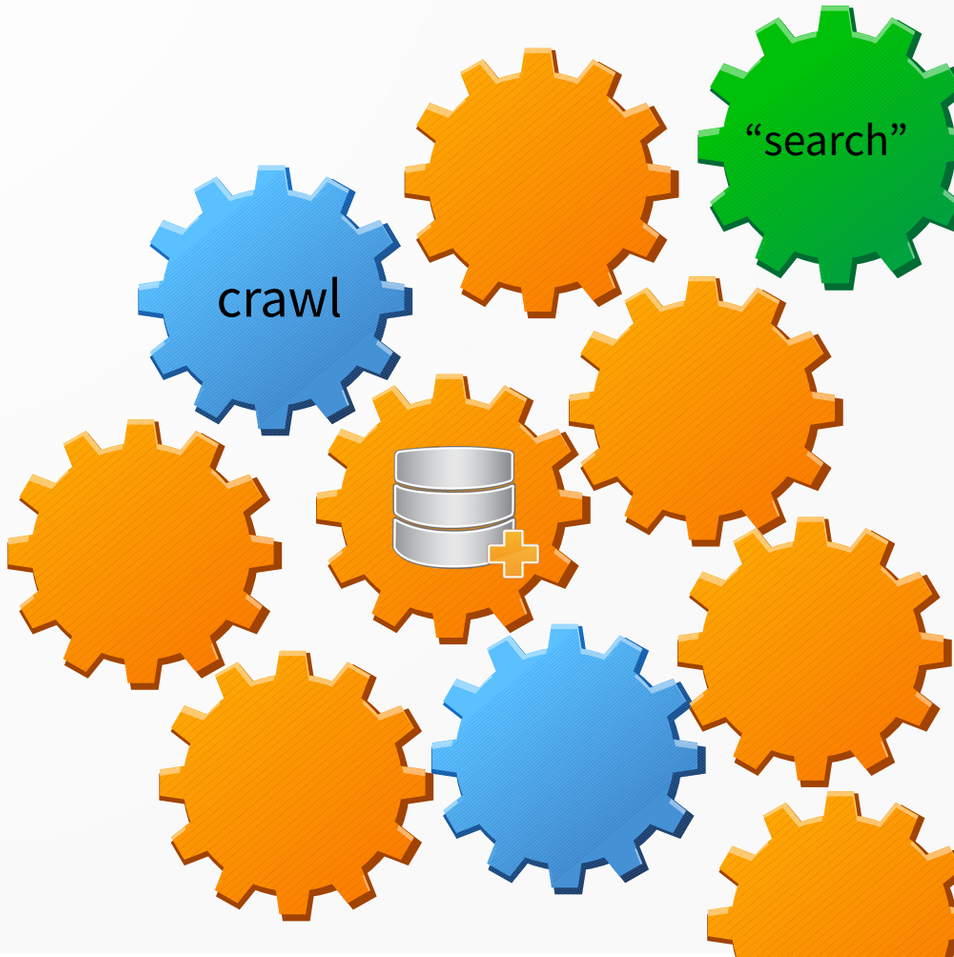
!

Good Behavior

SKRODON

-- established 2022 --

= ⊓



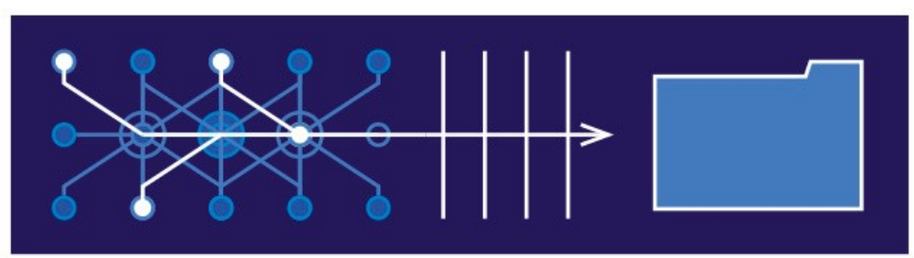
First components in **Skrodon**

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

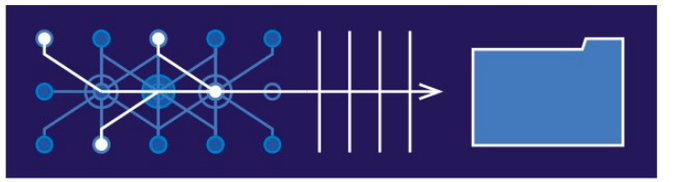
- Meshy Space

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



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!

Open Console

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

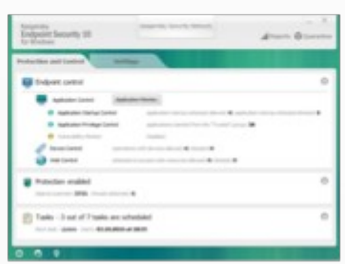
discovered
- new websites
- spam senders
- SEO networks
- meta-data
- phishing sites

Producers

website facts

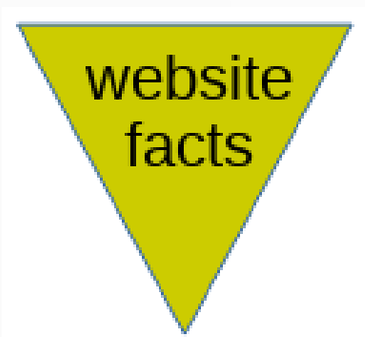
applications
- crawlers
- spam filters
- indexing
- research
- protection

Consumers



website owner
domain owner
network owner

Open Console, Facts

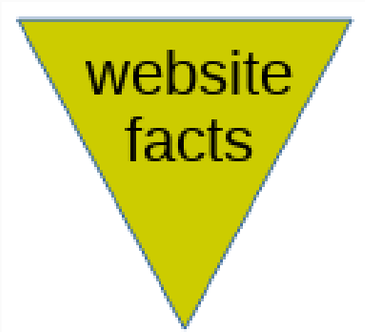


- Publisher maintains **facts**:

```
(  namespace-of-publisher,
   normalized-url,
   key-constant-by-publisher,
   values,
   expire
)
```

Open Console, Facts

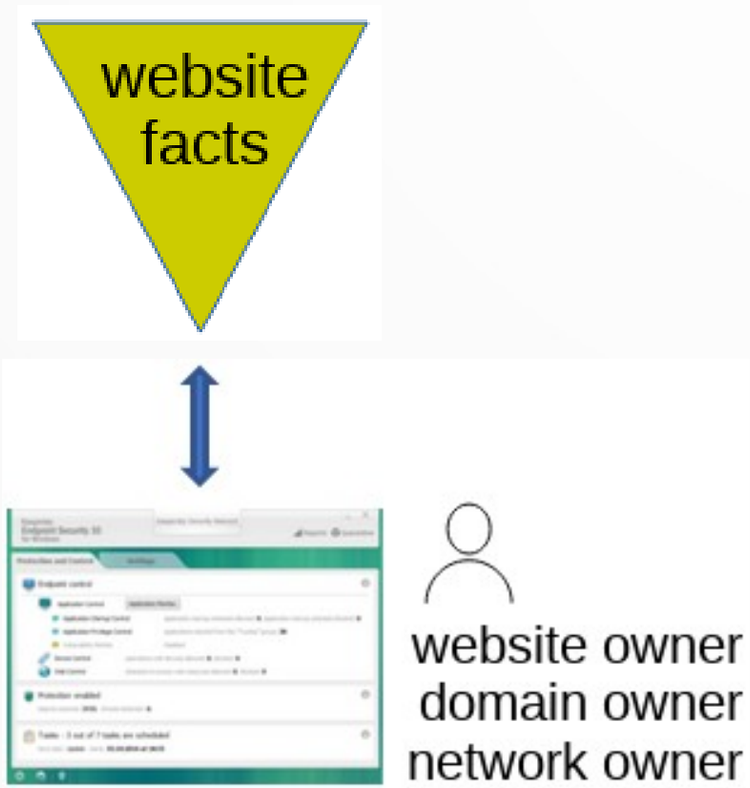
On the PetaFact scale!



Consumer:

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

Open Console, User Interface

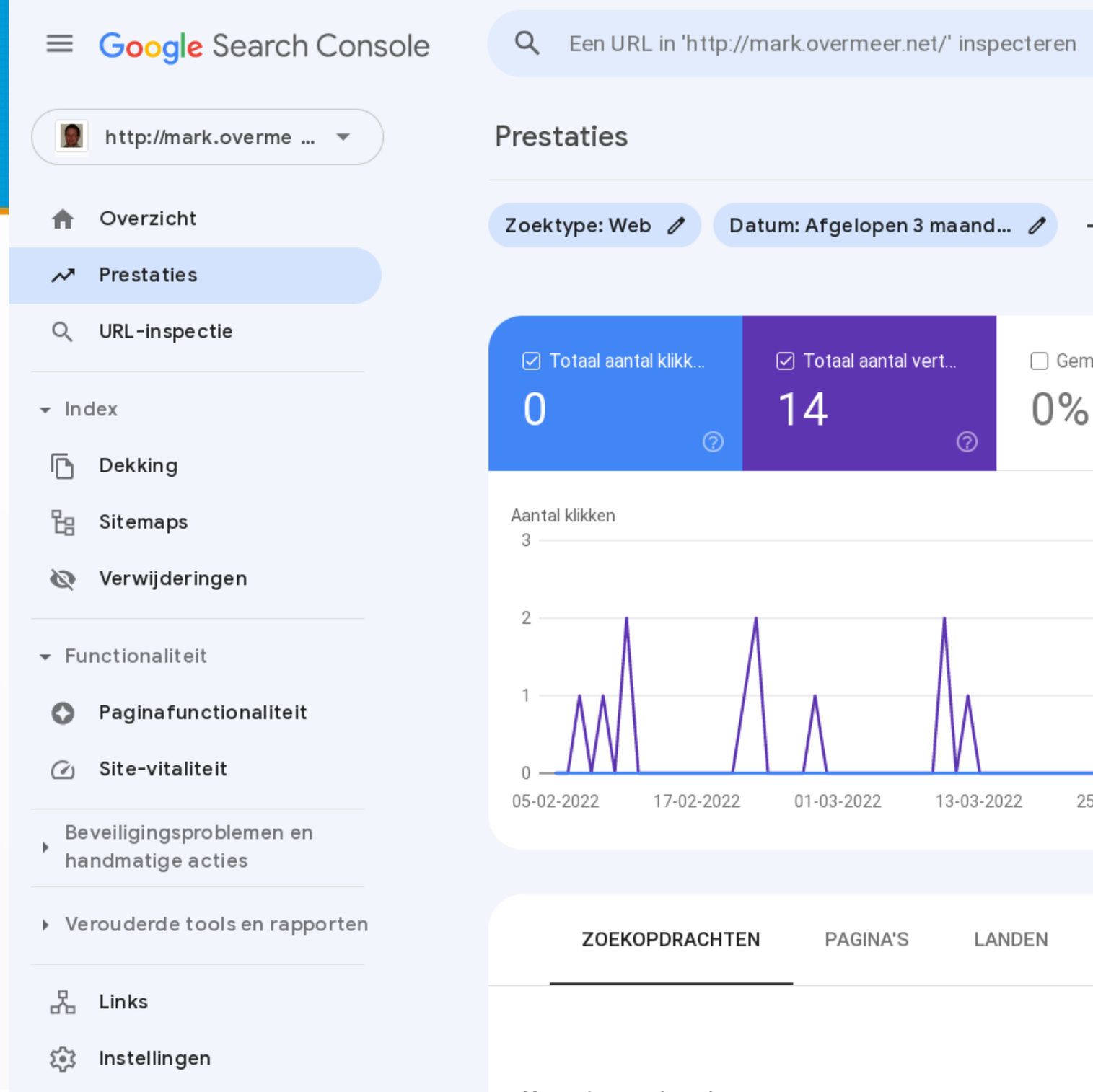


- Both producer & consumer
- Facts may be forms
- “Google Search Console”-alike
- Alternative implementations

@Google

- Site ownership
- Crawl optimization
- Feedback

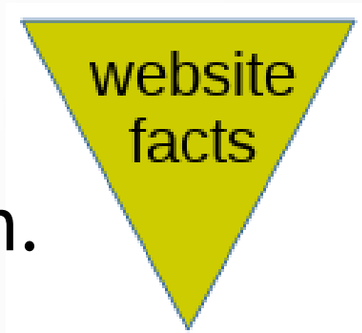
Major competative advantage



@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

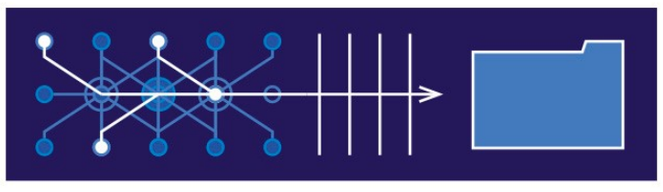
Generic interface towards the maintainers of the web, to

- 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!