Štěpán Škrob <stepan.skrob@firma.seznam.cz>

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

O čem bude přednáška?

Úvod

- Architektura
- Vyhledávání
- Robot
- Údaje z provozu

Konec

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Úvod

- Vyhledávače jsou si prakticky velmi podobné, liší se pouze v implementačních detailech :-)

- Jako auta…

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Část 1 – Architektura

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

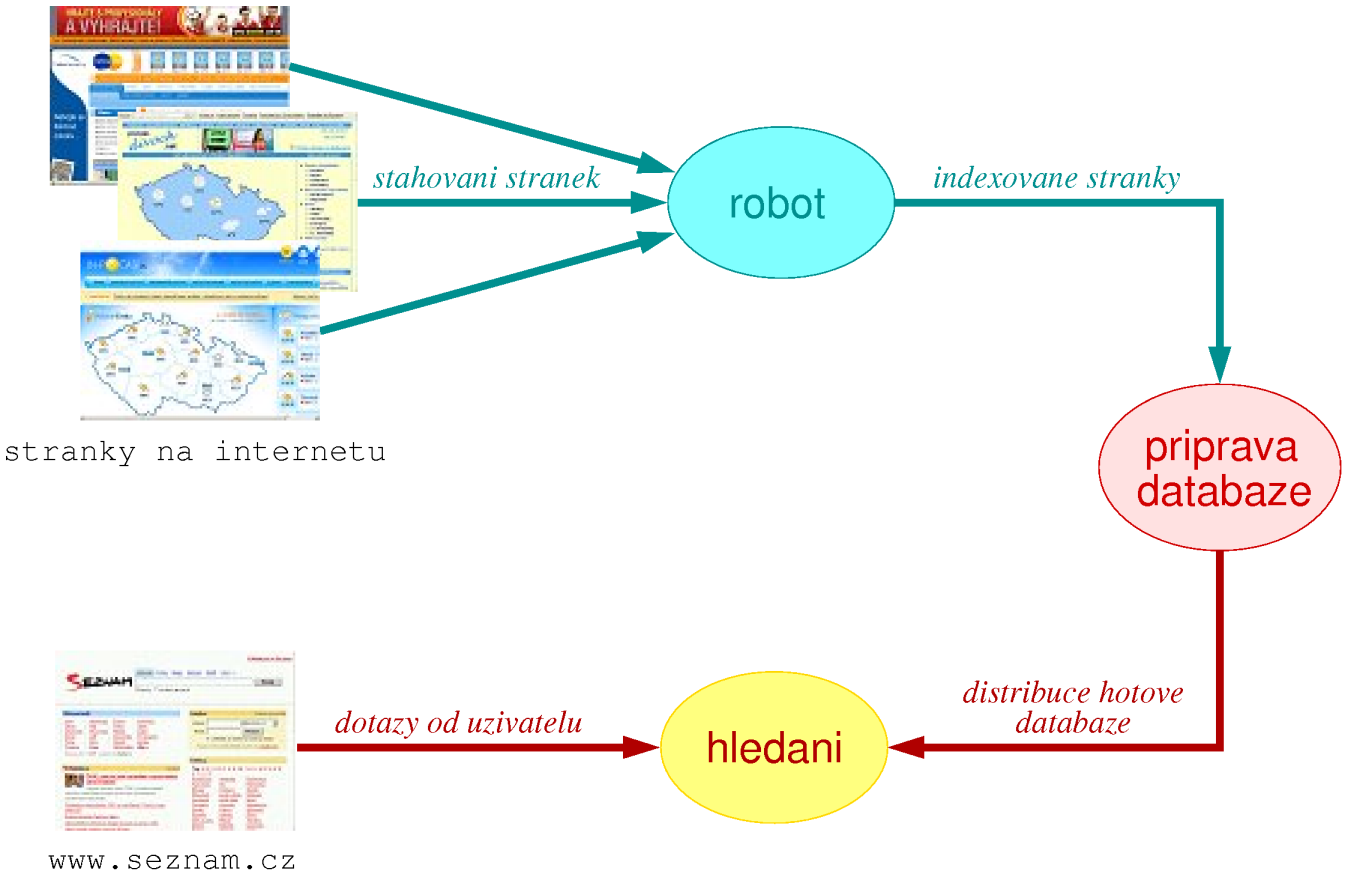
Architektura

1. Hlavní části
2. Dvě serverovny
3. Blokové schéma
4. Hardware

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

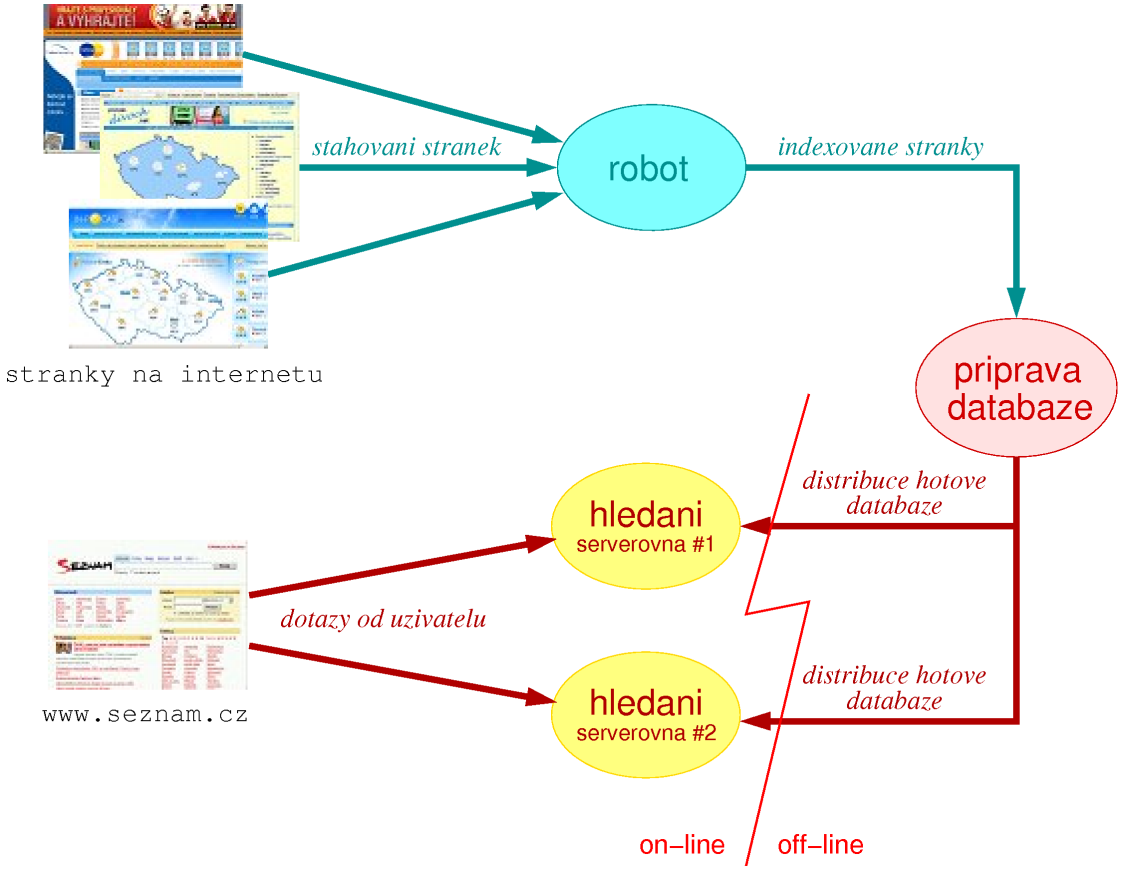
SEZNAM

Hlavní části



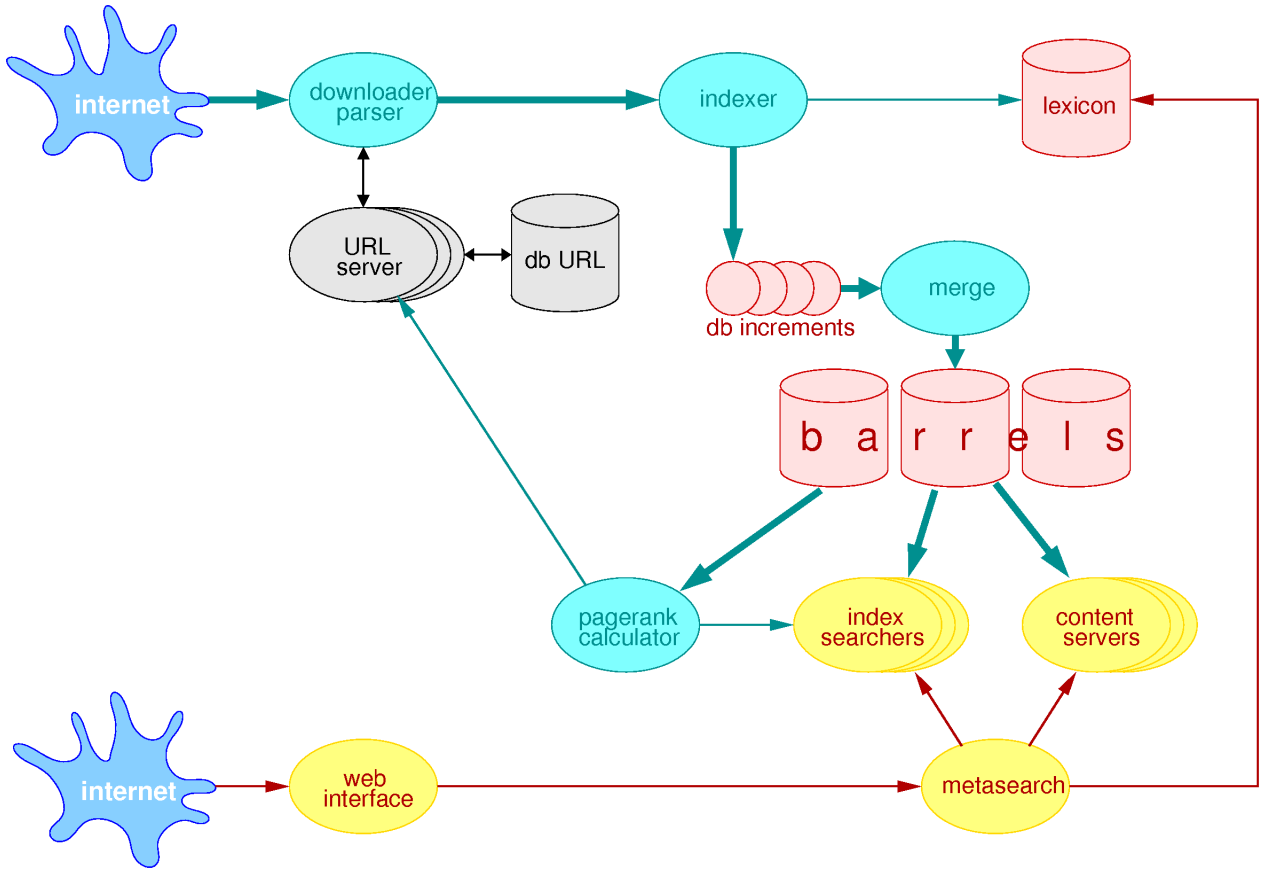
www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Dvě serverovny



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Blokové schéma



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hardware

- Robot + příprava databáze
  - 10 serverů, různé konfigurace:
    2x dual core CPU, 2-4 GB RAM, SAS/SATA disky
- Vyhledávání
  - 20 serverů x 2 serverovny, většina je:
    2x dual core CPU, 2 GB RAM, 6x140 GB SAS

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

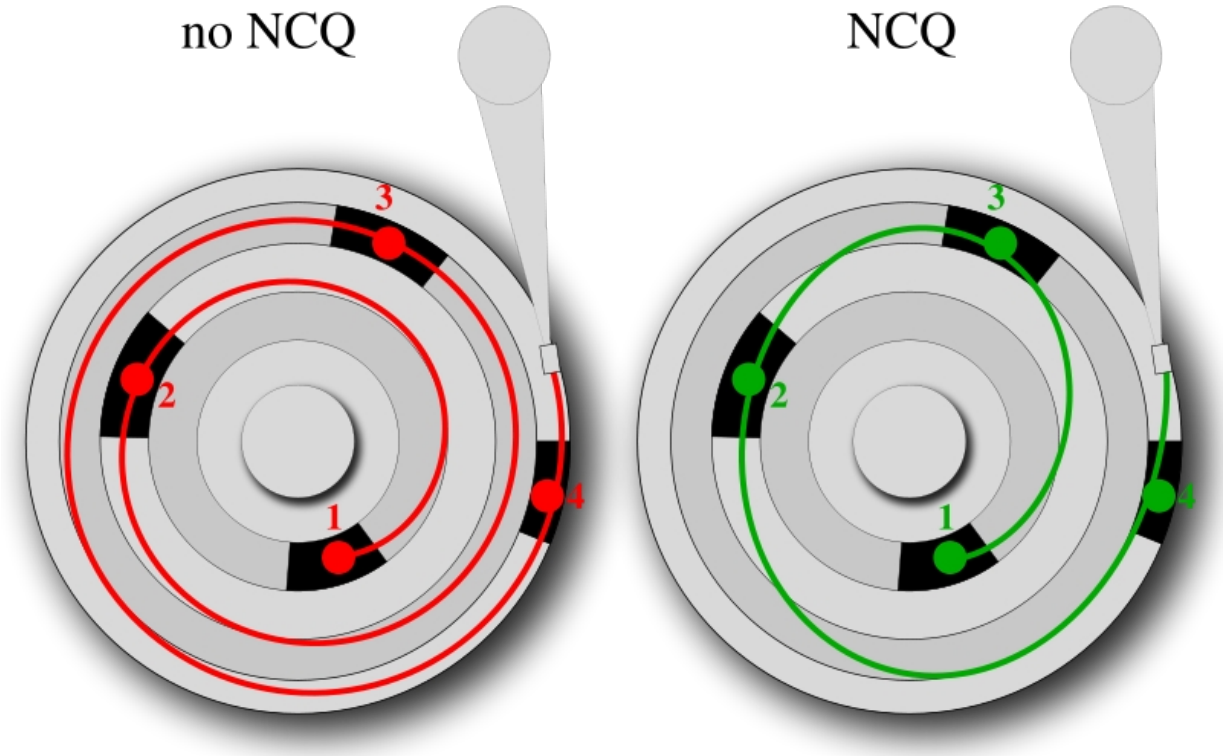
Hardware: SAS vs. SATA

	SAS serial attached SCSI	SATA serial ATA
Kapacita	70-140 GB	750-1000 GB
Otáčky	15,000	7,200
Seek	3 msec	9 msec
Optimalizace req.	TCQ	NCQ
IO operace/sec	300	180

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hardware: TCQ, resp. NCQ



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Část 2 – Vyhledávání

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Vyhledávání

1. Zadávané dotazy
2. Lemmatizace
3. Hodnocení stránek

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Zadávané dotazy (1)

- 10 náhodných dotazů
  - posilovna
  - plné hry ke stažení zdarma
  - plemena koní
  - planovac tras
  - petra němcová fotky
  - paragrafy a zákony
  - papírové vystřihovánky
  - panenka chou chou
  - paintball bazar
  - oplocení

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Zadávané dotazy (2)

- Forma dotazů:
  - přídavná a podstatná jména,
  - 1. pád,
  - jednotné i množné číslo,
  - občas bez diakritiky.

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Lemmatizace (1)

- Lemma = základní tvar slova
- Lemmatizátor
  - Vstup: slovo
  - Výstup: lemma, morfologické informace (slovní druh, pád, číslo, osoba…)
- Nejednoznačnost: stát, ženu, tancích…

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Lemmatizace (2)

- Věta:
  Jeden z nejlepších zdrojů o německých tancích.

- Lemmatizováno:
  Jedna/Jíst z dobrý zdroj o německý tank/tanec.

- Disambiguace = vyloučení nejednoznačnosti

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Hodnocení stránek (1)

Úvod - Statutární město **Brno**

... návrhu multifunkčního hodinového stroje s meteorologickou stanicí pro náměstí Svobody v **Brně** 11.10.2007. 2007 Deklarace Rady města **Brna** na podporu rodiny 31.10.2007.

www.**brno**.cz/

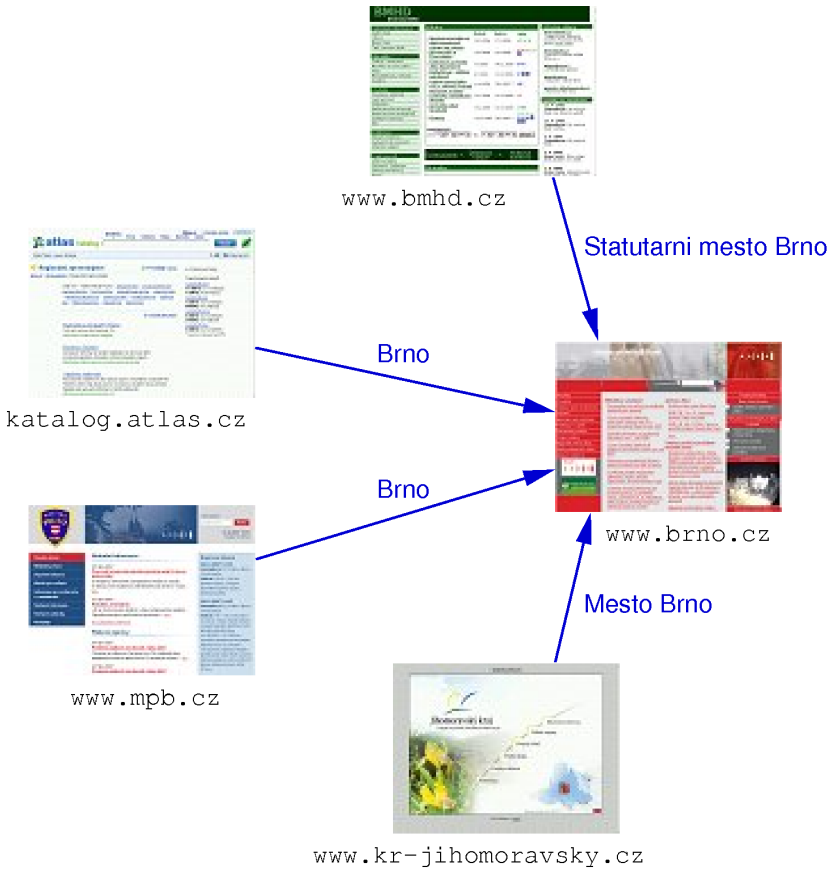
Další nalezené stránky na www.brno.cz

- **Titulek !!**
- Obsah stránky
- URL

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hodnocení stránek (2)



Citační analýza pro dotaz „brno“

Hodnocení stránek (3)

- Pagerank = statická „důležitost“ stránky založená na citační analýze

- Předpoklad: statisticky náhodné chování
- SPAM není jenom e-mail
  (každý systém se přizpůsobí kritériím, podle kterých je hodnocen :-)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Část 3 – Robot

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Robot

1. Hledání nových stránek
2. Reindexace stránek
3. Ne-HTML formáty

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hledání nových stránek (1)

- Před 3 lety start na www.seznam.cz
- Od té doby procházení nalezených odkazů

- Domény .cz, .sk, .com, .org, .net, .info, …
- Hledá stránky v českém jazyce

- Alternativní zdroje: RSS, články.cz, apod.

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hledání nových stránek (2)

- Robots.txt – standardní protokol pro zakázání přístupu robotů (www.robotstxt.org)
- Textový soubor http://example.com/robots.txt

```
# comment
User-Agent: *
Disallow: /statistiky

User-Agent: AnnoyingBot
Disallow: /
```

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hledání nových stránek (3)

- Sitemap.xml – nový protokol pro manifestaci existence stránek
- Obvykle na http://example.com/sitemap.xml

```
… <url>
        <loc>http://www.developstudio.com/</loc>
        <lastmod>2007-10-30T16:31:04+00:00</lastmod>
        <changefreq>daily</changefreq>
        <priority>1.0</priority>
    </url> …
```

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Reindexace stránek (1)

- Každý den se vybere množina stránek pro reindexaci
- Při výběru se hodnotí
  - Datum poslední návštěvy
  - Rank
  - Frekvence změn

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Reindexace stránek (2)

- Přetěžování webserverů
  - Shapování podle IP adresy
  - Omezení max počet URL / sec

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Ne-HTML formáty

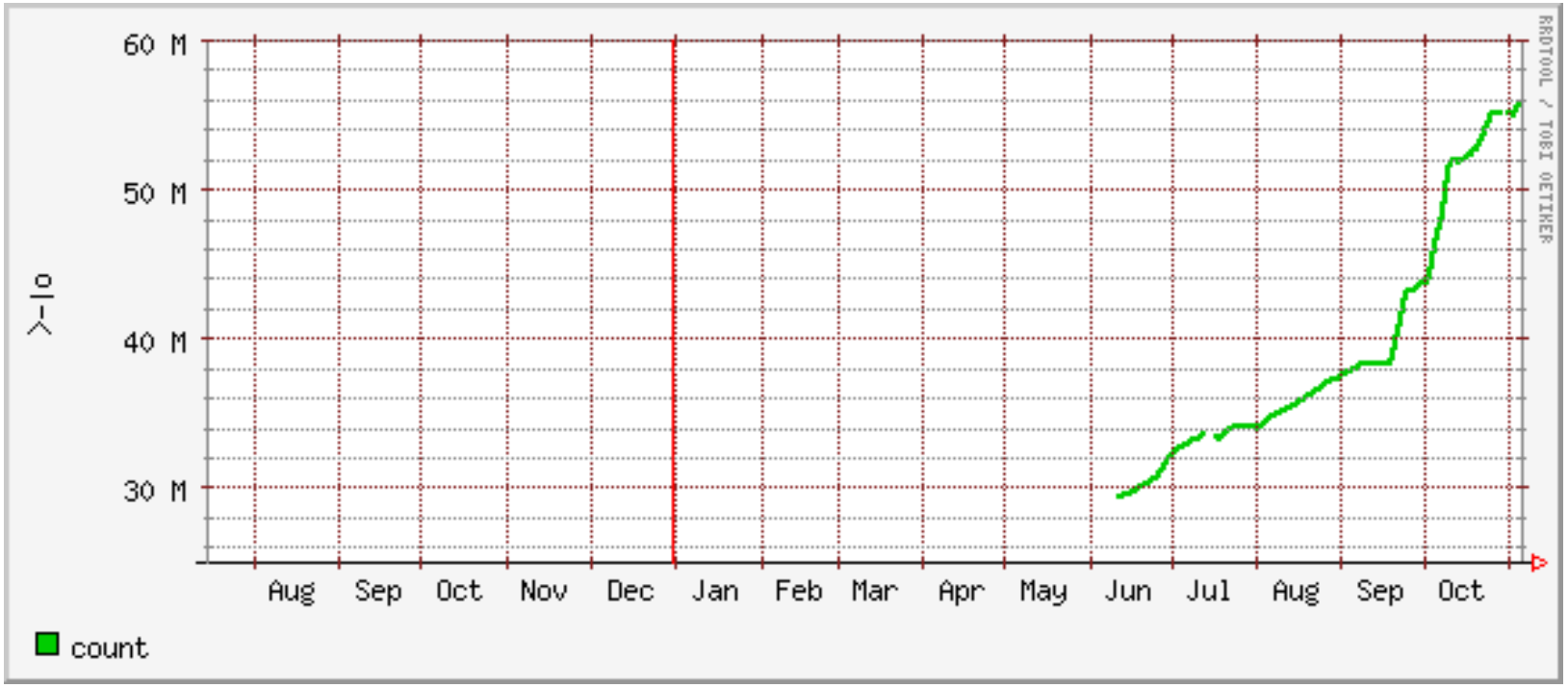
- XML (XSLT šablony ještě ne)
- PDF
- DOC (MS Word)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Část 4 – Údaje z provozu

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Velikost databáze (1)



- Počet dokumentů

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

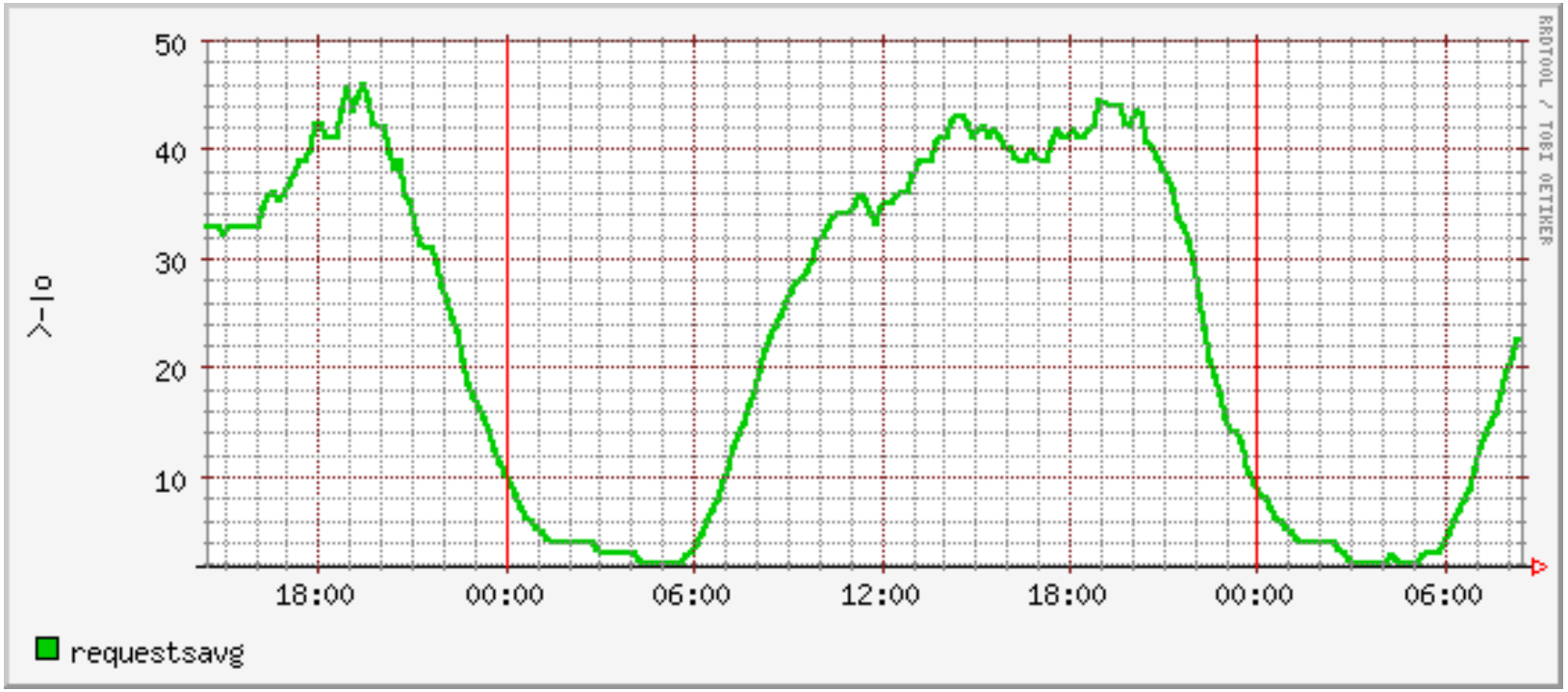
Velikost databáze (2)

Počet dokumentů	55 miliónů
Indexy	120 GB
Obsah dokumentů (texty)	200 GB
Průměrný text	4 kB / dokument

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

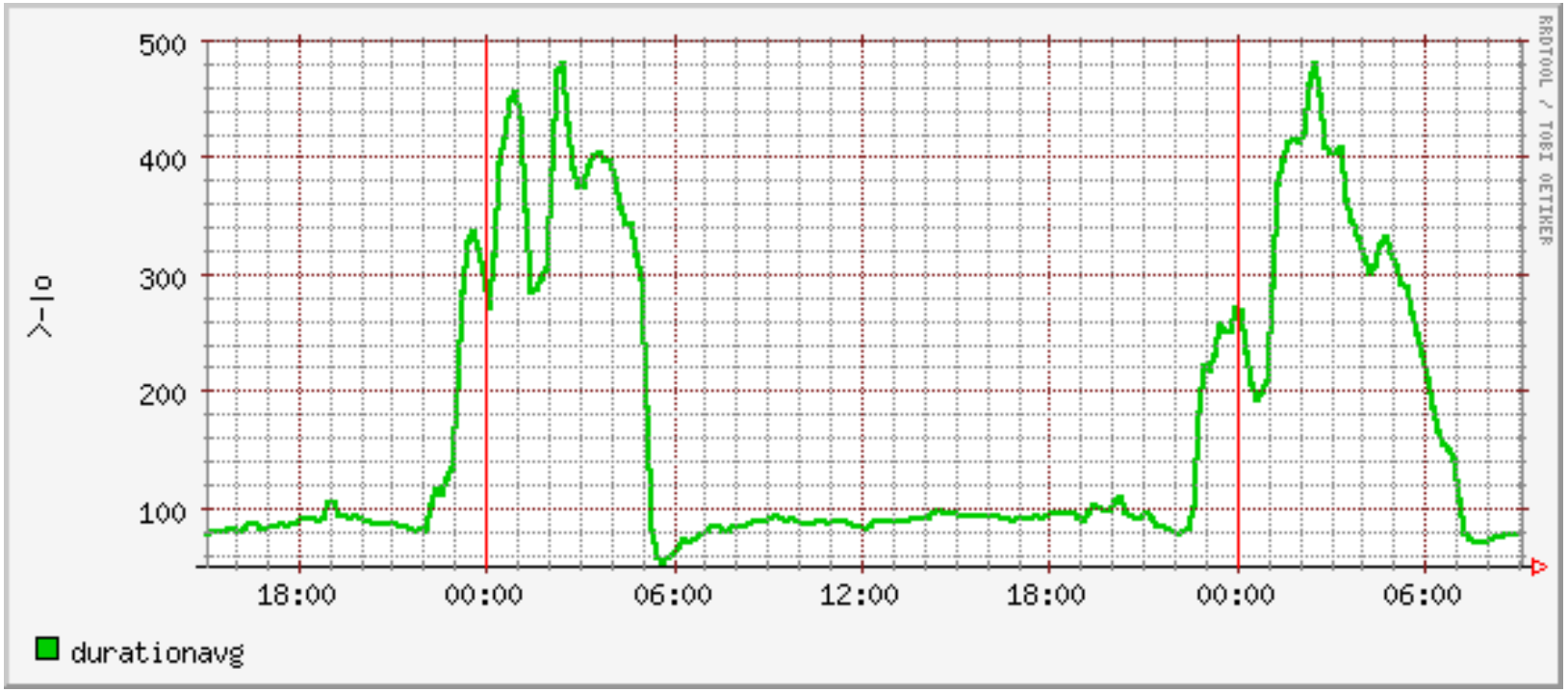
Zátěž během dne



- (pondělí) 1/6 zátěže => 250 dotazů/s

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Doba odezvy během dne



- Doba odezvy v msec

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Výkon robota

Rychlost stahování	> 100 stránek / sec
Průměrná stránka	~10 kB
Denní objem	10 miliónů dokumentů 100 GB dat

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Stáří dokumentů ve dnech

Minimální	0,9
Maximální	125
Průměr	8,8
Medián (prostřední)	3,8
Modus (nejčastější)	(1,7; 13,8)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Reklama

- **Uncle Sam wants you!**

http://vyvojari.seznam.cz

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Budoucnost?

- Lingvistické nástroje
  - Opravy překlepů,
  - Synonyma,
  - Hodnocení textu…
- Statistické zpracování dat
  - Logy,
  - Odkazové sítě…

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Konec

Děkuji za pozornost.
More info http://fulltext.sblog.cz

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM