

Inspektor webových stránek

Crawlcheck - rozšiřitelný webový robot

Problém

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

Cíle práce

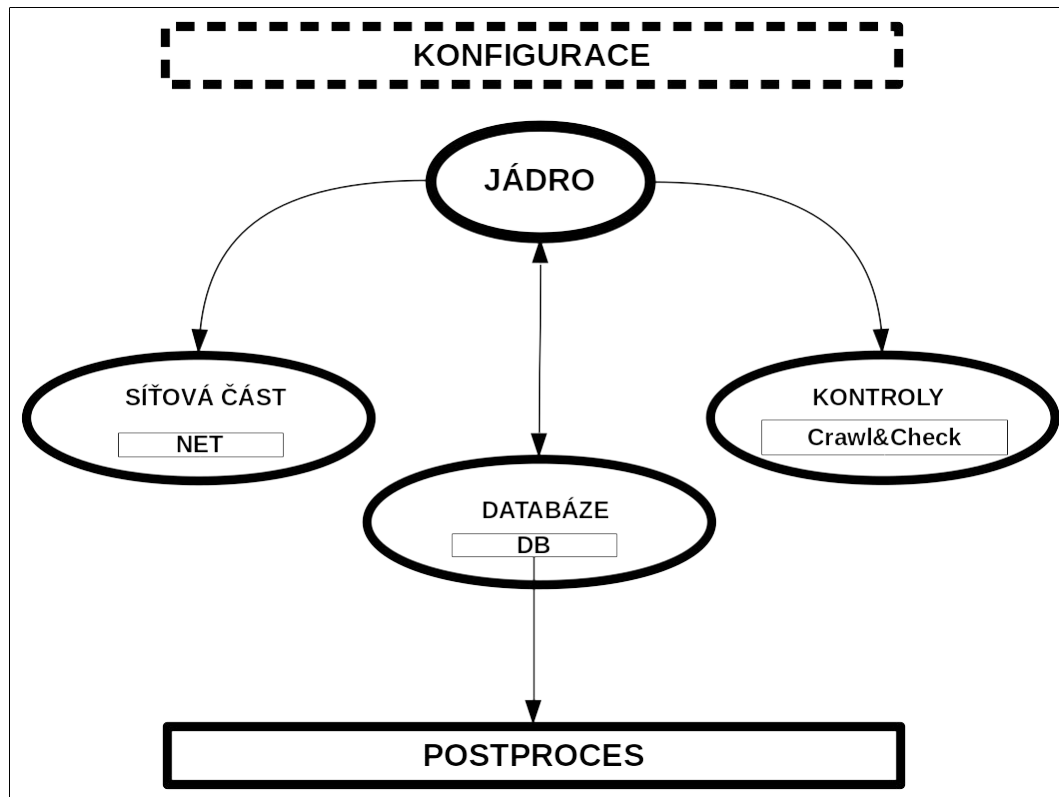
- Automatická kontrola webových celků.
- Konfigurovatelnost.
- Různé typy kontrol, rozšiřitelnost.
- Výstup.

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"
```

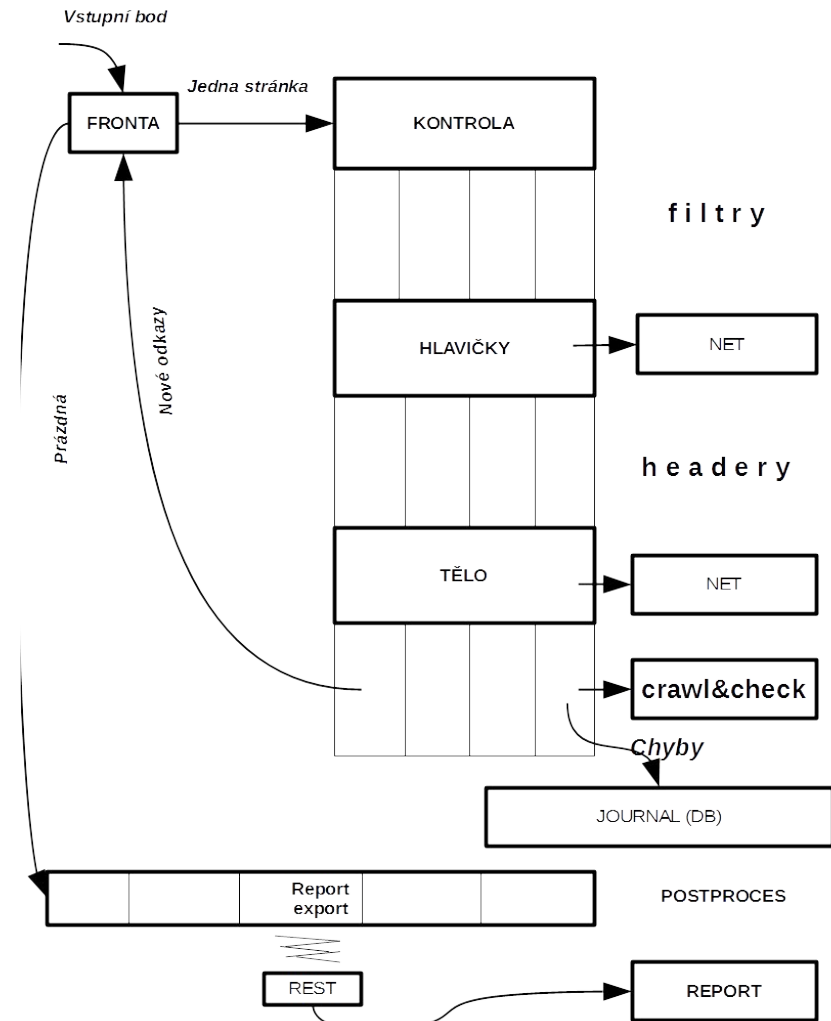
```
  plugins:
```

```
    - linksFinder
```

```
    - tidyHtmlValidator
```

Plugins

- 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

Crawcheck's report Transactions Defects Links

Findings

Transactions Defects Links

Defects

Defect quantities

Following chart shows absolute quantities of individual defect type. The colour represents severity of the defect type.



-> Back to data

Findings

Transactions Defects Links

Defects

-> Visualization

URI	Type	Evidence	Severity	Go to transaction
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://doodle.com/poll/wdqr9K3dgedfgd	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://ksp.mff.cuni.cz/about/anketa.cgi	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://smf.mff.cuni.cz/node/208	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://ks.math.muni.cz/Intersob2012/about	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://smf.mff.cuni.cz/node/202	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://doodle.com/74exprzlp7wxmndi	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://www.velkyvuz.cz/karie/	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://smf.mff.cuni.cz/node/147	1.0000000000	>
http://ksp.mff.cuni.cz/znovinky.html	badlink: invalid link	http://smf.mff.cuni.cz/node/146/	1.0000000000	>
http://ksp.mff.cuni.cz/vyvy.html	badlink: invalid link	http://mka.gkl.cz/	1.0000000000	>
http://ksp.mff.cuni.cz/kucharky/geometrie/	badlink: invalid link	http://mj.uow.cz/vyuka/ads43-geom.pdf	1.0000000000	>
http://ksp.mff.cuni.cz/kucharky/boky-v-sitci/	badlink: invalid link	http://mj.uow.cz/vyuka/ads41-boky.pdf	1.0000000000	>
http://ksp.mff.cuni.cz/tasks/27/tasks3.html	badlink: invalid link	http://cs.wikipedia.org/wiki/dev/random	1.0000000000	>
http://ksp.mff.cuni.cz/kucharky/lezle-problemy/	badlink: invalid link	http://mj.uow.cz/vyuka/ads49-prevody.pdf	1.0000000000	>
http://ksp.mff.cuni.cz/tasks/24-0/	badlink: invalid link	http://ksp.mff.cuni.cz/tasks/24-0/ksp24-0.ps	1.0000000000	>
http://ksp.mff.cuni.cz/tasks/24-0/	badlink: invalid link	http://ksp.mff.cuni.cz/tasks/24-0/ksp24-0p.ps	1.0000000000	>
http://ksp.mff.cuni.cz/tasks/23/solution1.html	badlink: invalid link	http://www.ma2.upc.es/~geocom-lalparp-83.pdf	1.0000000000	>
http://ksp.mff.cuni.cz/tasks/23/tasks5.html	badlink: invalid link	http://www.elehenewoltfram.com/publications/sectant/traumann/	1.0000000000	>

Path to page

Transaction information

URI	Method	Response status	Content type	Depth
http://mj.uow.cz/vyuka/ads43-geom.pdf	GET	None	None	3

Back to transaction details

1. <http://mj.uow.cz/vyuka/ads43-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.

