Inspektor webových stránek

Crawlcheck - rozšiřitelný webový robot

Alexandr Mansurov

Problém

- Dlouhodobě udržovaný, rozsáhlý webový celek.
- Provázanost odkazů.
- Standardy.

Cíle práce

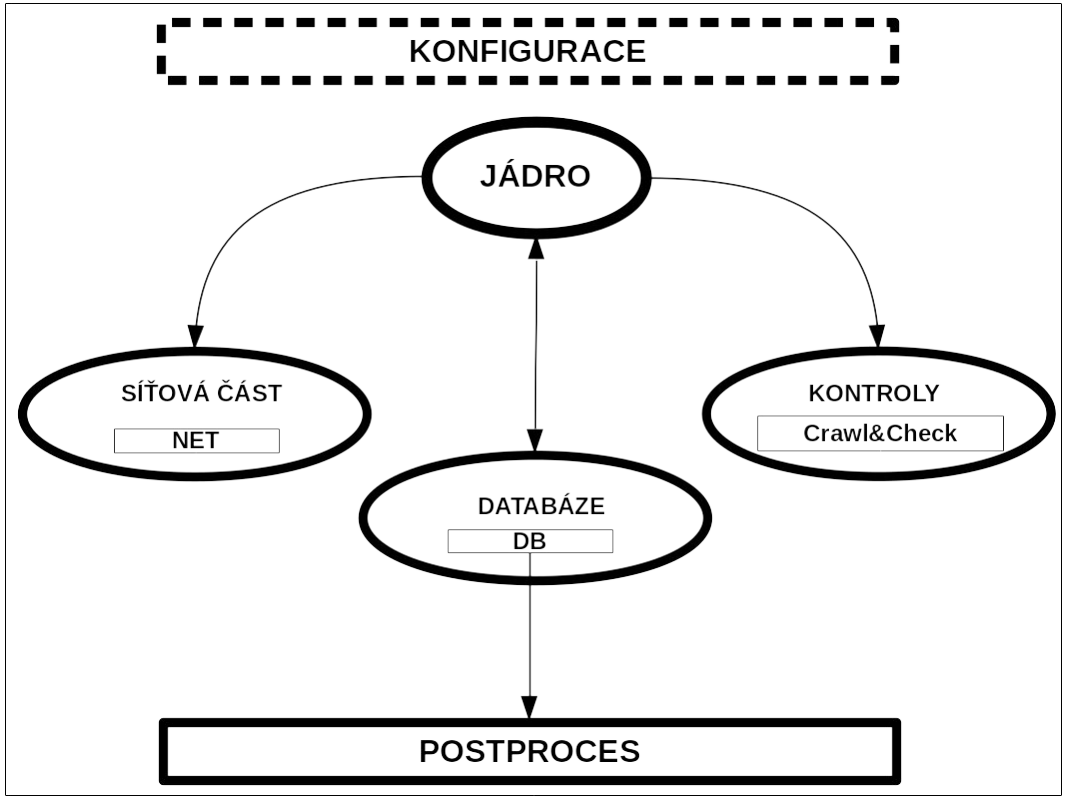
- Automatická kontrola webových celků.
- Konfigurovatelnost.
- Různé typy kontrol, rozšiřitelnost.
- Výstup.

Existující nástroje

- Kontrola odkazů
  - omezený počet
  - pouze z jedné stránky
- Multifunkční
  - velmi omezený počet stránek
- Specializované
  - jeden typ kontroly, jedna stránka
- Knihovny

Řešení

- Webový robot.
- Pluginy a existující nástroje.
- Konfigurace.



Konfigurace

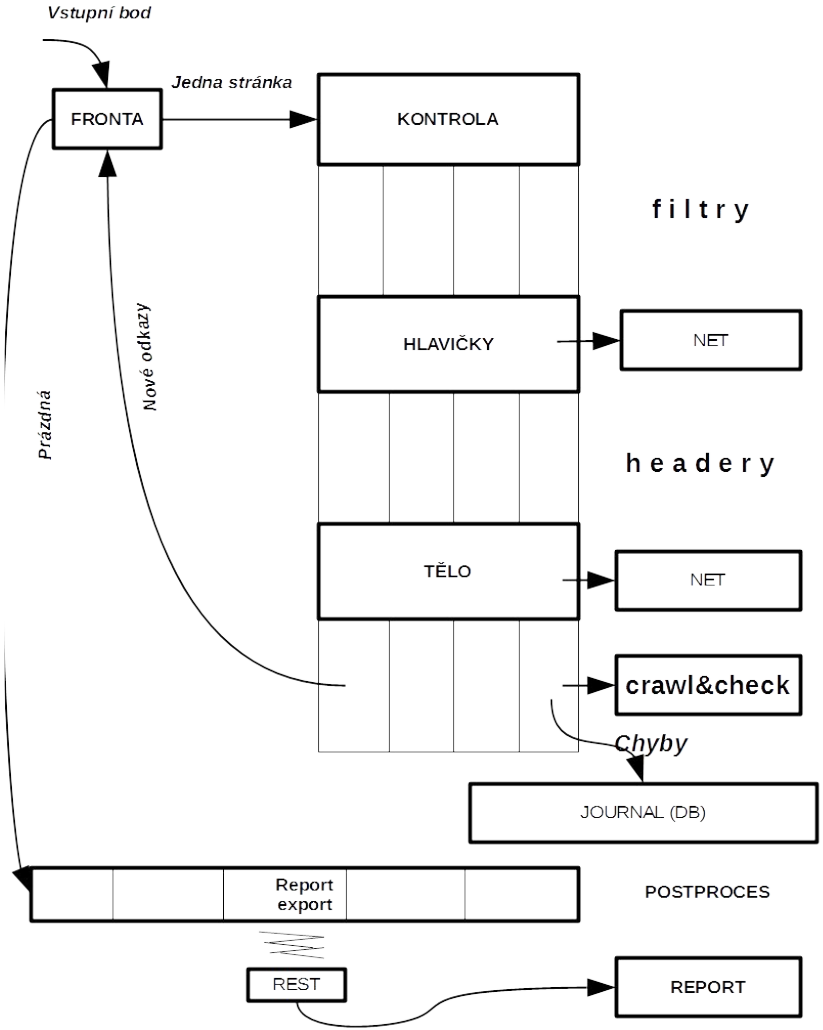
- Rozsah kontrol.
- Pravidla pro adresy.
- Pravidla pro typy obsahu.
- Filtry.
- Parametry.

```
regexes:
 -
       regex: "http://ksp.mff.cuni.cz/(?!profil|forum).*"
       plugins:
       - linksFinder
       - tidyHtmlValidator

content-types:
 -
   "content-type": "text/html"
   plugins:
       - linksFinder
       - tidyHtmlValidator
```

Pluginy

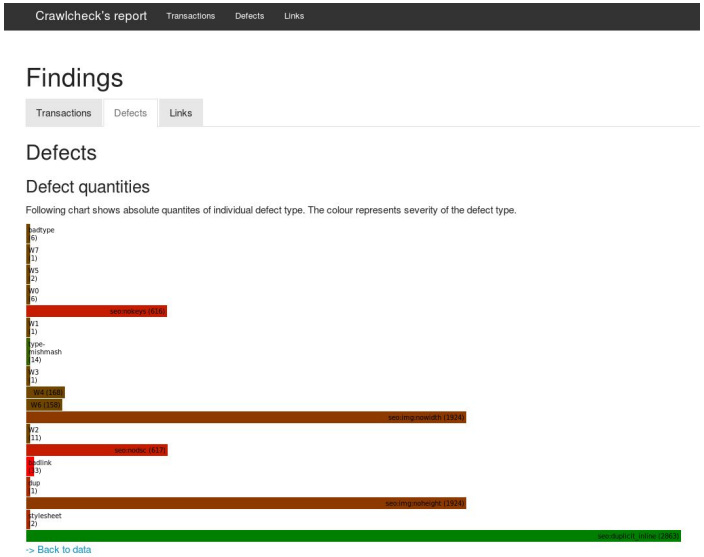
- Checker.
  - Kontrola obsahu.
- Crawler.
  - Hledání odkazů.
- Filter.
  - Filtr (např. robots.txt)
- Header.
  - Na základě hlaviček HTTP.
- Postprocessor.
  - Výstup.

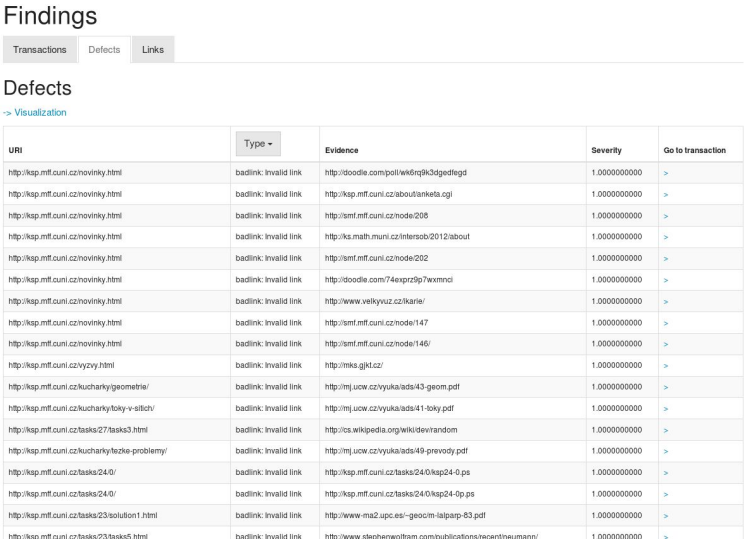


Typy kontrol

- Existující knihovny, které řeší problém.
  - Validace syntaxe HTML, CSS.
- Existující knihovny, které parsují formát.
  - Nutná vlastní logika nebo větší integrace.
  - Robots.txt, sitemaps.
  - HTML: detekce odkazů a další kontroly.
- Implementace.
  - Detekce duplikátů.

Zobrazení výsledků





Path to page

Transaction information

URI	Method	Response status	Content type	Depth
http://mj.ucw.cz/vyuka/ads/43-geom.pdf	GET	None	None	3

Back to transaction details

1. http://mj.ucw.cz/vyuka/ads/43-geom.pdf (Level: 3)
2. http://ksp.mff.cuni.cz/kucharky/geometrie/ (Level: 2)
3. http://ksp.mff.cuni.cz/tasks/ (Level: 1)
4. http://ksp.mff.cuni.cz/ (Level: 0)

Shrnutí

- Rozšiřitelný, konfigurovatelný webový robot.
  - Průchod webem, rozsah kontrol, výstup.
- Kontrola syntaxe HTML a CSS.
- Kontrola odkazů.
  - Detekce odkazů v HTML a ze sitemap.
- Kontrola sitemap.xml - omezení daná standardem.
- Podpora robots.txt, detekce sitemap.
- Detekce duplikátů.
- Detekce použití nesémantických značek a atributů.